VILNIUS UNIVERSITY
MATHEMATICS AND INFORMATICS FACULTY
INSTITUTE OF COMPUTER SCIENCE
INFORMATICS MASTER STUDIES

SPATIAL-TEMPORAL CHANGE DETECTION IN SATELLITE IMAGING-MONITORING

Edvinis ir Laikinis Aptikimas Palydovinių Vaizdų Stebėjime

Master Thesis

Author: Kürşat Kömürcü (parašas)

Supervisor: assoc. prof. dr. Linas Petkevičius (parašas)

Reviewer: prof. dr. Aistis Raudys (parašas)

Vilnius – 2025

Acknowledgements

I would like to express my deepest gratitude to my esteemed supervisor,
assoc. prof. dr. Linas Petkevičius, for his unwavering support throughout the preparation of this thesis. Author also thanks to Lithuanian Research Council for partially funding research ”New generation multi-task recognition from satellite image algorithms for climate monitoring” (Nr. S-MIP-23-45)

Santrauka

Baigiamajame darbe analizuojami palydoviniai vaizdai ir pokyčių aptikimas taikant skirtingus giliojo mokymosi metodus. Pirmasis eksperimentas buvo panašus į semantinį segmentavimą, suporuotas su vektorine autoregresija, skirta erdvinių ir laiko pokyčių analizei, o antrajame eksperimente naudojama CLIP pagrįsta klasifikavimo be pavyzdžių klasifikacija, kad būtų galima veiksmingai aptikti pokyčius be duomenų anotavimo. Be to, buvo atlikta aprašų pakitimų klasifikacija, kur aprašai buvo sugeneruoti naudojant Lamos modelį. Darbe toliau tiriamas MiniCPM-V modelio naudojimas palydovinio vaizdo atpažinimui įvairiuose duomenų rinkiniuose. Galiausiai, pristatomas novatoriškas algoritmų apjungimas, integruojantis tiksliai suderintą stabilią difuziją ir apmokytus CycleGAN modelius, siekiant sujungti nevienalyčius duomenų rinkinius, sujungiančius aprašus, RGB vaizdus ir daugiaspektrinius Sentinel-2 duomenis, taip sukuriant sintetinius daugiaspektrinius vaizdus.

Raktiniai žodžiai: palydoviniai vaizdai, gilus mokymasis, didelių kalbų modeliai, generatyvus modeliai, vaizdų aprašai, vaizdo transformacija

Abstract

In this master thesis, satellite imagery and change detection analyzed and by using different deep learning techniques. The first experiment was a UNet-like semantic segmentation paired with vector autoregression for spatial-temporal change analysis, and the second experiment employs CLIP-based zero-shot classification to effectively detect changes without the need for extensive labeling. Additionally, caption based change classification done which these captions generated using Llama model. The work further investigates the use of the MiniCPM-V model for satellite image recognition across diverse dataset. Finally, an innovative pipeline that integrates fine-tuned Stable Diffusion and our trained CycleGAN models is introduced to unify heterogeneous datasets merging captions, RGB images, and multispectral Sentinel-2 datathereby generating synthetic multispectral imagery.

Keywords: Satellite Imagery, Deep learning, Large Language Models, Generative AI, Image Captioning, Image Transformation

Contents

1. Introduction

The rapid advancement of remote sensing technology has significantly increased our capacity to monitor changes on the Earth through high-resolution satellite images. These technological advancements offer a wide range of applications, from monitoring natural disasters to managing agricultural areas, leveraging data obtained from space with advanced remote sensing methods and algorithms [AA19].

Focusing on a crucial area within remote sensing, this thesis explores the detection of spatial-temporal changes in satellite imagery—a topic of growing importance. Successful detection of these changes is of paramount importance to a number of important applications, such as early warning systems for natural disasters and tracking urbanization processes.

Spatial-temporal detection is the capability of detecting and analyzing the changes in the spatial, along with the temporal, aspects of satellite images. Spatial changes can be land cover, infrastructure development, or deforestation, whereas temporal changes are changes over time like seasonal changes, long-term trends, or frequency of abrupt natural occurrences like earthquakes and floods [AA19].

The accomplishment of correct and timely detection of these changes is not only a technological hurdle but also imperative in solving some of the most pressing societal and environmental challenges. To illustrate, early detection of land use changes can provide knowledge for sustainable urban planning, and the timely detection of natural environment changes can enhance disaster response and recovery efficiency.

The advent of sophisticated remote sensing technology has significantly enhanced our capacity to monitor and assess these dynamic landscapes at unprecedented resolutions and scales. This thesis focuses on the central role of spatial-temporal change detection in remote sensing that comprises both detecting and deciphering the subtle changes taking place in spatial and temporal dimensions of the Earth's surface [AA19].

As environmental issues at the global level intensify and the imperative for sustainable development grows, satellite imagery's aerial perspective becomes ever more important to support well-informed decision-making and effective policy-making across a variety of critical applications [JC13].

This research deals with state-of-the-art remote sensing techniques to improve detection and interpretation of spatial and temporal changes in satellite imagery, thus opening up avenues to more effective monitoring systems. The research also mentions the application of novel machine learning algorithms, such as visual language models, cycle gan and stable diffusion.

2. Aims and Tasks

The main goal of this research is to investigate and propose a new algorithms based on visual language models to identify spatial-temporal changes

1. Make a scientific literature review and identify state of the art algorithms.

2. Formulate mathematical problem by incorporating visual language models for change detection.

3. Create a dataset for satellite image captioning to support future research and applications in automated image caption generation.

4. Run empirical experiments to validate proposed algorithms on publicly available new datasets.

3. Scientific Literature Review

Constructing a dataset necessitates a comprehensive understanding of the specific domain. Numerous sources of remote sensing data are available, thanks to various missions conducted by government agencies that collect extensive satellite imagery and make it freely accessible to the public. These sources will serve as the foundation for the dataset. Once compiled, the dataset will need to be benchmarked to evaluate its relevance and effectiveness. This includes the formulation of tasks, which are critical as they guide machine learning models toward improved performance.

This section aims to provide an overview of the dataset creation process. Section 3.1 identifies key sources of satellite imagery data from which multi-spectral data is harvested. Section 3.2 discusses existing datasets, which are significant in understanding the optimum characteristics of a new dataset. Progressively, Section 3.3 discusses common issues faced by developers of these datasets. Section 3.4 examines literature review by methodologies. Section 3.5 discusses several loss functions, providing a first glimpse of possible tasks that can be done to optimize outcomes of dataset analysis. Lastly, Section 3.6 presents some examples of visualization.

3.1. Data Resources

Creating a comprehensive dataset entails exploring various sources of satellite images. The next section is an overview of key satellite missions used for remote sensing and their respective applications in environmental monitoring and analysis:

- Sentinel-1 is a mission of the Copernicus program including two satellites with C-band Synthetic Aperture Radar (SAR) ensuring imagery regardless of the weather. This is vital for continuous monitoring with a six-day revisit time [1].
- Sentinel-2 is also under the Copernicus program, the mission delivers high-resolution, multispectral imagery with up to 10-meter resolutions, which is suitable for detailed land cover and vegetation health monitoring. The twin satellites guarantee a five-day revisit time at the equator [2].
- Sentinel-3 mission is meant to observe Earth's ocean, land, water, and atmosphere. It delivers high-quality optical, radar, and altimetry data that are essential in tracking sea surface topography, sea and land surface temperature, and ocean and land color. Its instruments are essential for environmental and meteorological monitoring [3].
- MODIS (Moderate Resolution Imaging Spectroradiometer) Onboard NASA's Terra and Aqua satellites, MODIS provides daily global coverage, facilitating the monitoring of the

[1] https://www.copernicus.eu/en/about-copernicus/infrastructure/space-component/sentinel-1

[2] https://www.copernicus.eu/en/about-copernicus/infrastructure/space-component/sentinel-2

[3] https://www.copernicus.eu/en/about-copernicus/infrastructure/space-component/sentinel-3

atmosphere, oceans, and land with its broad swath width of 2330 km [4].

- Landsat is a longstanding series of Earth observation satellites since 1972, providing multispectral and thermal imagery crucial for environmental research, with a resolution as fine as 30 meters and a 16-day revisit cycle [5].

- GOES (Geostationary Operational Environmental Satellites) provide real-time data important for weather forecasting, severe storm tracking, and meteorological research. They are geostationary, thus constantly monitoring the same area of the Earth [6].

- WorldView is operated by Maxar Technologies, this constellation offers very high-resolution optical imagery, up to 30 cm, facilitating applications in urban planning, mapping, and defense and intelligence [7].

- NOAA Satellites is managed by the National Oceanic and Atmospheric Administration, these satellites monitor the climate and the environment, contributing significantly to weather forecasting and environmental monitoring [8].

Each satellite system provides unique capabilities that are instrumental in addressing specific research needs. Researchers rely on these platforms to build datasets that underpin studies in a wide array of fields including environmental science, urban development, and disaster response.

3.2. Datasets

Datasets are the primary source of knowledge when training the model. Their unique characteristics allow scientists to select the best-suiting one. It also forces researchers to carefully weigh each dataset from a wide variety of properties. The dataset creation process also benefits from the research of relevant examples. To gather insight on dataset creation progress we need to formulate criteria. These measures will allow emphasizing datasets that have had similar challenges and in general are related in terms of geographic information, and data gathering processes. The following table shows that commonly used datasets and their features 1.

3.3. Challenges Creating Datasets

Creating comprehensive datasets for spatial-temporal change detection in satellite imaging is fraught with numerous challenges that can impact the accuracy and applicability of research findings. One of the foremost challenges is the availability and accessibility of high-quality satellite imagery. While numerous satellite missions provide data, access to high-resolution imagery often involves substantial costs or restrictive licensing, limiting its use in academic research [WCR+19].

[4] https://modis.gsfc.nasa.gov/
[5] https://www.usgs.gov/core-science-systems/nli/landsat/landsat-9
[6] https://www.goes.noaa.gov/
[7] https://www.maxar.com/products/worldview-legion
[8] https://www.noaa.gov/satellites

Dataset Name	Images	Coverage	Use Case	Task	Reference
LEVIR-CD	637 pairs	Urban Areas	Building Change Detection	Segmentation	[CS20]
LEVIR-CC	10077 pairs	Urban Areas	Image-Text Matching	NLP	[LZC+22]
DSIFN	3600 pairs	Varied	Semantic Change Detection	Segmentation	[ZYT+20]
S2Looking	5000 pairs	Global	Multi-temporal Change Detection	Segmentation	[SYC+21]
OSCD	26 Multispectral Pairs	Global	Multi-temporal Change Detection	Segmentation	[CLB+19]
HRSCD	291 pairs	Urban Areas	High-resolution Change Detection	Segmentation	[DLB+19]
SMARS	24 pairs	Multimodal	Aerial Remote Sensing	Segmentation	[RXY+23]
GVLM	17 pairs	Global Landslides	Landslide Mapping	Segmentation	[ZYP+23]
SI-BU	1328 pairs	Urban Areas	Building Usage Analysis	Segmentation	[LHY+23]
CNAM-CD	2503 pairs	Urban Areas	Urban Change Detection	Segmentation	[ZWD+23]
BANDON	2283 pairs	Off-nadir	Building Change Detection	Segmentation	[PWD+23]
DynamicEarthNet	552 pairs	Global	Dynamic Change Detection	Segmentation	[TKW+22]
CLCD	600 pairs	Agricultural	Crop Land Change Detection	Segmentation	[LCD+22b]
SYSU-CD	20000 pairs	Urban Areas	Urban Change Detection	Segmentation	[SLL+21]
S2MTCP	1520 pairs	Urban Pairs	Multi-temporal Urban Analysis	Self-Supervised Learning	[LMB+21]
Hi-UCD	359 pairs	Urban Areas	Urban Change Detection	Segmentaion	[TMZ+20]
Hyperspectral-CD	9986 samples	Varied	Hyperspectral Imaging	Segmentation	[LGH+18]
GETNET	3750 samples	Traffic Networks	Traffic Network Analysis	Segmentation	[WYD+19]
3DCD Dataset	472 pairs	Urban Areas	3D Change Detection	Segmentation	[CMR22]
URB3DCD	Did not mention	Urban Simulated	3D Urban Change Detection	Segmentation	[dLC21]
RSITMD	3603 images	Varied	Image-Text Matching	NLP	[YZF+21]
RSICD	10921	Varied	Image-Text Matching	NLP	[LWZ+]
UCMerceed	2100	Varied	Image-Text Matching	NLP	[YN10]

Table 1. Overview of Datasets for Satellite Imagery Change Detection

Furthermore, geopolitical restrictions can impede access to satellite data over certain regions, creating significant gaps in global datasets [Woo06].

Another major challenge is the heterogeneity of data sources. Satellite datasets from different missions vary significantly in terms of spatial and temporal resolutions, spectral bands, and data formats. This variation necessitates extensive preprocessing to standardize the datasets for analysis, which can be both time-consuming and computationally expensive [GHD+17].

Cloud coverage and adverse atmospheric conditions can also severely affect the quality of optical satellite imagery. Techniques such as Synthetic Aperture Radar (SAR) are used to overcome these issues; however, integrating SAR data with optical images requires sophisticated fusion techniques that are still under active development [SZS15].

Accurate change detection in satellite imagery is also made more difficult by a number of complicating factors, such as the occurrence of noise in the data and the natural variability of landscapes. Such factors introduce false positives and false negatives into change detection algorithms. Robust techniques for filtering noise and interpreting the data accurately are still a major challenge [RAA+05].

Synchronization of images over time, which is so important for successful change detection, is yet another challenge. The revisit rate of the satellite can be too low to observe high speed changes, especially in dynamic environments such as cities or disaster areas [ZTM+17]. This can result in missed alerts or stale information that compromises the integrity of change analysis.

Lastly, scalability of processing large satellite datasets is a fundamental challenge. The computational demand of processing, storing, and analyzing huge volume of satellite data requires strong infrastructure and effective algorithms. With datasets increasing with more satellites launched at ever reducing intervals, the demand for scalable solutions becomes even more pressing [LDC16].

3.4. Literature Review According To Methodologies

This section of the literature review provides a comprehensive overview of diverse methodologies, each of which is characterized by its own analytical strengths and application contexts. We discuss Spectral Index Based Methods in Section 3.4.1 which employ specific wavelengths to monitor changes in vegetation, water bodies, and built-up regions. Statistical Analysis Methods are examined for their strengths in identifying significant changes based on time-series data in Section 3.4.2. Section 3.4.3 discusses in detail the development and refinement of Change Detection Algorithms and their role in automating detection. Furthermore, we investigate the cutting-edge realms of Machine Learning & Deep Learning Techniques, which have revolutionized predictive accuracy and efficiency in Section 3.4.4. In Section 3.4.5 we will review Object-Based Image Analysis (OBIA) is highlighted for its precision in segmenting high-resolution images into meaningful, analyzable objects. Moreover, Synthetic Aperture Radar (SAR) Techniques are examined for their capability to penetrate cloud cover and provide all-weather, all-time monitoring capabilities which we will look at in Section 3.4.6. In Section 3.4.7, we will discuss about Visual Language Models and Zero-Shot Learning are considered for their potential to transform traditional methodologies by integrating advanced computational models with minimal human supervision. Together, these methodologies depict a vibrant spectrum of technological advancements driving the future of satellite change detection.

3.4.1. Spectral Index Based Methods

Spectral index based methods have become a cornerstone in the field of spatial-temporal change detection in satellite imaging. These methods exploit the distinctive spectral signatures recorded in multi-temporal satellite imagery for the detection and tracking of temporal changes, establishing its value in a wide range of applications from environmental monitoring to urban planning.

The combination of multi-spectral scale-invariant feature transform (M-SIFT) and robust statistical change detection methods has transformed the accuracy of geometric registration of multi-temporal satellite image analysis significantly. Integration is very important for trustworthy change detection, as per [AAE+11], to increase the reliability of subsequent analysis by effective alignment of multi-temporal data sets.

In addition, [BHC15] investigated the application of image fusion methods for change detection, this time in flood management. Through the utilization of local spectral distortions in guiding image fusion, their method is a new way of change detection, highly valuable in disaster response activities wherein prompt and correct assessment is extremely important.

New techniques of unsupervised change detection have also been proposed to improve detection accuracy. [RMA13] proposed a new technique based on the use of the ERGAS index for multi-temporal satellite image change detection. By processing all available spectral bands at the local level, this technique provides an improved detection mechanism that greatly improves the accuracy of subtle change detection.

Lastly, the integration of spectral and statistical indices, as developed by [SS18], surmounts the issues of heterogeneous image acquisition times and the mixed pixel problem. Their method

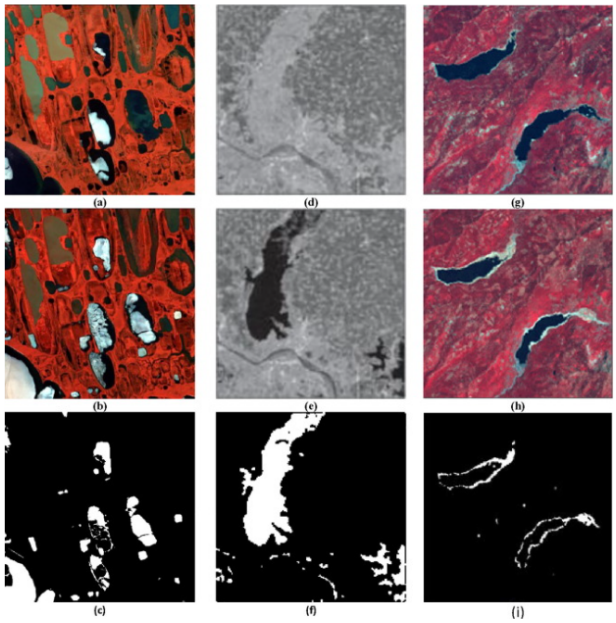


Figure 1: Satellite images (a) and (b) for ”Alaska”, Bitemporal images for ”Alaska” (c), Satellite images (d) and (e) for ”Bangladesh”, bitemporal images for ”Bangladesh” (f), Satellite images (g) and (h) for ”Reno and Lake Tahoe”, bitemporal images for ”Reno and Lake Tahoe” (i)[SS18].

demonstrates a comprehensive manner of change detection with high precision and robust classification of changes in heterogeneous landscapes 1.

These advances point to the growing sophistication of spectral index-based satellite image processing approaches, with a trend towards more integrated and computationally efficient methods that utilize both spectral and spatial data attributes to monitor and analyze spatial-temporal change.

3.4.2. Statistical Analysis Methods

Different statistical methods for spatial-temporal change detection in satellite images have been developed and tested. Beurs and Henebry (2005) designed an extensive statistical system to decompose long time series images of coarse spatial resolution satellites. The system provides strong procedures for multiple-comparison testing, seasonally adjusted Mann–Kendall trend estimation, and a sequence of orderly chained tests for quadratic land surface phenology models [BH05].

In addition, to detect land cover change from MODIS image time series, Lu et al. (2016) applied multidimensional arrays and the BFAST change detection algorithm. The algorithm is very effective in overcoming the spatial and temporal autocorrelation that characterizes satellite image time series and has strong change detection capability [LPS+16].

Further, Verbesselt et al. (2010) presented the BFAST (Breaks For Additive Season and Trend) method, where time series is decomposed into trend, season, and residuals. The procedure is highly effective in detecting and describing change in satellite image time series and thereby serves as a reliable tool for ongoing monitoring of environmental change through time [VHN+10] 2.

Overall, statistical methods for satellite image spatial-temporal change detection are complex and suit varying data complexity and uses. Statistical methods for satellite image spatial-temporal change detection are strong tools of remote sensing research that enhance our ability for monitoring

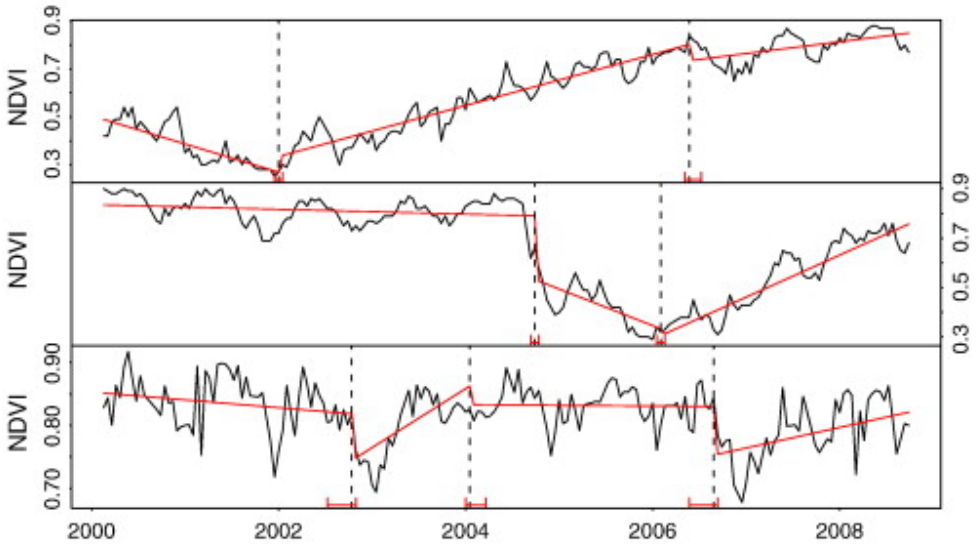


Figure 2: An individual MODIS pixel over a pine plantation, planted in 2001 (top), harvested in 2004 (middle), and with tree mortality in 2007 (bottom), showed changes in trend components (red) in a 16-day NDVI time series (black). The change date (—) and its red-colored confidence intervals are provided [VHN+10].

and change analysis in space and time.

3.4.3. Change Detection Algorithms

The development of change detection algorithms in satellite imagery is at the forefront of monitoring and understanding temporal changes on the Earth's surface. Various new methods have been proposed to improve the accuracy and reliability of the algorithms.

One of them is an unsupervised change detection algorithm that employs the nonsubsampled Contourlet transform and a pulse coupled neural network. The algorithm is very effective in maintaining stability and accuracy despite noise interferences, such as Gaussian and speckle noise [LJ22].

Moreover, the use of genetic algorithms for unsupervised change detection demonstrates an impressive accomplishment. By cost function minimization defined to detect changed and unchanged regions in satellite images, the method is proved to be efficient without any a priori assumptions about data [Çel10].

Another novel method includes the dual-tree complex wavelet transform (DT-CWT), which takes advantage of the multiscale and directional properties of satellite images to enhance change detection. The method has more accurate detection of subtle changes and better robustness under diversified types of noise conditions [ÇM08].

Further, integration of machine learning algorithms such as relevance feedback and query-answer models guarantees an interactive process that adapts according to the change detection requirements individual to a user. It thus constitutes a highly automated process with significantly low manual effort for hand-searching appropriate changes, guaranteeing highly efficient use with extensive applications [Sah13].

These innovations underscore the evolving essence of change detection algorithm research, emphasizing the continued demand for techniques capable of effectively processing and analyzing the increasing volume of satellite imagery available.

3.4.4. Machine Learning & Deep Learning Techniques

Machine learning and deep learning techniques are increasingly being used in satellite image change detection with far-reaching sophisticated methodologies for accuracy and efficiency. Some of the notable contributions in this regard are touched upon in this section.

A movel hybrid machine learning method integrates supervised and unsupervised learning techniques to improve the change detection using satellite images. The proposed technique applies clustering, soft labeling with fuzzy logic, Support Vector Machine (SVM), and Genetic Algorithm (GA) to substantially improve the change detection performance [PPT+20].

Deep learning has revolutionized remote sensing change detection and performs far better than the traditional methods. A comprehensive review of deep learning-based methods, including supervised, unsupervised, and semi-supervised approaches, highlights their effectiveness across various datasets such as SAR, multispectral, hyperspectral, and high-resolution images [SCK+22].

Moreover, convolutional neural networks (CNNs) have been successfully employed to generate change detection maps directly from bi-temporal satellite imagery. This strategy does away with hand-engineered features by harnessing deep feature learning to boost change detection performance [ALW16].

End-to-end methodologies, particularly using improved UNet++ architectures, facilitate direct learning of change maps from co-registered satellite image pairs, eliminating the need for intermediate processing steps and reducing error propagation [PZG19] 3.

Another innovative approach utilizes self-supervised learning to overcome the challenges of semantic supervision in satellite imagery. By transforming images into consistent feature representations without labeled data, this method enhances the robustness and accuracy of change detection [DMW+20].

The development of the Irregular-Time-Distanced Recurrent Convolutional Neural Network (IRCNN) addresses the challenge of non-uniform time intervals in satellite image sequences, a common issue arising from obstacles like cloud cover. The IRCNN model cleverly integrates a Siamese CNN with irregular-time-distanced LSTM and fully connected layers, significantly enhancing the ability to manage the temporal dependencies in satellite time series. This approach outperformed state-of-the-art methods [YQL+22]. Moreover, the study titled ”Optical Satellite Image Change Detection Via Transformer-Based Siamese Network” explores the application of Transformer architectures to the field of optical satellite image change detection. Transformers are adapted here to address the limitations of traditional convolutional neural network (CNN)-based methods, particularly their challenges in capturing long-range dependencies across images. This research introduces a novel approach by utilizing a Siamese network architecture that incorporates Vision Transformers (ViT). The method processes bitemporal satellite images as inputs through the Transformer-based network, effectively handling the complexities of change detection. The application of Siamese

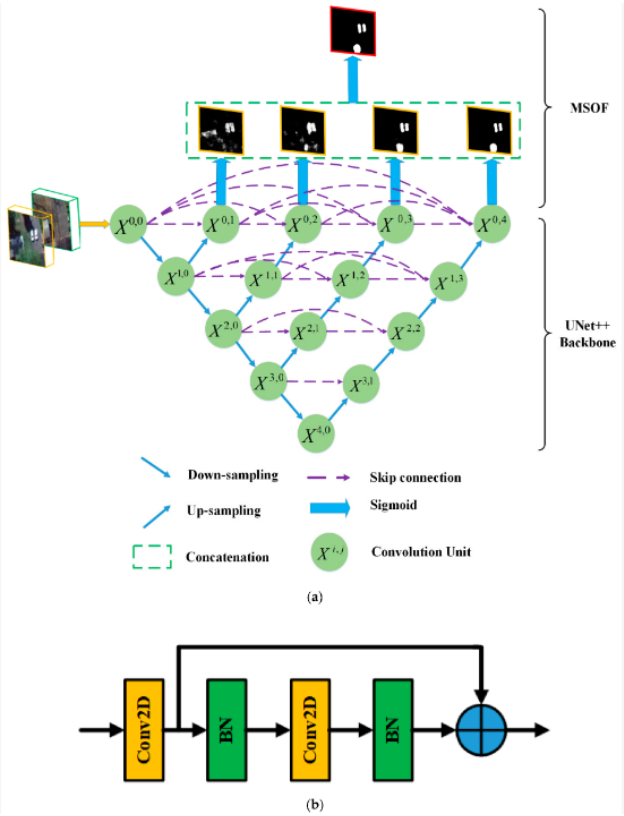


Figure 3: Flowchart of the UNet++: (a) schematic of the primary flowchart; (b) schematic of the convolutional block[PZG19].

ViT networks has demonstrated promising results in open change detection datasets, showcasing superior effectiveness and improvements over existing CNN-based models [WWL+22].

In addition, the Self-Supervised Multisensor Change Detection approach solves the problem of change detection in bi-temporal satellite images captured by various sensors, such as optical and SAR. Utilizing the deep clustering and contrastive learning in a self-supervised context, the approach maximizes the change detection efficiency in multi-modal data without requiring labeled data [SEZ21].

Besides, the Large-Area Land-Cover Change Monitoring with Time-Series Remote Sensing Imagery based on Transferable Deep Models proposes a new deep transfer learning approach to adaptive change detection. Dynamic time warping is applied to similar time series clustering and a time convolutional network to non-linear prediction of time series, and thus greatly enhances the efficiency and effectiveness of land-cover change monitoring [YWH+22].

These advancements highlight the great promise of machine learning and deep learning to make the process of change detection more accurate and efficient using satellite images.

3.4.5. Object-Based Image Analysis (OBIA)

Object-Based Image Analysis (OBIA) is an active method of satellite imaging, particularly for very high resolution (VHR) change detection. The method contrasts the traditional pixel-based methods as OBIA makes use of objects or groups of objects, and thus accuracy of change detection increases.

A significant advancement in OBIA for change detection is presented in a 2020 study, where a

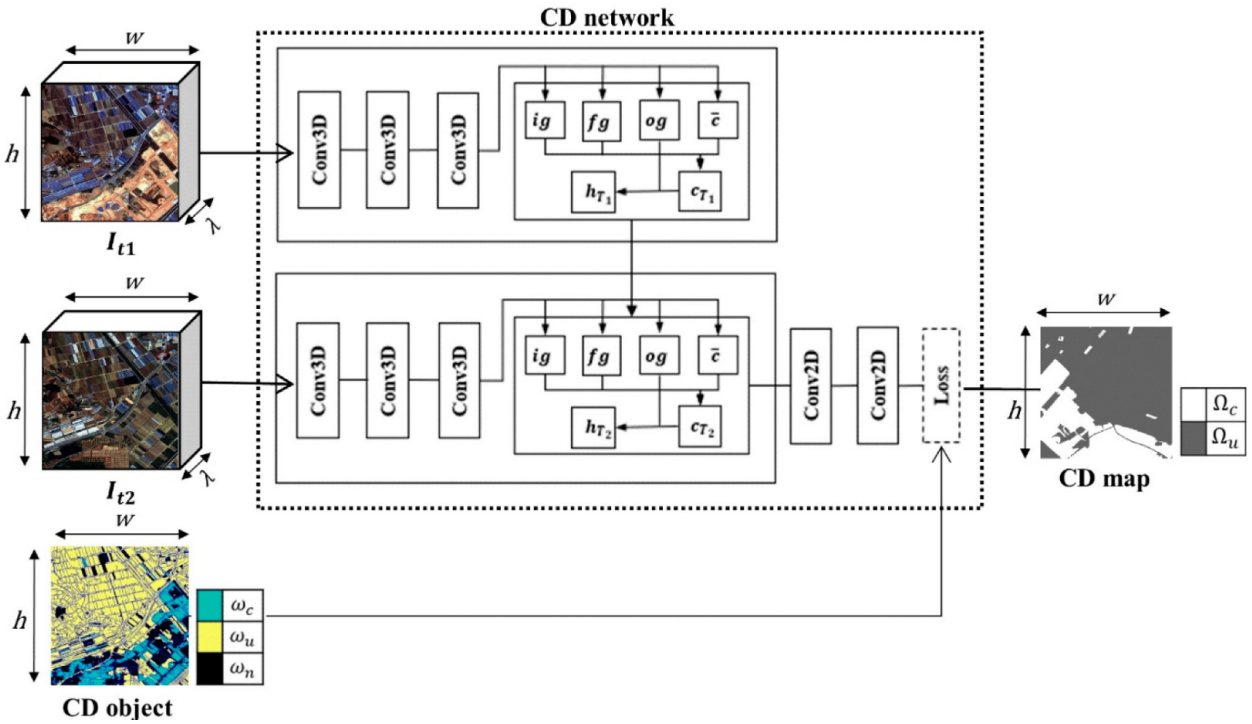


Figure 4: The CD network's structure integrates 3D convolutional layers to capture spatial–spectral features and convolutional LSTM layers to model temporal dependencies between the two feature sets. This is followed by two 2D convolutional layers that produce the final score map. As input, the network receives a pair of temporal images along with the CD object. After training, it outputs a binary change detection map. Here, w, h, and λ denote the image width, height, and number of spectral bands; Ω_c and Ω_u represent the change and no-change categories; and ω_c, ω_u, and ω_n refer to the change, no-change, and no-data classes in the initial CD map, respectively [SKH20].

deep learning network is employed to address uncertainties associated with objects in VHR satellite images. This approach defines the uncertainty linked to each object, leveraging three-dimensional convolutional layers and convolutional long short-term memory layers to improve change detection without relying on ground truth data. The process involves generating change detection objects through unsupervised methods, which are then used to train and update the change detection network iteratively, ensuring precise classification of changes within the area [SKH20] 4.

Another noteworthy development in OBIA is the integration of fuzzy knowledge systems to automate the detection of building changes in suburban areas from high-resolution satellite imagery. This method involves multi-resolution segmentation and fuzzy classification to distinctly categorize image objects into buildings and non-buildings, thereby identifying changes effectively between two temporal states. The OBIA framework employed here allows for a qualitative and quantitative evaluation of changes, which is crucial for accurate urban planning and monitoring [AMT13].

These studies highlight the utility of OBIA in leveraging structural and contextual information from satellite images, proving particularly useful in environments where high-resolution data is available, thus pushing the boundaries of traditional change detection methods.

3.4.6. Synthetic Aperture Radar (SAR) Techniques

In the area of SAR-based satellite image change detection, the latest progress has centered on applying deep learning to boost both accuracy and efficiency. A notable paper entitled "A Deep Learning Method for Change Detection in Synthetic Aperture Radar Images" proposes a novel CNN architecture that works directly on raw SAR data without any preprocessing. By sidestepping the generation of differential images, this approach markedly lessens their influence on final outcomes and proves resilient on both simulated and real-world datasets [LPC+19]. Another notable technique involves using coherence measurements between two SAR images to automatically detect changes, especially useful in applications like monitoring urban development or environmental changes [ZV92]. Additionally, a hierarchical approach has been developed to address change detection in SAR imagery with very high resolution, useful for surveillance applications, by utilizing multiscale techniques and semantic modeling to enhance the detection of changes in complex scenarios [BMB13].

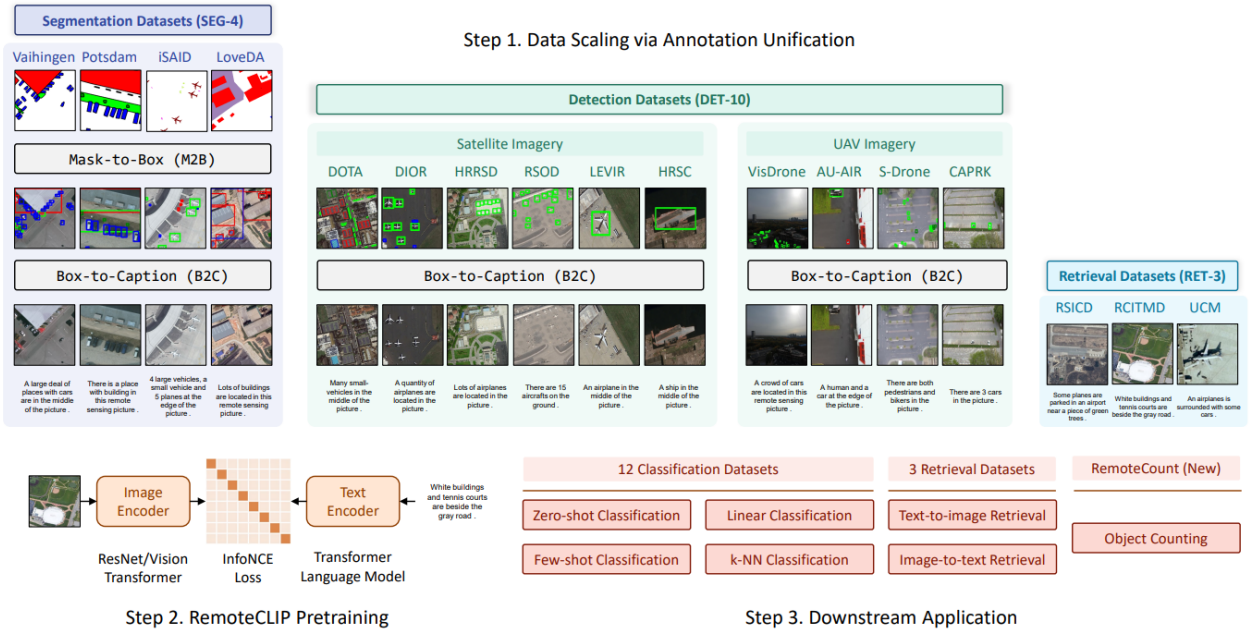


Figure 5: RemoteCLIP pipeline. Step 1: RemoteCLIP is first pretrained on a wide array of remote-sensing collections—10 object detection benchmarks (six from satellites and four from UAVs), four semantic-segmentation datasets, and three image–text corpora. To exploit the varied label formats, we introduce Box-to-Caption (B2C) generation and Mask-to-Box (M2B) conversion, and expand the total image–text training pool to twelve times the size of the original combined datasets. Step 2: We then carry out continual pretraining of the standard CLIP backbone, adapting it specifically for remote-sensing imagery. Step 3: Finally, we rigorously evaluate RemoteCLIP across seven distinct task categories using sixteen downstream datasets—including our newly assembled RemoteCount benchmark—to validate its strong performance and generalization. [LCG+23].

3.4.7. Visual Language Models and Zero-Shot Learning

Investigating vision–language architectures combined with zero-shot learning for satellite image change detection opens up exciting opportunities to advance remote sensing capabilities. Recent

work taps into the powerful synergy of pre-trained vision–language models like CLIP [RKH+21] and zero-shot techniques to tackle complex change detection challenges without relying on large, annotated datasets.

Zero-shot methodologies have been effectively utilized in remote sensing and satellite imagery for tasks such as scene categorization in high–spatial–resolution data [CXH+20; LLW+17; LWH+23] and monitoring urban transformations [FKG+21; FKG+22]. By leveraging semantic embedding vectors together with directed-graph representations, these approaches can recognize previously unseen scene classes without labeled examples [LLW+17]. Moreover, vision–language frameworks have been applied to zero-shot classification of remote sensing imagery: by exploiting pre-trained models that learn image–text correspondences, they achieve higher accuracy than existing techniques [KRR+23; LCG+23; LWH+23; RBE+23] (see Figure 5).

Additional zero-shot learning methods focusing on transferable object proposal mechanisms and vision-language knowledge distillation present innovative approaches to overcome the challenges of domain shift in zero-shot detection [GLK+21; SLW19].

3.5. Loss Functions for Satellite Imagery Analysis

In satellite imagery analysis, the choice of loss functions is pivotal in training machine learning models, as they quantify the discrepancy between model predictions and actual outcomes, thus guiding the optimization process. Mean Squared Error (MSE), given by $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, is commonly used for regression tasks such as environmental parameter prediction from satellite data [Smi20].

For classification tasks, Cross-Entropy Loss is essential; it calculates the loss by $L = -\sum_i y_i \log(\hat{y}_i)$, ideal for pixel-wise land cover classification or cloud detection [Joh19].

Dice Loss and Jaccard Loss are particularly beneficial for segmentation, addressing class imbalances prevalent in satellite imagery. Dice Loss, defined as $D = \frac{2|X \cap Y|}{|X|+|Y|}$, enhances the overlap between predicted and actual segmentation maps [Tho21]. Similarly, Jaccard Loss or Intersection over Union (IoU), formulated as $J = \frac{|X \cap Y|}{|X \cup Y|}$, is used to segment distinct regions accurately [Fis22].

Focal Loss, introduced to focus more on difficult, misclassified cases within an imbalanced dataset, modifies Cross-Entropy Loss by adding a focusing parameter γ, as $FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$, proving effective for rare event detection like oil spills or deforestation in vast satellite datasets [LGG+17]. Lastly, Huber Loss combines the properties of MSE and absolute error through the formula:

$$L_\delta(a) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{1}{2}\,a_i^2, & |a_i| \le \delta, \\ \delta\big(|a_i| - \frac{1}{2}\delta\big), & \text{otherwise.} \end{cases} \tag{1}$$

which is less sensitive to outliers and useful for tasks like surface elevation modeling where outliers are expected due to atmospheric conditions [Hua18].

Each of these loss functions is selected based on specific data characteristics and project objec-

tives to optimize the performance of machine learning models in handling the unique challenges of satellite imagery analysis.

3.6. Visualization

The visualization of spatial-temporal changes detected in satellite imaging is crucial for interpreting and communicating the dynamics of various phenomena on Earth. Efficient visualization techniques facilitate the understanding of complex data and support decision-making in domains like disaster management, urban planning, and environmental monitoring. This section clarifies several visualization techniques that are extensively used in the context of satellite imagery spatial-temporal change detection.

Time-series animation is a powerful technique that involves creating a sequence of images or frames that illustrate change over time. The method is particularly well-suited to illustrating cumulative changes such as urbanization, deforestation, or changes in vegetation with the seasons. Through processing satellite imagery into video format, it is simple to track and analyze the evolution of changes within the specified time period [LMB+04].

Change detection maps highlight the areas which have undergone great change between two time periods. Change detection maps are usually color-coded to reflect the type and extent of change. For instance, it is a common convention to use red for areas where vegetation cover has decreased and green for areas where vegetation cover has increased. Such maps are often generated using techniques like principal component analysis (PCA), post-classification comparison, and image differencing. This visual method enables quick identification of change patterns in large areas [Sin89].

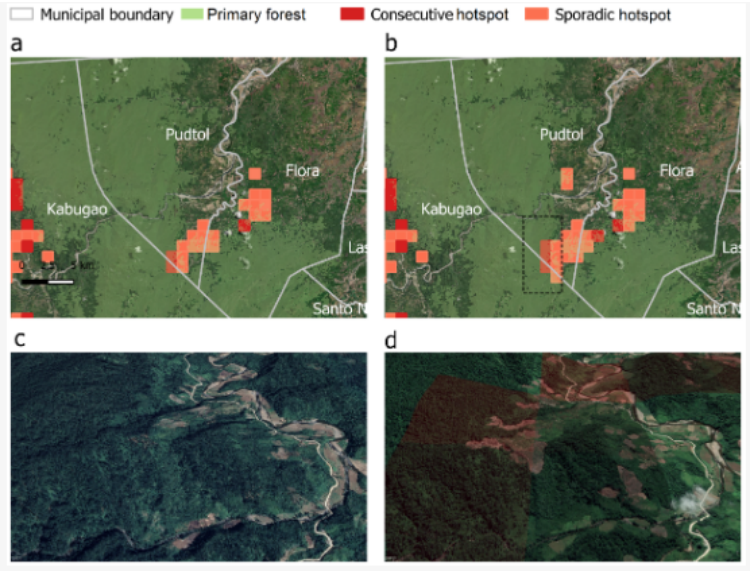


Figure 6: An example of hotspot progression detected by GLAD alerts in 2020—triggered by slash-and-burn activities across multiple municipalities in Apayao province. Panels (a) and (b) depict the hotspot maps for the first and second quarters, respectively, while panels (c) and (d) show the actual landscape in mid-2019 and mid-2020 (images from Google Earth Pro) [BHH+21].

Heatmaps are used to illustrate the magnitude of change over an area. Used on satellite imagery,

heatmaps can illustrate areas of high change, such as high urbanization or hotspots of deforestation. The technique of projecting different colors or shades based on different degrees of change makes it convenient to identify and target critical areas [BHH+21] 6.

Three-dimensional (3D) visualization adds depth to the analysis by providing a more realistic picture of landscape changes. Techniques such as digital elevation models (DEMs) and 3D rendering can be used to visualize topography, land cover, and land use changes. The method makes data more understandable by allowing users to view changes from various angles and perspectives [Str+21].

Multi-temporal composite images are a composite of satellite images at different times in one composite image. The method may involve techniques like image stacking or the use of spectral indices like the Normalized Difference Vegetation Index (NDVI). Composite images help to detect patterns and trends by visual comparison of different time points in a single frame [PVM+05].

Web interactive maps provide an active way to study spatial-temporal differences. Through the use of Geographic Information Systems (GIS) and web technology, interactive maps provide zooming, panning, and data interaction. Features such as layer toggling, time sliders, and query tools enable users to customize their view and perform detailed analysis [HSP08].

False color composites use combinations of different spectral bands to highlight specific features or changes in satellite images. For example, using near-infrared (NIR), red, and green bands can help in distinguishing vegetation health and land cover changes [LKC15].

Visualization techniques play a important role in spatial-temporal change detection in satellite imaging by transforming raw data into comprehensible and actionable insights. The selection of visualization technique depends on the specific application and the nature of the changes being studied. Whether through animations, heatmaps, 3D models, or interactive maps, effective visualization helps in better understanding and communication of the dynamic processes occurring on our planet.

4. Change Detection Experiments

4.1. Experiment 1: Semantic Segmentation for Change Detection in Satellite Imaging

This study uses advanced semantic segmentation techniques to address challenges in the detection of spatial-temporal changes in satellite imagery. The methodology integrates deep learning models for semantic segmentation and statistical approaches for temporal analysis [KP24a].

The primary components of the proposed methodology are the following.

- **Semantic Segmentation:** The semantic segmentation is of computer vision problem of assigning a class label to each pixel in an image from a predefined set of classes. Let us assume the input of format $\mathbf{X} \in \mathbb{R}^{c \times w \times h}$ of image consistent of tensor $\mathbf{X}$ with c - number channels, and width/height w, h, respectively. The semantic segmentation mask $\mathbf{X} \in \mathbb{R}^{L \times w \times h}$ contains L number of classes, where each pixel is assigned to one of the classes. Such models could predict the class of each pixel in the image [DK23]. In our experiments, we used the UNet-like model[9]. The pre-trained model had Building, Land, Road, Vegetation, Water and Unlabeled classes. For generic segmentation models, like Segment Anything Model [KMR+23] provide object mask prediction confidence score.

- **Vector Autoregression (VAR):** Vector autoregression (VAR) models the linear relationships among several time series simultaneously. As a multivariate extension of the univariate autoregressive (AR) model [HNR88], VAR is widely employed for forecasting. The VAR model can be expressed as follows:

$$y_t = \beta + \sum_{i=1}^{p} \Omega_i y_{t-i} + \epsilon_t$$

  where y_t is a $k \times 1$ vector of endogenous variables at time t, β is a $k \times 1$ vector of bias, Ω_i is a $k \times k$ matrix of coefficients for i-th lag, p is the order of the VAR process, and ϵ_t is a $k \times 1$ vector of error terms at time t. The confidence interval of VAR models could be used either dynamic, either fixed. In our case use t-distribution confidence interval which is same for each time step. The critical value for the confidence level α, in our experiments we used $\alpha = 0.05$. The VAR model was used using `Python` package `statsmodel`.

The investigation of change detections was applied on covering wide range of diverse cases. We randomly chose 100 coordinates over Baltic region (53,53100 - 59,69747 latitude values and 20,49722 - 28,22760 longitude) using uniform distribution. After that, we used COPERNICUS/S2 satellite in Google Earth Engine API for collect images of random chosen coordinates over 2022 - 2023 time period. In our experiments, we used pixel intensities of B4, B3 and B2 bands which represent red, green and blue colors. For each coordinates, we made predictions using `geospatial`

[9] https://github.com/ayushdabra/dubai-satellite-imagery-segmentation

Segment Anything Model [WO23] and collect IOU and score values. Class probabilities are collected using UNet-model. Cloud Probabilities collected using Google Earth Engine API. In such dataset for each coordinate consist of 11 features of Raw Pixel Intensity of B4, Pixel Intensity of B3, Pixel Intensity of B2, IOU, Scores, Probabilities of 6 classes and Cloud Probabilities.

For each point, we collected sentinel-2 RGB images using scale 10 zoom rate. Then, having surounding enviroment around the segmentation predictions was made using relevant models for each images. Such enables to have semantic information for each investigative pixel. After creating our dataset, we used VAR model for selected index and forecast $h = 12$ steps. The experiment we calculating root mean square error (RMSE), akaike information criterion(AIC) and confidence intervals for each feature using t distribution.

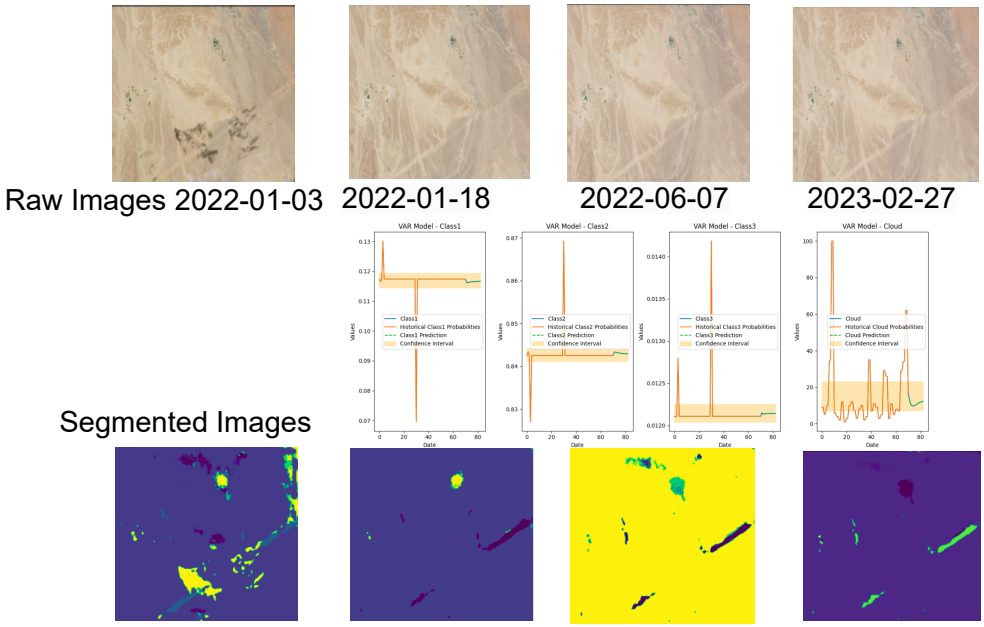


Figure 7: The illustrative example of confidence interval of prediction of the VAR model, which is used to detect the changes in the landscape.

The Fig. 7 presents the general pipeline of approach. The segmented image semantic information are added to vector time series models, thus while raw image data seems unchanged significantly, the semantic information allows to be additional control mechanism for quality assessment. Cloud probability are often used to remove untruthfull images, the same could be done by tracking unchanged situations. The illustrative case in Fig. 7 can be seen for index 6 in the Table 2 below. Also one can be seen in Table 2 that some testing images have high variation in raw data or some data was not overlaped (black/empty image) over specific flight and 0 observation fell in confidence interval.

Table 2. Summary Table

Index	Lat	Lon	RMSE	AIC	Fall In CI
0	57.9822	27.5759	0.206	-105.333	0
1	54.8303	21.8945	2.98e-15	None	0.9761
2	59.1785	24.5851	0.02	-112.393	0.4444
3	57.2123	24.1739	0.008	-58.051	0.5
4	55.4973	23.1317	5.84e-04	None	0.988
5	59.5876	25.7885	0.029	-114.48	0.3048
6	57.2948	22.5929	3.07e-04	-205.4	0.9761
7	53.6124	27.2380	5e-04	-51.1965	0.4352
8	54.6356	22.8023	0.115	-136.188	0
9	56.6370	20.7791	0.098	None	0

Note: RMSE and Fall In CI columns are values for Class2

4.2. Experiment 2: Zero Shot Classification for Change Detection in Satellite Imagery

In this study [KP24c], we introduce the primary model we utilized Contrastive Language–Image Pre-training (CLIP) to analyze satellite imagery [RKH+21]. CLIP, a model developed by researchers of OpenAI, is designed to form expressive features of images in the context of natural language descriptions, making it particularly suited for tasks such as zero-shot classification where the objective is to classify classes that the model had not encountered during its training phase [RKH+21]. Model consist from two encoder models $f_{image}(I|\theta_{image}) = T_{image} : \mathbb{R}^{d_w \times d_h \times d_c} \rightarrow \mathbb{R}^{d_e}$, $f_{text}(T|\theta_{text}) = T_{text} : \mathbb{R}^{K \times d_Z} \rightarrow \mathbb{R}^{d_e}$, where f_{text}, f_{image} neural networks encoding raw data to embeddings vectors T_{text} and T_{image}, respectively. The θ represents neural networks unknown parameters, $d_w \times d_h \times d_c$ - image dimensions, $K \times d_Z$ - text input. The model is pretrained on predicting $\hat{Y} = T_{image}^T T_{text}$ identification of corresponding pairs of images/text.

As input to the text encoder model, we provided an array containing the selected classes names common in satellite imagery of size $K = 32$ objects which are *landscape, forest, building, road, vehicle, bridge, river, lake, farmland, airport, runway, ship, railway, parking lot, cloud, wind turbine, stadium, school, hospital, industrial site, park, beach, mountain, glacier, desert, volcano, crater, island, wetland, quarry, dam* and *residential area*. These objects could potentially be identified within the satellite images. This array served as the textual descriptions against which the model evaluates the imagery, predicting the likelihood $\hat{Y} = T_{text}$ of each object's presence within the image. By transforming outputted logits via softmax to probabilities for each of these objects, indicating their presence within each image see Fig 8 9.

This process was systematically applied across all images within our datasets. By doing so, we were able to assess the model's ability to identify changes in satellite imagery, many of which were not included in the training dataset of the model.

Upon completing the analysis with the CLIP model, the next step involved processing the

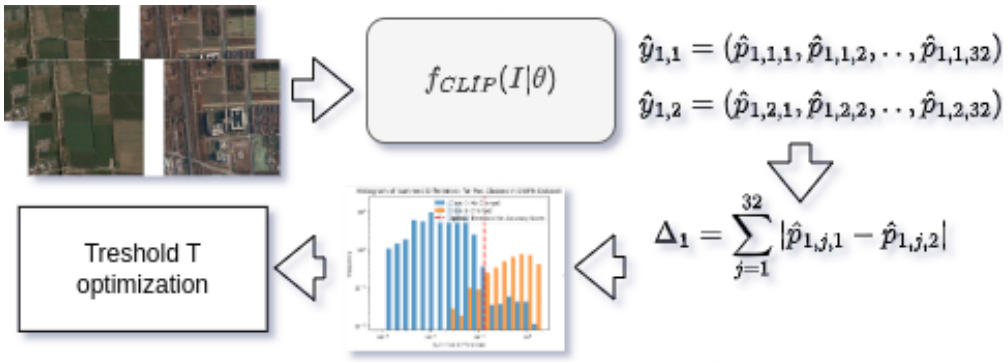


Figure 8: The pipeline of zero-shot learning: using pre-trained CLIP model from learned embeddings to classity in selected number of classes and using it differences for threshold optimization.

model's output probabilities for each satellite image. The core of our methodology was to determine the presence of any significant changes within each image based on a assessment of probabilities assigned to potential classes identified by the CLIP model, by calculating difference (2), for each image i in dataset. Where $\hat{Y}_{i,k} = (\hat{p}_{i,k,1},\hat{p}_{i,k,2},...,\hat{p}_{i,k,32})$, and $k = 1,2$, the reference and query images, respectively.

To achieve this, we first summed the probabilities difference across all identified objects for each image to create a composite likelihood score. This score was intended to reflect the overall presence of change-indicative features within the image, as recognized by the CLIP model.

The critical part of our methodology was to classify images into two categories: 'change' and 'no change'. This classification was based on a threshold optimization process, which aim to identify the optimal probability threshold T that distinguishes between the two categories based on selected F1-score metric. The optimization was formulated as follows:

1. **Summation of differences:** For each image, sum the probabilities differences of all potential objects detected by the CLIP model.

$$\Delta_i = \sum_{j=1}^{n} |\hat{p}_{i,j,1} - \hat{p}_{i,j,2}| \tag{2}$$

   where Δ_i is the sum of probabilities difference for the i-th image and $n = 32$ is the total classes.

2. **Threshold Optimization:** Determine the optimal threshold T by evaluating a range of threshold values to maximize classification metrics. For each candidate threshold value, classify images as 'change' if their summed probability exceeds the threshold, or 'no change' otherwise.

$$C_i(T) = \begin{cases} 1 & \text{if } \Delta_i > T \\ 0 & \text{otherwise} \end{cases}$$

   where $C_i(T)$ represents the classification of the i-th image under threshold T, with 1 indicating 'change' and 0 indicating 'no change'.

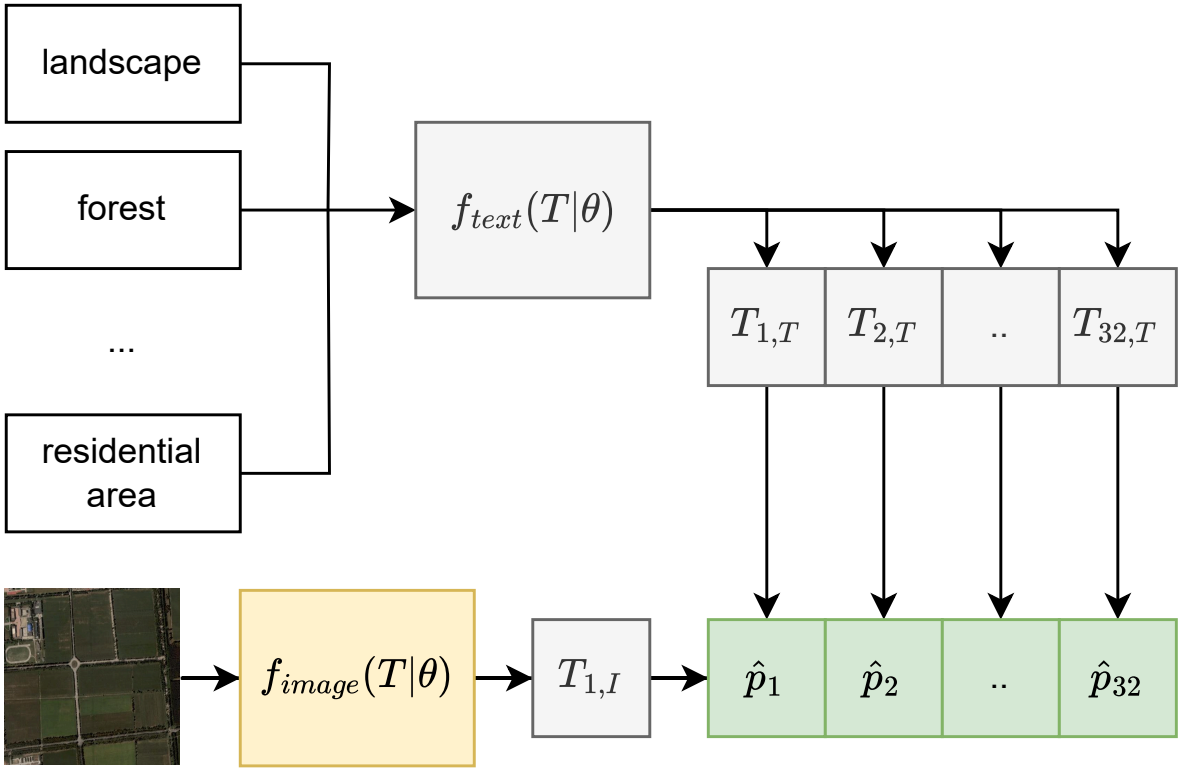


Figure 9: CLIP model inference.

3. **Evaluation Metrics:** For each threshold T, compute key evaluation metrics such as F1 score, recall, precision, and accuracy. These metrics evaluate the performance of each threshold in accurately classifying images into 'change' and 'no change' categories.

$$F1(T), Recall(T), Precision(T), Accuracy(T)$$

4. **Optimal Threshold Selection:** Identify the threshold that maximizes the desired metrics.

$$T_{\text{opt},F1} =_T F1(T)$$

$$T_{\text{opt},R} =_T Recall(T)$$

$$T_{\text{opt},P} =_T Precision(T)$$

$$T_{\text{opt},Acc} =_T Accuracy(T)$$

This optimized thresholds T_{opt} is subsequently applied to categorize every image in the dataset, providing a systematic and quantitatively justified method for detecting changes within satellite imagery. Through this proposed methodology, we ensure that our classification process remains both reliable and consistent with the overarching goal of identifying significant changes in the observed landscapes.

Following the zero-shot classification based on threshold optimization, we compared the results with those obtained from several tree-based machine learning algorithms which datasets divided by 70% train 30% test data and total amount of image pairs are 746 for LEVIR-CD dataset, 6601

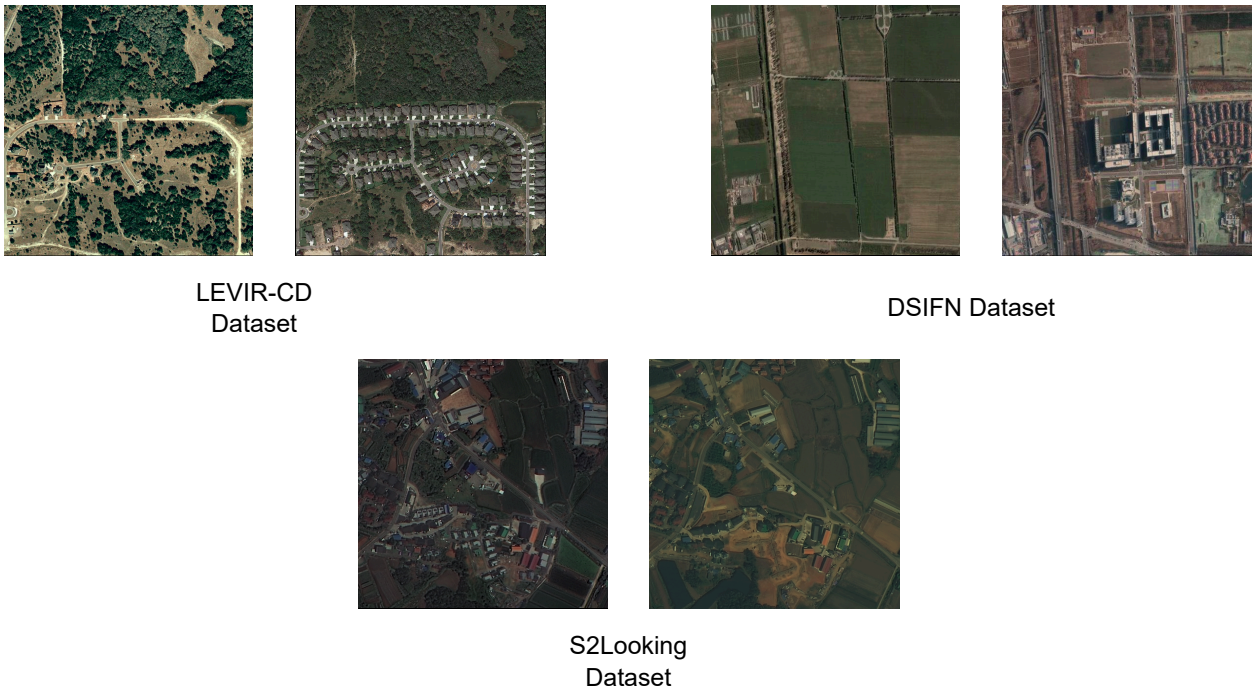


Figure 10: Examples of Image Pairs for Each Datasets LEVIR-CD, DSIFN, S2Looking.

for DSIFN dataset, 6501 for S2Looking dataset with augmented images. This comparative analysis aimed to assess the efficacy of our zero-shot learning approach against traditional supervised learning methods in the context of satellite image classification. The tree-based algorithms selected for this comparison were:

- **Decision Tree:** A basic tree-structured algorithm that splits the data into subsets based on feature values, applying decision rules from the root down to the leaf nodes [Qui86].
- **Random Forest:** A collection of decision trees whose aggregated predictions boost classification accuracy and curb overfitting by training each tree on a different subset of the data [Bre01].
- **Gradient Boosting:** An additive model that sequentially adds weak decision trees to improve the model by focusing on instances that were misclassified in previous rounds [Fri01].
- **XGBoost:** A distributed, optimized gradient boosting library offering an exceptionally efficient implementation of the gradient boosting framework [CG16].

This comparative study was designed to underscore the potential benefits of applying a zero-shot learning approach to detecting changes in satellite imagery, particularly in scenarios where labeled data for certain classes might not be available or are scarce.

We utilized three prominent datasets for evaluating our zero-shot classification methodology for change detection in satellite imagery: LEVIR-CD, DSIFN, and S2Looking. Illiustrative examples of images can be seen Fig 13. Each dataset offers a unique set of challenges and characteristics, making them ideal for assessing the robustness and effectiveness of our approach across different scenarios and environments.

4.2.1. LEVIR-CD

The LEVIR-CD dataset, tailored for building change detection, contains 637 pairs of high-resolution aerial images, each measuring 1024×1024 pixels. CThese images, which were collected from Google Earth, represent both urban and rural environments, featuring a wide range of building structures in various stages of development. This dataset works exceptionally well to test the performance of a model's capacity to detect changes in human-made structures within intricate landscapes [CCL+20].

4.2.2. DSIFN

The DSIFN (Deeply Supervised Image Fusion Network) dataset, though primarily intended for image fusion, presents an ideal platform for the study of change detection. Multi-temporal, multi-spectral, and multi-resolution images within the dataset represent a rich field for investigating zero-shot learning models' performance regarding the identification of subtle changes not necessarily visible from single-temporal or single-spectral data [FLB20].

4.2.3. S2Looking

The S2Looking dataset is a large-scale remote sensing scene change detection dataset with over 5000 pairs of high-resolution images. They span different geographical sites and environmental conditions, such as from urban development to natural disasters. Because of the variety of scenes and changing types, the dataset is an excellent test bed for the generalizability of change detection algorithms in different domains and change scenarios [SWZ+20].

To combat the issue of imbalanced labels in the LEVIR-CD, DSIFN, and S2Looking datasets, we employed a data augmentation strategy to balance the number of images with and without changes. Augmentation was accomplished through a custom Python function that randomly alters each image's scale and rotation. By resizing the images with a scaling factor of between 0.5 and 1.5 and rotating them by random degrees in the range of -90 to 90, we generated diverse versions of the original datasets. This approach not only introduced more balance between the classes but also made the datasets more rich with a higher variety of perspectives and scales, which further challenged and hence made our model more robust and generalizable to a wide range of change detection situations. Datasets in the form of pairs of images and their corresponding masks for segmentation were utilized. Having applied the Contrastive Language–Image Pre-training (CLIP) model on each of the images, the thus resultant probability arrays were noted and stored in the form of CSV files. Manual labeling of the images in these CSV files as '1' in case of change, and '0' in the absence of any change was done next.

They were selected according to their fit for the goals of the study and potential in providing in-depth information about the performance of zero-shot classification algorithms in change detection in satellite images. By using these diverse datasets whose examples are interpretable in image pairs in Fig 13, it is our hope to demonstrate the flexibility and adaptability of our proposed approach in a variety of environments and problems in satellite image analysis.

Table 3. Algorithm Comparison For Training Data

	LEVIR-CD				DSIFN				S2Looking				Average			
	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec
Threshold Optimization	0.9369	0.9439	1.0	0.9496	0.9648	0.9669	1.0	0.9664	0.9507	0.9535	1.0	0.9985	0.9508	0.9547	1.0	0.9715
Decision Tree	0.875	0.8703	0.8468	0.8952	0.9606	0.9609	0.9570	0.9647	0.9374	0.9419	0.9428	0.9410	0.9244	0.9243	0.9155	0.9336
Random Forest	0.9464	0.9473	0.9729	0.9230	0.9757	0.9764	0.9950	0.9586	0.9697	0.9718	0.9695	0.9741	0.9639	0.9651	0.9791	0.9519
Gradient Boosting	0.9508	0.9502	0.9459	0.9545	0.9732	0.9739	0.9870	0.9611	0.9723	0.9741	0.9676	0.9806	0.9654	0.9660	0.9668	0.9654
XGBoost	0.9553	0.9553	0.9639	0.9469	0.9782	0.9788	0.9930	0.9650	0.9743	0.9760	0.9714	0.9807	0.9692	0.97	0.9761	0.9642

Table 4. Algorithm Comparison For Testing Data

	LEVIR-CD				DSIFN				S2Looking				Average			
	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec	Acc	F1	Rec	Prec
Threshold Optimization	0.9375	0.9669	1.0	0.9767	0.9352	0.9665	1.0	0.9642	0.94	0.9690	1.0	0.9960	0.9375	0.9674	1.0	0.9789
Decision Tree	0.8593	0.9203	0.8888	0.9541	0.9294	0.9623	0.9504	0.9746	0.947	0.9727	0.9488	0.9978	0.9119	0.9517	0.9293	0.9755
Random Forest	0.9296	0.9620	0.9743	0.95	0.9617	0.9802	0.9969	0.9640	0.971	0.9852	0.9729	0.9979	0.9541	0.9758	0.9813	0.9706
Gradient Boosting	0.8984	0.9432	0.9230	0.9642	0.9588	0.9785	0.9907	0.9667	0.974	0.9868	0.9759	0.9979	0.9437	0.9695	0.9632	0.9762
XGBoost	0.9296	0.9613	0.9572	0.9655	0.9764	0.9877	0.9969	0.9787	0.978	0.9888	0.9799	0.9979	0.9613	0.9792	0.978	0.9807

Our research systematically analyzes the performance of the zero-shot classification method for change detection from satellite imagery. Applying the Contrastive Language–Image Pre-training (CLIP) model, we tested its effectiveness on three varying datasets: LEVIR-CD[CCL+20], DSIFN[FLB20], and S2Looking[SWZ+20]. Additionally, we provide a comparative analysis with traditional supervised learning algorithms to underscore the zero-shot approach's relative performance.

For a comprehensive perspective, we compared the zero-shot classification's results with those obtained from several tree-based supervised learning algorithms with different metrics as train data in Table 3 as test data in Table 4.

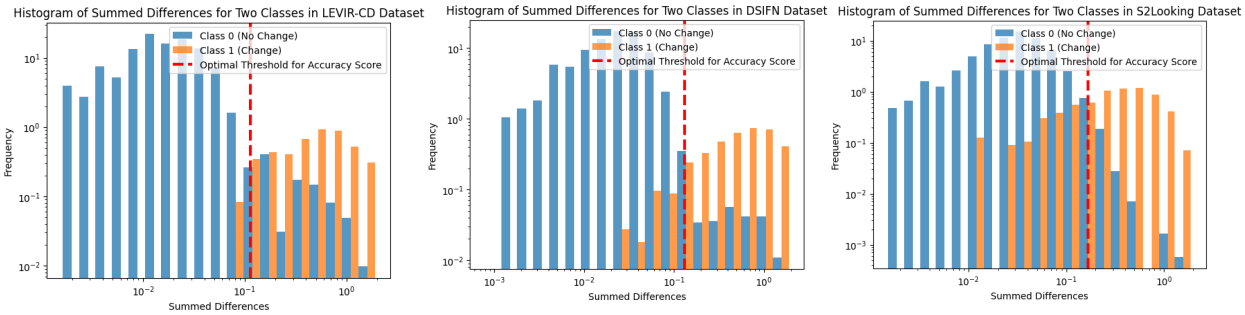


Figure 11: Histogram of Classes for Each Datasets in Log Scale

The histograms in Figure 11 illustrate the frequency distribution of the summed differences for classes 0 and 1 across the LEVIR-CD, DSIFN, and S2Looking datasets, with the optimal threshold for accuracy delineated, showcasing the effectiveness of the zero-shot classification method in differentiating between the change classes.

In Figure 12, we present an illustrative example of change detection using the CLIP model on satellite imagery. The left side shows the original paired images, while the graph on the right quantifies the changes detected with increased probability of *industrial site*.

In our analysis of the performance metrics, it is evident that the zero-shot classification using the CLIP model yields competitive accuracy when benchmarked against traditional algorithms. Particularly, the XGBoost algorithm demonstrates high efficacy, as reflected by its F1 score and precision across all datasets for both training and testing data. However, the zero-shot approach

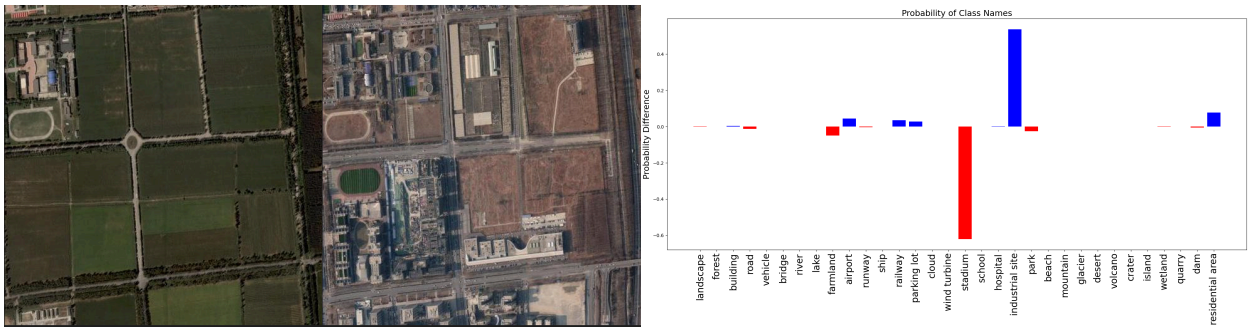


Figure 12: A Change Detection Example for an Image Pair

stands out in its ability to maintain consistent performance without the need for extensive training data, a notable advantage in scenarios where labeled data is scarce or expensive to procure. The histograms further reveal the zero-shot method's robustness, indicating a significant frequency of accurate classifications at the optimal threshold. This threshold optimization appears to be a critical factor in enhancing the model's performance, suggesting that the zero-shot methodology could be finely tuned to achieve greater efficacy in change detection tasks. These findings suggest that while supervised methods continue to be reliable, the zero-shot learning provides a training-free method for remote sensing applications where adaptability and quick deployment are crucial.

4.3. Experiment 3: Change Detection in Satellite Imagery Using Transformer Models and Machine Learning Techniques: A Comprehensive Captioning Dataset

In this work [KP24b], we used the Llama model [TLI+23] to generate captions for each pair of images [YYZ+24b] using CLCD, LEVIR-CD, DSIFN, and S2Looking datasets.

4.3.1. Datasets

This study makes use of four main datasets, as illustrated in Figure 13. The CLCD (Cross-Sensor Land Cover Change Detection) dataset is designed to assess land-cover dynamics captured by multiple sensors, enabling comprehensive multi-temporal and multi-sensor analyses by providing diverse sensor inputs [LCD+22a]. The LEVIR-CD (LEveraging VIdeo for Remote sensing Change Detection) dataset focuses on building change detection with high-resolution aerial image pairs optimized for urban monitoring, effectively identifying structural modifications such as new constructions and demolitions to support detailed urban planning [CS20]. The DSIFN (Dual-Stream Interactive Feature Network) dataset, created for high-resolution change detection, utilizes a dual-stream architecture to enhance feature extraction and interaction between pre- and post-event images, thereby improving sensitivity to subtle changes and boosting detection accuracy [ZYT+20]. Finally, the S2Looking dataset comprises Sentinel-2 satellite imagery and offers a large-scale repository for extensive change detection research, facilitating broad environmental monitoring and analysis [Liu+21]. Each of these datasets contributes uniquely to the development of robust methodologies for automated change detection and caption generation in remote sensing.

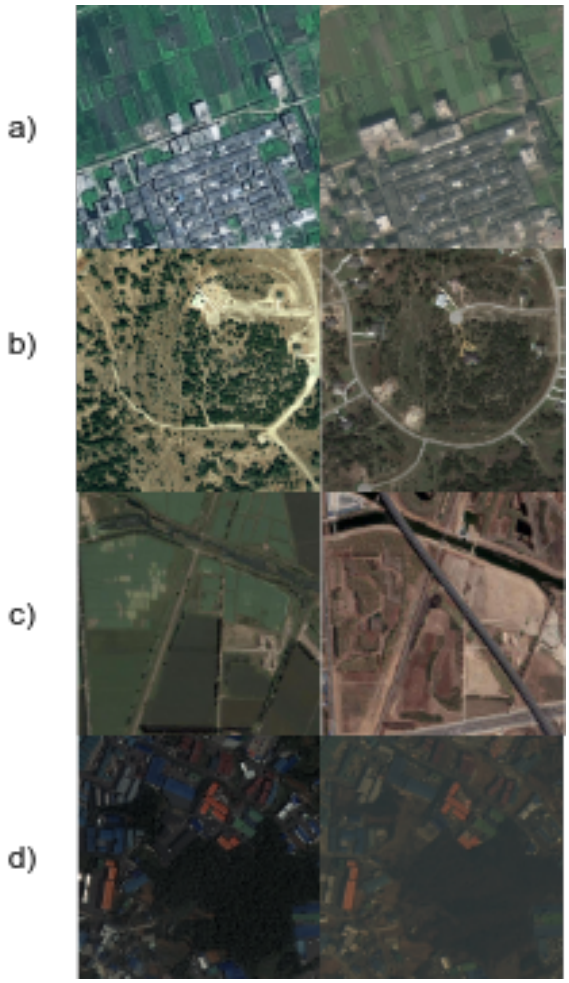


Figure 13: Image pair examples for a) CLCD, b) LEVIR-CD, c) DSIFN, d) S2Looking datasets

4.3.2. Data Augmentation

The datasets exhibited imbalances in terms of the number of image pairs and the types of changes represented. This imbalance could potentially lead to biased model training, where the model becomes proficient at detecting more frequent changes but performs poorly on less represented change types. To mitigate this issue, we performed data augmentation using several techniques aimed at increasing the diversity and quantity of training samples.

First, we applied random rotations to each image pair, rotating them randomly between -90 and 90 degrees. This technique helps the model become invariant to the orientation of the images, thereby improving its ability to generalize across different viewing angles.

Additionally, we used random scaling, where each image pair was randomly scaled between 0.5 and 1.5 times its original size. This approach simulates varying distances from the satellite to the Earth's surface, which can occur due to different orbital positions or satellite types. By training the model on images of varying scales, we improve its resilience to variations in image resolution and size.

These augmentation techniques helped in enhancing an imbalanced dataset via greater variability and volume of training instances, thereby the model strength 14. Using this balanced dataset, the model is able to learn a broader perception of the kind of changes occurring in the satellite images

Figure 14: Data augmentation examples of a) CLCD, b) LEVIR-CD, c) DSIFN, d) S2Looking datasets

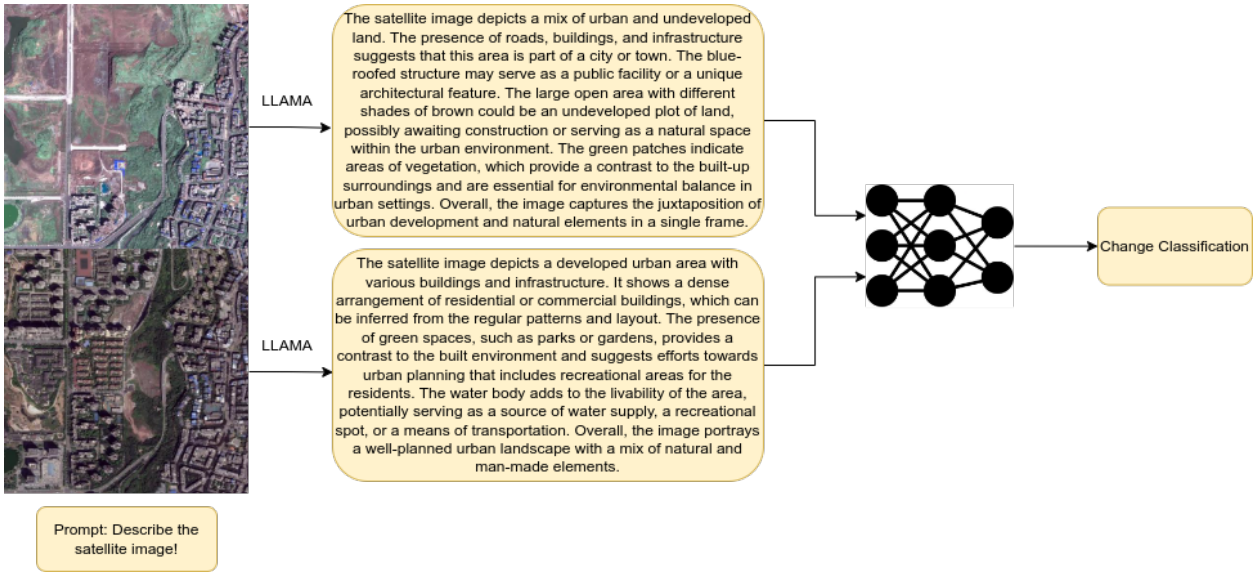


Figure 15: Our methodology for change classification

to identify changes with increased accuracy and reliability along with caption generation.

After applying data augmentation, the total number of images for each dataset is as follows:

- **CLCD:** 710 training image pairs and 220 validation image pairs
- **LEVIR-CD:** 745 training image pairs and 228 validation image pairs
- **DSIFN:** 6600 training image pairs and 550 validation image pairs
- **S2Looking:** 6500 training image pairs and 1800 validation image pairs

4.3.3. Caption Generation Using MiniCPM-V Llama Model

For captioning, we utilized the MiniCPM-Llama3-V-2_5 model [TLI+23][YYZ+24b], which is ideally suited for translating detailed natural language from visual information. With an extremely powerful Transformer-based architecture, the model works best for reading complex visual details and translating them into understandable, contextually accurate text. As such, it is particularly suited to read satellite images, which are likely to have complex details and varied contexts.

The process was to generate captions for each pair of images according to a unique prompt that was appropriate for satellite images. All pairs of satellite images were inputted directly into the Llama model with the prompt "Describe the satellite image!" The prompt asked the model to generate captions that describe the contents of the images in great detail.

Following the initial round of caption generation, the outputs were refined for coherence and clarity. This was necessary to ensure that the descriptions were mapped accurately onto the contents of the images and to eliminate any ambiguities. Refining the captions generated involved ensuring the correctness of the language used, ensuring that key features and changes in the satellite images were accurately captured and described 15.

The model relies on the transformer architecture [VSP+17], the self-attention mechanism of which changed NLP by enabling the network to assign different weights to tokens of the input sequence. This enables the model to produce more fluent, context-dependent text. Formally, this operation is described as:

Given an image I, the model generates a sequence of words $\{w_1, w_2, \ldots, w_n\}$. The probability of each word in the sequence is given by:

$$P(w_t|I, w_1, w_2, \ldots, w_{t-1})$$

where $P(w_t|I, w_1, w_2, \ldots, w_{t-1})$ is the conditional probability of the word w_t given the image I and the previous words in the sequence.

The model is optimized to maximize the probability of the ground-truth word sequence, which can be expressed as:

$$\mathcal{L} = \sum_{t=1}^{n} \log P(w_t|I, w_1, w_2, \ldots, w_{t-1})$$

This training objective compels the model to generate the most probable word sequence based on a given input image and all past tokens in the caption. Additionally, the self-attention mechanism of the Transformer [AP18] allows it to capture long-range dependencies and intricate interrelations within the data that is especially useful for comprehending complex satellite images.

Using this method, the Llama model produces captions with rich, detailed descriptions of satellite images. It aims to create detailed, contextually informed accounts, thereby enhancing the applied usefulness of these images across many domains. Automatically generating precise and informative captions can greatly support efforts in environmental monitoring, disaster response, and urban planning. This automated process not only saves time but also lessens dependence on specialized expertise, making satellite image analysis more accessible and efficient.

To assess how well our model-generated captions are an indicator of real image changes, we applied four traditional machine learning algorithms—Logistic Regression [Cox58], Naive Bayes [Min61], Support Vector Machine (SVM) [HDO+98], and K-Nearest Neighbors (KNN) [CD04]—after transforming the text into a TF–IDF vectorizer [Spa72]. Additionally, we fine-tuned four transformer models—BERT [DCL+19], DistilBERT [SDC+19], RoBERTa [LOG+19], and XLNet [YDY+19]—to forecast whether a change had occurred between the pre-event and post-event images

based only on their generated captions.

Table 5. Accuracy Results for Train Data

Datasets	CLCD	LEVIR-CD	DSIFN	S2Looking
Machine Learning Models				
Logistic Regression	0.8521	0.8644	0.8087	0.8263
Naive Bayes	0.8183	0.7704	0.7348	0.7406
Support Vector Machine	**0.976**	**0.9838**	**0.943**	**0.9443**
K-Nearest Neighbors	0.7169	0.804	0.791	0.7769
Transformer Models				
BERT	**0.7394**	**0.6846**	**0.7970**	**0.7485**
DistilBERT	0.7113	0.6779	0.7848	0.7438
RoBERTa	0.7183	0.6443	0.7856	0.7438
XLNET	0.7042	0.6644	0.7939	0.74

Table 6. Accuracy Results for Validation Data

Datasets	CLCD	LEVIR-CD	DSIFN	S2Looking
Machine Learning Models				
Logistic Regression	**0.7727**	0.7105	0.7762	0.7859
Naive Bayes	0.7545	0.6754	0.7254	0.7181
Support Vector Machine	0.7681	**0.7192**	**0.8084**	**0.7893**
K-Nearest Neighbors	0.609	0.6447	0.705	0.6592
Transformer Models				
BERT	**0.6773**	0.693	0.8169	0.7522
DistilBERT	0.6727	**0.7018**	0.8102	0.7383
RoBERTa	0.6455	0.636	**0.8288**	**0.7628**
XLNET	0.6455	0.6711	0.8	0.7357

The performance of the machine learning models is also varied. The Support Vector Machine (SVM) consistently outperforms the other traditional machine learning models, with the highest accuracy on all of the datasets for the training and validation phases. This suggests that SVM's capability to handle high-dimensional spaces and its effectiveness with small to medium-sized datasets make it a robust choice for change detection tasks. Logistic Regression and Naive Bayes perform reasonably well, but their simpler assumptions about data distribution limit their accuracy compared to more complex models. K-Nearest Neighbors (KNN), as easy to use and interpret, shows lower accuracy, perhaps due to its vulnerability to the choice of 'k' and the method of distance calculation, which might not capture the complexity of the datasets.

Performance variations of transformer-based models arise due to differences in model architectures and pre-training tasks of the models. BERT, since it is pre-trained bidirectionally on masked language modeling, is effective at capturing context information, and that translates into superior performance across other datasets when it is being trained. On the other hand, RoBERTa, which is the optimized version of BERT with increased training and larger batch sizes, appears to generalize better on validation sets, particularly those such as DSIFN and S2Looking which would benefit from such optimization.

DistilBERT, while being smaller and faster, performs competitively but is behind BERT and RoBERTa, suggesting the model size vs. accuracy trade-off. XLNET, using its autoregressive approach, performs well but not consistently better than the other models, suggesting that the bidirectional context of BERT and RoBERTa is more beneficial to change detection tasks.

All in all, machine learning model performance test for the task of change detection illustrates Support Vector Machines' (SVM) superiority amongst the classic models Transformer-based models, particularly BERT and RoBERTa, perform incredibly with their bidirectionally trained and optimized, having RoBERTa show great generalization ability. DistilBERT has a comparative, though speedier, option, though being slightly less precise, while autoregressive nature of XLNET provides good outcomes without continuously topping the bidirectionally trained ones. These insights underscore the need to balance model complexity, speed, and precision according to particular dataset demands.

5. Experiment 4: MiniCPM-V LLaMA Model for Image Recognition: A Case Study on Satellite Datasets

This study compares the performance of the MiniCPM-V approach in satellite image identification, in terms of pattern classification capability using satellite datasets [KP25]. We selected MiniCPM-Llama3-V version 2.5 and tested its capacity to generalize using various satellite imaging types and how it performs when compared to typical deep learning frameworks such as convolutional neural networks. By probing the strengths of MiniCPM-Llama3-V 2.5 here, we aim to advance the edge of large language models for satellite image analysis and to highlight their value in enhancing the accuracy and efficiency of the recognition processes.

The main contributions of this paper are:

a) This research investigates the application of the MiniCPM-V model, which can process visual and textual data, as a new method of satellite image analysis and compares it with current techniques.

b) The MiniCPM-V model is tested using various satellite image datasets (MAI, RSICD, RSSCN7) and an aggregated dataset to examine its ability to generalize and how data variability affects model performance.

c) The work analyzes how well the model can perform with multiple labels within multi-label classification problems, identifying the issues with overlap and class imbalance and solutions that can improve model accuracy.

5.1. Model

The MiniCPM-V model [YYZ+24a] is a smaller, optimized variant of the CPM (Chinese Pre-trained Language Model) family, specifically designed for multi-modal tasks that require the integration of both visual and textual data. Built on transformer-based architecture, MiniCPM-V uses both Vision Transformers (ViTs) [Ale20] and text transformers [VSP+17], allowing it to handle tasks like image captioning, image classification, and visual question answering with natural language prompts.

Table 7. Performance metrics on standard multimodal benchmarks. "RW QA" denotes RealWorldQA, "Obj HalBench (Res./Men.)" refers to the Object Hallucination Benchmark with separate response- and mention-level hallucination rates, and "*" indicates results we obtained using the official checkpoints. The best open-source results are highlighted in **bold** [YYZ+24a].

Model	**Size**	**Open-Compass** [Con23]	**MME** [FCS+23]	**MMB test (en)** [LDZ+24]	**MMB test (cn)** [LDZ+24]	**MMMU val** [YNZ+24]	**Math-Vista** [LBX+24]	**LLaVA Bench** [LLW+23]	**RW QA** [LLL+23; MKJ21; SNS+19]	**Obj HalBench (Res./Men.)** [RHB+18; YYZ+24c]
Proprietary Models										
GPT-4V (2023.11.06)	-	63.5	1711.5	77.0	74.4	53.8	47.8	93.1	63.0	13.6/7.3*
Gemini Pro	-	62.9	2148.9	73.6	74.3	48.9	45.8	79.9	60.4	-
Claude 3 Opus	-	57.7	1586.8	63.3	59.2	54.9	45.8	73.9	48.4	-
Open-source Models										
DeepSeek-VL-1.3B [LLZ+24]	1.7B	46.2	1531.6	66.4	62.9	33.8	29.4	51.1	49.7	16.7/9.6*
Mini-Gemini [LZW+24]	2.2B	-	1653.0	-	-	31.7	-	-	-	-
Yi-VL-6B [AY+24]	6.7B	48.9	1915.1	68.4	66.6	40.3	28.8	51.9	53.5	19.4/11.7*
Qwen-VL-Chat [BBY+23]	9.6B	51.6	1860.0	61.8	56.3	37.0	33.8	67.7	49.3	43.8/20.0*
Yi-VL-34B [AY+24]	34B	52.2	**2050.2**	72.4	70.7	45.1	30.7	62.3	54.8	20.7/14.0*
Phi-3-vision-128k-instruct [AAA+24]	4.2B	-	-	-	-	40.4	44.5	64.2*	58.8*	-
XTuner-Llama-3-8B-v1.1 [XTu23]	8.4B	53.3	1818.0	71.7	63.2	39.2	40.0	69.2	-	-
CogVLM-Chat [WLY+24]	17B	54.2	1736.6	65.8	55.9	37.3	34.7	73.9	60.3	26.4/12.6*
Bunny-Llama-3-8B [HLW+24]	8.4B	54.3	1920.3	77.0	73.9	41.3	31.5	61.2	58.8	-
DeepSeek-VL-7B [LLZ+24]	7.3B	54.6	1765.4	73.8	71.4	38.3	36.8	77.8	54.2	11.4/6.5*
LLaVA-NeXT-Llama3-8B [LZZ+24]	8.4B	-	1971.5	-	-	41.7	-	80.1	60.0	-
Idefics2 [LTC+24]	8.0B	57.2	1847.6	75.7	68.6	45.2	52.2	49.1	60.7	-
Cambrian-8B [TBW+24]	8.3B	58.8	1802.9	74.6	67.9	41.8	47.0	71.0	60.0	-
CogVLM2-19B-Chat [WLY+24]	19B	62.3	1869.5	73.9	69.8	42.6	38.6	83.0	62.9	-
LLaVA-NeXT-Yi-34B [LLL+24]	34B	62.7	2006.5	**81.1**	79.0	48.8	40.4	81.8	66.0	-
Cambrian-34B [TBW+24]	34B	64.9	2049.9	80.4	**79.2**	**50.4**	50.3	82.0	**67.1**	-
MiniCPM-V 1.0 [YYZ+24a]	2.8B	47.5	1650.2	64.1	62.6	38.3	28.9	51.3	51.2	21.6/11.5
MiniCPM-V 2.0 [YYZ+24a]	2.8B	54.5	1808.6	69.1	66.5	38.2	38.7	69.2	55.8	14.5/7.8
MiniCPM-Llama3-V 2.5 [YYZ+24a]	8.5B	**65.1**	2024.6	77.2	74.2	45.8	**54.3**	**86.7**	63.5	**10.3/5.0**

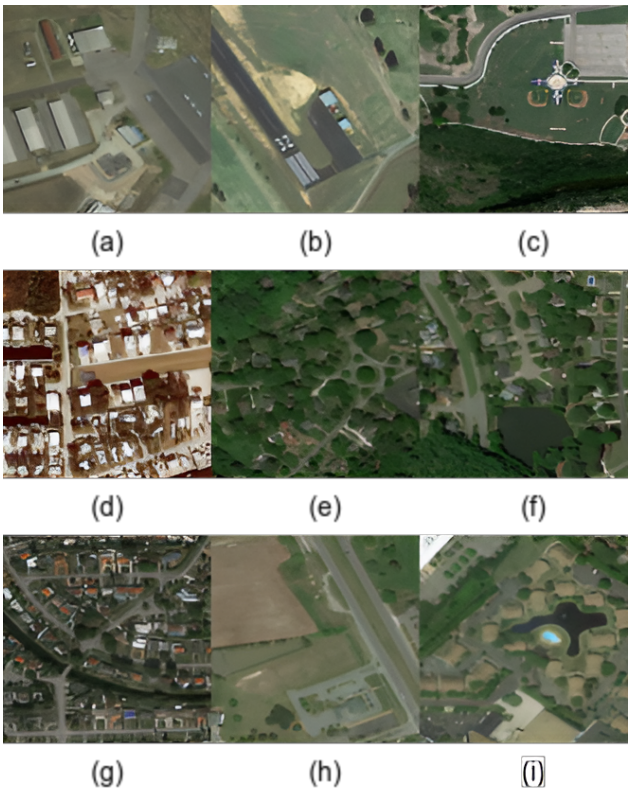


Figure 16: Example images for MAI Dataset (a) ”apron, parking lot, residential, runway”, (b) ”apron, residential, runway”, (c) ”baseball field, parking lot, river, park”, (d) ”residential, runway”, (e) ”residential, roundabout”, (f) ”residential, lake, park”, (g) ”residential, bridge, roundabout”, (h) ”commercial, farmland, residential”, (i) ”commercial, parking lot, residential, lake”

Vision Transformers [Ale20] are used for process visual data by splitting images into patches, embedding each patch into a vector space, and feeding them through transformer layers. This approach enables the model to grasp spatial dependencies and feature representations within images, similar to how transformers handle textual data. Unlike traditional CNNs, ViTs [Ale20] in MiniCPM-V [YYZ+24a] can model long-range dependencies across the image, enhancing its ability to recognize complex patterns.

MiniCPM-V [YYZ+24a] is designed to be more computationally efficient than larger models in the CPM family, with fewer parameters but similar performance. This makes it suitable for use where computational resources are limited, such as real-time image classification or edge device deployment. The model’s parameter-sharing techniques help ensure that model size and accuracy can be balanced well so that the model can suitably be deployed in real-life applications. MiniCPM-V [YYZ+24a] was utilized in satellite image classification in this study. The training data included satellite imagery in conjunction with structured prompts to guide the model in identifying specific patterns or objects in the images. The model could combine visual and textual knowledge, enhancing its precision in classifying diverse landscapes and objects in satellite imagery.

We selected MiniCPM-V due to its performance on handling vision-language tasks at the cost of a good trade-off between performance and computational cost. Unlike larger models, MiniCPM-V is optimized for multi-modal learning that makes it efficient in satellite image recognition without requiring large computational resources [YYZ+24a]. From the various variants of the MiniCPM

model listed in Table 7, MiniCPM-Llama3-V 2.5 is the most appropriate model for our study. This model possesses the highest Open-Compass score (65.1) and performs better on a number of benchmarks, including MME (2024.6), MMB test (en) (77.2), and LLaVA Bench (86.7). It also possesses the lowest hallucination rate in the Object HalBench (10.3/5.0), which indicates high reliability in object recognition tasks. Compared to MiniCPM-V 1.0 and 2.0, the Llama3-V 2.5 variant possesses significantly higher accuracy with relatively small size (8.5B parameters). Thanks to its enhanced multimodal understanding, balanced computational efficacy, and top-notch classification capability, MiniCPM-Llama3-V 2.5 was selected as the top performer among our satellite image recognition experiment. The test metrics in Table 7 were reported from [YYZ+24a], an extensive evaluation of the MiniCPM family and its performance over various multimodal benchmarks.

5.2. Datasets

We employ three prominent datasets for remote sensing image recognition and classification: MAI[10], RSICD[11], and RSSCN7[12]. These collections span a wide variety of aerial scenarios and are extensively utilized in remote sensing and image processing studies.

Table 8. Number of images for each class in the MAI dataset [HML+21].

Class	Number	Class	Number	Class	Number
Residential	2387	Parking Lot	2007	Woodland	1610
Commercial	1610	Farmland	1222	Bridge	878
River	764	Lake	756	Park	638
Sparse Shrub	336	Soccer Field	302	Roundabout	281
Baseball Field	271	Runway	230	Storage Tanks	219
Apron	211	Works	186	Beach	165
Stadium	136	Tennis Court	114	Sea	80
Golf Course	75	Port	10	Train Station	9

Table 9. Number of images for each class in the RSICD dataset [LWZ+17].

Class	Number	Class	Number	Class	Number
Airport	420	Farmland	370	Playground	1031
Bare Land	310	Forest	250	Pond	420
Baseball Field	276	Industrial	390	Viaduct	420
Beach	400	Meadow	280	Port	389
Bridge	459	Medium Residential	290	Railway Station	260
Center	260	Mountain	340	Resort	290
Church	240	Park	350	River	410
Commercial	350	School	300	Sparse Residential	300
Dense Residential	410	Square	330	Storage Tanks	396
Desert	300	Parking	390	Stadium	290

The MAI (Multi-Scene Aerial Image Dataset) is a dataset [HML+21] designed for understanding aerial scenes, particularly focusing on recognizing various scene types. The dataset contains a wide variety of images aimed at multi-scene recognition tasks. It was introduced in the context of

[10] https://github.com/Hua-YS/Prototype-based-Memory-Network
[11] https://github.com/201528014227051/RSICD_optimal
[12] https://github.com/palewithout/RSSCN7

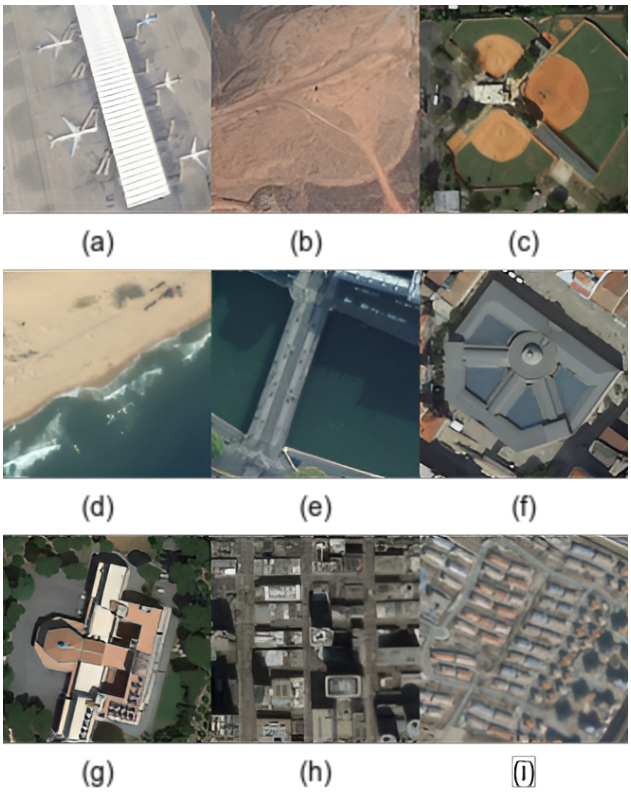


Figure 17: Example images for RSICD Dataset (a) airport, (b) bare land, (c) baseball field, (d) beach, (e) bridge, (f) center, (g) church, (h) commercial, (i) dense residential

prototype-based memory networks for scene classification, demonstrating the effectiveness of this method in identifying different scene types in aerial images. To accelerate advancements in aerial scene interpretation under real-world conditions, we introduce the MAI dataset, comprising 3923 large-scale images sourced from Google Earth covering regions in the United States, Germany, and France. Each image measures 512×512 pixels, with spatial resolutions ranging from 0.3 m/pixel to 0.6 m/pixel. The dataset encompasses 24 categories—apron, baseball field, beach, commercial area, farmland, woodland, parking lot, port, residential zone, river, storage tanks, sea, bridge, lake, park, roundabout, soccer field, stadium, train station, industrial site, golf course, runway, sparse shrubland, and tennis court (see Fig.16 and Table8).

The RSICD (Remote Sensing Image Captioning Dataset) [LWZ+17] was specifically assembled to train models that generate natural-language descriptions of remote sensing imagery. It comprises a diverse collection of images used to teach captioning systems how to produce accurate annotations. In total, RSICD contains 24333 sentences with a combined vocabulary of 3323 words, making it a key resource for advancing image-captioning methods in the remote sensing field. The dataset is organized into 30 categories—airport, bare land, baseball field, beach, bridge, central region, church, commercial district, dense residential area, desert, farmland, forest, industrial zone, meadow, medium-density residential area, mountain, park, school, square, parking lot, playground, pond, viaduct, port, railway station, resort, river, sparse residential area, storage tanks, and stadium (see Fig.17 and Table9).

The RSSCN7 dataset [ZNZ+15] is a widely adopted benchmark for deep-learning–driven scene

classification in remote sensing. It comprises 2 800 high-resolution images distributed across seven representative land-cover categories—grassland, forest, farmland, parking lot, residential area, industrial zone, and river/lake. For each category, 400 images were harvested from Google Earth and organized into four distinct scale levels, with 100 images per scale. All images measure 400×400 pixels. See Fig. 18.

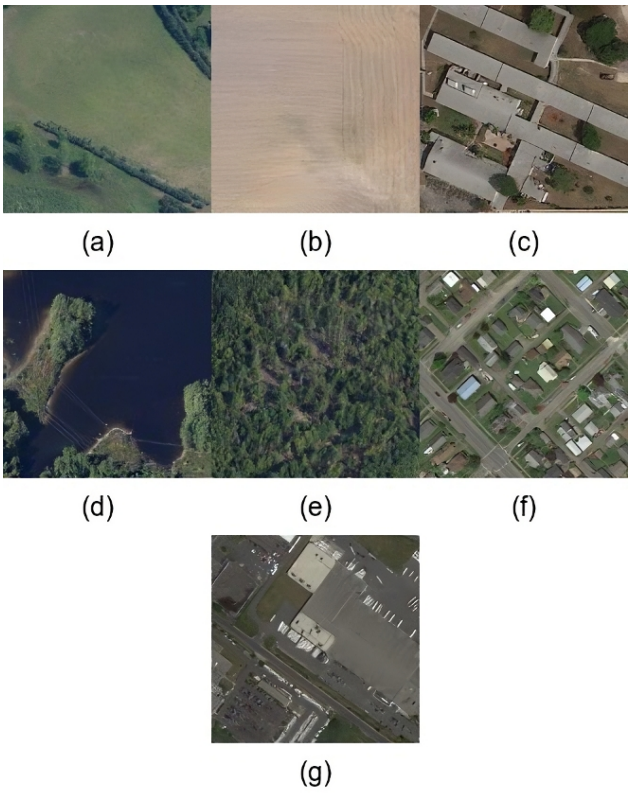


Figure 18: Example images for RSSCN7 Dataset (a) grass, (b) field, (c) industry, (d) river/lake, (e) forest, (f) resident, (g) parking

To ensure consistency and fairness across the merged dataset, labels from the MAI, RSICD, and RSSCN7 datasets were systematically standardized by grouping synonymous or overlapping categories under unified labels. For example, ”dense residential,” ”medium residential,” ”sparse residential,” and ”resident” were consolidated into a single category named ”Residential.” Similarly, ”soccer field” and ”playground” were unified under the label ”Football field,” while ”woodland” from MAI and ”forest” from RSICD and RSSCN7 were combined as ”Forest.” ”Parking lot” and ”parking” were merged into ”Parking,” and ”train station” was standardized as ”Railway station” to align with other datasets. Additionally, ambiguous or inconsistently represented labels such as ”grass,” ”meadow,” and ”river/lake” were excluded due to their variability across datasets, which could introduce noise and affect the reliability of the results. This careful standardization ensured the merged dataset retained its diversity while providing a clear and consistent framework for model evaluation.

5.3. Methodology

Each meta-data of the datasets was first converted into CSV file format. For each dataset, we created a CSV file containing the image names and their corresponding labels. This format made it easy to process and manage the image data during the classification processes. This preprocessing allowed the datasets to be in a common format, making it easy to integrate into the classification pipeline.

For our classification model, we utilized the MiniCPM-V model [YYZ+24a]. We ran the model on the Google Colab environment with an L4 GPU. For input to the model, we used both the images and particular prompts that were intended to guide the classification action. These prompts provided context and additional guidance to enable the model to more accurately classify each image. By using a prompt-based, ordered approach, we ensured the model was provided with explicit instructions for object recognition to reduce classification vagueness. Data and reporduceable code can be found in kaggle.com/datasets/kursatkomurcu/minicpm-v-satellite-object-recognition/data.

Using this approach, we leveraged the MiniCPM-V model to classify each image in the datasets and subsequently collected the classification results. After this process, we merged three datasets and converted synonymous labels into unified category names to maintain consistency. Specifically, the label *residential* was used in place of *dense residential, medium residential, sparse residential, and resident*; *football field* replaced *soccer field and playground*; *forest* was substituted for *woodland*; *parking* replaced *parking lot*; *industrial* replaced *works*; *railway station* replaced *train station*; and *lake* was used instead of *pond*. Moreover, we did not include *grass* and *river/lake* labels in the MAI dataset [HML+21] in our merged dataset due to the unclear nature of their images (see Fig. 19).

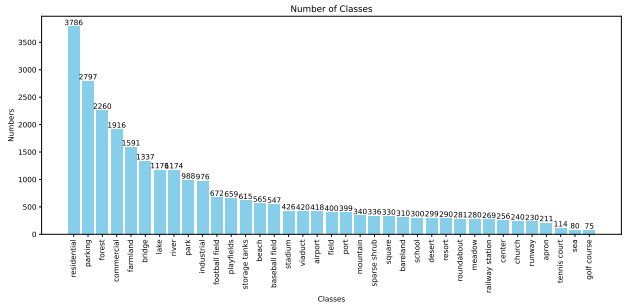


Figure 19: Distribution of the classes in the merged dataset

The MiniCPM-V classification process follows a structured workflow, outlined in Algorithm 1. The methodology is designed to preprocess data, apply classification, and collect results systematically.

To ensure the MiniCPM-V model effectively classifies images, we used a set of structured prompts tailored to each dataset:

1. Identify the following categories in the satellite image and list them in a comma-separated format enclosed in double quotation marks: ”apron, baseball field, beach, commercial, farmland, woodland, parking lot, port, residential, river, storage tanks, sea, golf course, runway, sparse shrub, tennis court, bridge, lake, park, roundabout, soccer field, stadium, train station,

Algorithm 1 MiniCPM-V Image Classification Pipeline

Input: Satellite image dataset D, MiniCPM-V model M, Prompts P
Output: Classified image labels L
Convert dataset D metadata into CSV format
for each image I in D **do**
    Generate classification prompt P_I for I
    Feed image I and prompt P_I into model M
    Collect classification result L_I
end for
Merge dataset results and standardize labels
Return classified labels L =0

works". Return the identified categories in double quotation marks without any explanations or additional text.

2. Identify the object in this satellite image. Respond with only one word from the following list: airport, bareland, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, meadow, medium residential, mountain, park, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, viaduct. Do not use any other words or phrases.

3. What do you see in this satellite image? Do not answer more than one word. Reply with only one word from these options: green areas, field, industry, river/lake, forest, resident, parking.

4. Identify the satellite image from the following list: field, industrial, river, lake, forest, residential, parking, airport, bareland, baseball field, beach, bridge, center, church, commercial, desert, farmland, meadow, mountain, park, playfields, port, railway station, resort, school, square, stadium, storage tanks, viaduct, apron, sea, golf course, runway, sparse shrub, tennis court, roundabout, football field. Note: Do not include any other words than the list and do not include any other additional information in your response. The image may contain one or multiple objects. Decide whether the image contains one or more objects and please list only the names of the detected object(s), if there are more than one objects and separate them by commas.

The MiniCPM-V model's performance was optimized by adjusting its hyperparameters to better align with the characteristics of the datasets. Two key hyperparameters, *temperature* and *sampling*, were carefully selected to control the model's output behavior, ensuring a balance between prediction diversity and determinism.

For the MAI, RSICD, and RSSCN7 datasets, the temperature was set to 0.7. This parameter allowed the model to explore other possibilities of predictions while retaining the most probable labels, a key feature for multi-label classification problems. On the combined dataset, however, a lower temperature of 0.1 was applied. This stricter setup ensured deterministic and stable predictions across the diverse and intricate class distributions in the combined dataset.

The parameter of sampling was made to influence the variability of the model's prediction. For the MAI, RSICD, and RSSCN7 datasets, sampling was enabled to encourage a broader search for

potential labels, particularly useful for cases that involve multi-label. Conversely, for the combined dataset, sampling was disabled to prioritize the most probable outputs, thereby constraining noise and enhancing reliability in prediction.

These adjustments were guided by empirical data to ensure the optimal possible match between the model's prediction and the unique needs of each dataset, without fine-tuning the internal weights of the model. By prioritizing hyperparameter choice and normalization of datasets, the MiniCPM-V model was effectively adjusted for satellite image recognition tasks with varying data and label complexities.. The inclusion of multi-label datasets enabled the proper testing of the capacity of the model to address complex object recognition tasks, as well as demonstrating the superiority of large language models in satellite image classification.

5.4. Results

Our study systematically evaluates the performance of the MiniCPM-V [YYZ+24a] model for satellite image recognition based purely on language information from vision-language model (see Table 10). We explored its efficacy across three distinct datasets: MAI [HML+21], RSICD [LWZ+17], RSSCN7 [ZNZ+15] and the merged dataset.

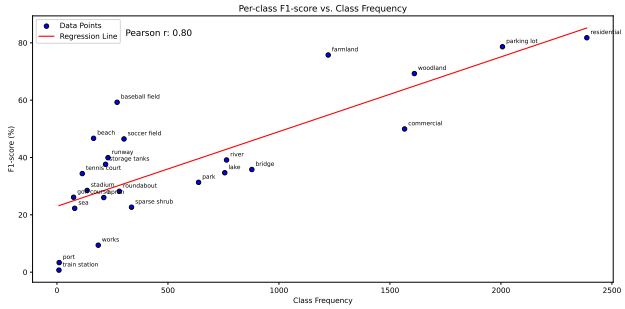


Figure 20: MAI Dataset F1 Scores-Class Frequency Graphics

For the MAI dataset, the model achieved a notably low Top-1 accuracy of 0.0701, despite moderate precision 0.485 and recall 0.6116, culminating in an F1 score of 0.541. This indicates that while the model identified a significant portion of relevant instances (high recall), it struggled to correctly predict the majority class labels, as evidenced by the low accuracy. The Top-5 accuracy, however, was significantly higher at 0.9783, showing that the true labels were often present within the top five predictions. Despite this, the Top-5 precision and recall dropped to 0.618 and 0.5, respectively, resulting in an F1 score of 0.11. This suggests that while the model could identify relevant predictions, ranking these predictions accurately remains a challenge.

In contrast, on the RSICD dataset the model achieved a Top-1 accuracy of 0.6219, with precision at 0.6784, recall at 0.5836, and an F1 score of 0.5575. Although the Top-5 accuracy remained 0.6219, precision fell to 0.311 and recall to 0.500, producing an F1 of 0.3834. These findings suggest that, while the model ranks its single best predictions quite well, it is less effective at capitalizing on the extra candidates in a Top-5 evaluation.

The RSSCN7 dataset demonstrated the best performance across the board, with a Top-1 accuracy of 0.7057 and corresponding precision 0.4257, recall 0.4117, and F1 score 0.4084.

When evaluating the model on the merged dataset, which combines the challenges of all individual datasets, the Top-1 accuracy was moderate at 0.4349, with precision 0.6026, recall 0.4551, and an F1 score 0.5186. In the Top-5 setting, the accuracy improved to 0.7023, although precision 0.1547, recall 0.5, and the F1 score 0.2363 showed declines. These results suggest that while the model is capable of identifying relevant predictions in a broader set of data, achieving high precision across multiple labels remains challenging.

For the MAI and merged datasets, the lower Top-1 accuracy scores are due to their multi-labeled nature—the MAI dataset is entirely multi-labeled, while the merged dataset is partially so. In multi-label classification, accuracy reflects the model's ability to correctly predict each individual label, which inherently complicates the task. However, the significantly higher Top-5 accuracy values demonstrate that the model performs considerably better when evaluated with leniency in ranking predictions (see Tables 11 and 14).

Table 10. Top-1/Top-5 Overall Metrics of the Datasets.

Dataset	Accuracy	Precision	Recall	F1
MAI	0.0701/0.9783	0.4850/0.6180	0.6116/0.5000	0.5410/0.1100
RSICD	0.6219	0.6784	0.5836	0.5575
RSSCN7	0.7057	0.4257	0.4117	0.4084
Merged	0.4349	0.6026	0.4551/	0.5186

5.4.1. Results for MAI Dataset

In the study by [HML+21], the authors employed various machine learning and deep learning models on the entire MAI dataset, which comprises 100,000 images. They achieved maximum overall precision, recall, and F1 scores of 0.801, 0.665, and 0.713, respectively. Additionally, they reported only average precision results for each class. In contrast, our study utilized a subset of 3,923 images from the MAI dataset. Although our results are lower than those reported in [HML+21] (see Tables 10 11). Furthermore, we observed a positive correlation with 0.8 p-value between class frequencies and F1 scores (see Fig. 20).

As shown in Table 11, the accuracy, precision, recall, and F1 scores for each class vary significantly. For example, the "Parking Lot" and "Residential" classes exhibit relatively high F1 scores, indicating that the model performs well in these categories. On the other hand, classes such as "Port," "Train Station," and "Works" have very low F1 scores. This could be due to the under representation of these classes in the dataset, potential confusion with other similar classes, or the model's difficulty in distinguishing specific features of these classes. Additionally, while the model demonstrates acceptable performance for certain classes like "Commercial," "Woodland," and "Farmland," lower recall and precision rates were observed for some classes. These findings highlight the impact of data imbalance and overlap between classes on the overall model performance. Therefore, creating a more balanced dataset or employing techniques such as data augmentation could potentially improve the model's performance. Furthermore, the prediction outcomes for sample images from the MAI dataset can be found in Fig. 29, which demonstrates the model's performance visually, highlighting both false positives and false negatives.

Table 11. Top-1 results for MAI dataset.

Class	Accuracy	Precision	Recall	F1
Apron	0.7204	0.1517	0.9147	0.2603
Baseball Field	0.9177	0.4502	0.8672	0.5927
Beach	0.9342	0.3542	0.6848	0.4669
Bridge	0.7798	0.5150	0.2745	0.3581
Commercial	0.6834	0.6769	0.3959	0.4996
Farmland	0.8218	0.6574	0.8936	0.7575
Golf Course	0.9207	0.1590	0.7333	0.2613
Lake	0.8022	0.4769	0.2725	0.3468
Park	0.8101	0.3803	0.2665	0.3134
Parking Lot	0.7693	0.7471	0.8301	0.7864
Port	0.9115	0.0172	0.6000	0.0334
Residential	0.7859	0.8480	0.7897	0.8178
River	0.7930	0.4579	0.3416	0.3913
Roundabout	0.6821	0.1683	0.8719	0.2821
Runway	0.9026	0.3128	0.5522	0.3994
Sea	0.9217	0.1397	0.5500	0.2228
Soccer Field	0.9014	0.3990	0.5563	0.4647
Sparse Shrub	0.8399	0.1933	0.2738	0.2266
Stadium	0.9118	0.1983	0.5074	0.2851
Storage Tanks	0.8792	0.2643	0.6530	0.3763
Tennis Court	0.9212	0.2269	0.7105	0.3439
Train Station	0.9286	0.0037	0.1111	0.0071
Woodland	0.7507	0.7008	0.6851	0.6928
Works	0.8871	0.0759	0.1237	0.0741

5.4.2. Results for RSICD Dataset

Table 12 presents the recall performance of various models on the RSICD dataset. Among the evaluated models, RemoteCLIP [LCG+24] achieved a recall of 0.3635, while GeoRSCLIP-FT [ZZG+24] slightly improved this metric to 0.3887. The AMFMN model [YZF+22], however, demonstrated a significantly lower recall of 0.1553. In contrast, our proposed LLaMA MiniCPM-V model attained a substantially higher recall of 0.5836, outperforming all compared models by a considerable margin.

Table 12. Comparison for RSICD dataset.

Model	Recall	Reference
RemoteCLIP	0.3635	[LCG+24]
GeoRSCLIP-FT	0.3887	[ZZG+24]
AMFMN	0.1553	[YZF+22]
HarMA	0.3895	[Hua24]
MiniCPM-Llama3-V 2.5	**0.5836**	[YYZ+24a] (Our Experiment)

Moreover, our model achieved high metric values for each class (see Fig. 21 and Table 16). This indicates that the MiniCPM-V model not only excels in overall recall but also maintains strong performance across individual classes. The superior recall of the LLaMA MiniCPM-V model highlights its enhanced capability to identify relevant instances within the RSICD dataset, which may be attributed to its advanced feature extraction and classification mechanisms. These results suggest that MiniCPM-V is highly effective in capturing the diverse and complex patterns present in satellite imagery, leading to improved identification and classification of relevant classes. The con-

sistent high performance across classes underscores the potential of the MiniCPM-V model for applications requiring robust and reliable satellite image recognition.

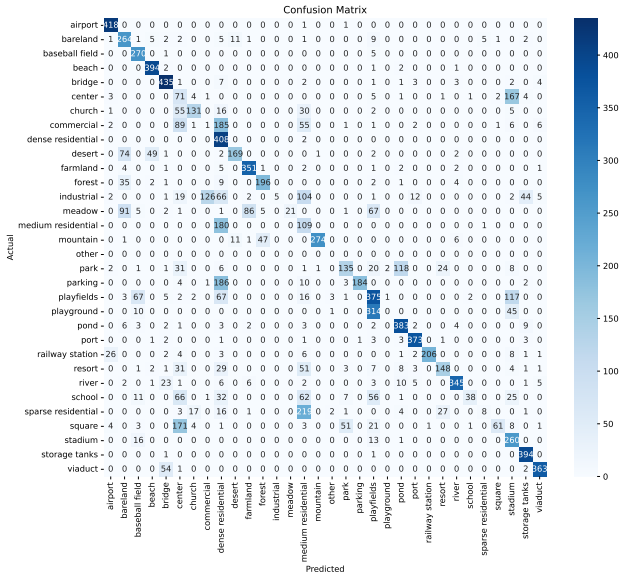


Figure 21: Confusion Matrix of RSICD Dataset Classification

5.4.3. Results for RSSCN7 Dataset

Table 14 demonstrates that our experimental model achieves a lower accuracy compared to other approaches. Notably, the LLaMA MINICPM-V model [YYZ+24a] incorrectly classified many instances of the **green areas** class, while exhibiting high accuracy for the remaining classes (see Table 15 and Fig. 22).

Despite the overall lower accuracy of the MiniCPM-V model on the RSSCN7 dataset compared to other approaches, the performance across most classes remains relatively strong. The exceptionally low performance on the **green areas** class suggests that the model struggles with distinguishing this class from others, potentially due to high visual similarity with classes like **field** and **forest**. This confusion may arise from overlapping features and textures that make it challenging for the model to accurately differentiate between these categories. Overall, while the MiniCPM-V model shows promise in handling several classes within the RSSCN7 dataset, targeted improvements are necessary to enhance its performance on more ambiguous or visually similar categories.

5.5. Results for the Merged Dataset

The evaluation of the MiniCPM-V model on the merged dataset, which integrates the MAI, RSICD, and RSSCN7 datasets, provides a comprehensive assessment of the model's ability to generalize across diverse satellite image distributions and class heterogeneities (see Table 16 and Fig. 19). The merged dataset presents a more complex classification task due to the varied characteristics and label distributions inherited from the individual datasets.

Table 13. Results for RSICD Dataset

Class	Accuracy	Precision	Recall	F1
Airport	0.9961	0.9107	0.9952	0.9511
Bare Land	0.9760	0.5500	0.8516	0.6684
Baseball Field	0.9886	0.6959	0.9783	0.8133
Beach	0.9940	0.8678	0.9850	0.9227
Bridge	0.9885	0.8101	0.9477	0.8735
Center	0.9385	0.1284	0.2731	0.1747
Church	0.9875	0.8239	0.5458	0.6566
Commercial	0.9562	0.0077	0.0029	0.0042
Dense Residential	0.9242	0.3306	0.9951	0.4964
Desert	0.9860	0.8848	0.5633	0.6884
Farmland	0.9892	0.7800	0.9486	0.8561
Forest	0.9902	0.7871	0.7840	0.7856
Industrial	0.9648	1.0000	0.0129	0.0254
Meadow	0.9763	1.0000	0.0750	0.1395
Medium Residential	0.9311	0.1603	0.3759	0.2247
Mountain	0.9938	0.9928	0.8059	0.8896
Park	0.9740	0.6618	0.3857	0.4874
Parking	0.9810	0.9946	0.4718	0.6400
Playfields	0.9246	0.4112	0.5673	0.4768
Playground	0.9658	0.0000	0.0000	0.0000
Pond	0.9826	0.7146	0.9119	0.8013
Port	0.9959	0.9279	0.9589	0.9431
Railway Station	0.9950	0.9952	0.7923	0.8822
Resort	0.9821	0.7363	0.5103	0.6029
River	0.9920	0.9401	0.8415	0.8880
School	0.9757	0.9048	0.1271	0.2229
Sparse Residential	0.9728	0.5714	0.0268	0.0511
Square	0.9750	0.9385	0.1848	0.3089
Stadium	0.9609	0.3957	0.8966	0.5491
Storage Tanks	0.9934	0.8491	0.9949	0.9163
Viaduct	0.9926	0.9380	0.8643	0.8996

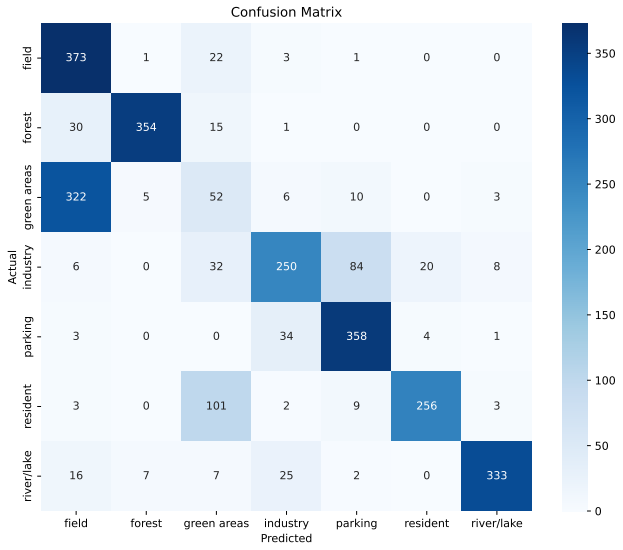


Figure 22: Confusion Matrix of RSSCN7 Dataset Classification

Across the combined dataset, the model attained a modest accuracy of 0.4349 noticeably below its performance on the separate datasets. This drop likely stems from the greater heterogeneity and complexity introduced by merging the data, which strains the model's generalization ability.

Table 14. Comparison for RSSCN7 dataset.

Model	Accuracy	Reference
DBN	0.7700	[ZNZ+15]
GLNet	**0.9507**	[SLZ+21]
CPM	0.5000	[WRK+16]
MiniCPM-Llama3-V 2.5	0.7057	[YYZ+24a] (Our Experiment)

Table 15. Results for RSSCN7 dataset.

Class	Accuracy	Precision	Recall	F1
Green Areas	0.1300	0.1250	0.0162	0.0288
Field	**0.9325**	0.2000	0.1865	0.1930
Industry	0.6250	0.1667	0.1042	0.1282
River/Lake	**0.8325**	0.1250	0.1041	0.1136
Forest	0.8850	0.2500	0.2212	0.2347
Resident	0.6400	0.1250	0.0800	0.0976
Parking	**0.8950**	0.2000	0.1790	0.1889

Despite this, certain classes within the merged dataset exhibit high accuracy scores, indicating that the model can effectively recognize specific types of satellite imagery when sufficient distinguishing features are present.

For instance, classes such as **Airport** (Accuracy: 0.9899), **Bridge** (0.9463), **Mountain** (0.9953), and **Port** (0.996) demonstrate exceptionally high accuracy, precision, and recall values. These results suggest that the model performs well on classes with distinct and consistent visual patterns. Conversely, classes like **Apron** (Accuracy: 0.9873), **Bare Land** (0.9832), and **Commercial** (0.8913) show lower performance metrics, highlighting areas where the model struggles, possibly due to overlapping features with other classes or insufficient training samples.

The F1 scores across classes reveal a similar trend, with higher scores for well-defined classes and lower scores for more ambiguous or less represented classes. For example, the **Residential** class achieved an F1 score of 0.7417, while the **Sparse Shrub** class only reached 0.0435. These disparities indicate that while the model is capable of handling certain categories effectively, it faces challenges with classes that have subtle distinctions or limited representation in the merged dataset.

Figure 23 illustrates the relationship between class frequency and accuracy scores, demonstrating a positive correlation where classes with higher frequency tend to achieve better accuracy.

The combined dataset results also reflect the inherent difficulty of multi-label classification, where the model must predict more than one class for each image simultaneously. The task difficulty increases with the diversity of the combined dataset because the model must contend with a broader range of visual patterns and class overlaps. Regardless of these issues, the MiniCPM-V model shows strong performance on a range of primary classes, highlighting the potential for tasks involving holistic satellite image recognition from diverse environments.

Overall, the findings based on the combined dataset emphasize both the strengths and weaknesses of the MiniCPM-V model. While it excels at correctly predicting some well-defined and adequately represented classes, performance is compromised in more heterogeneous and less represented classes. These observations emphasize the need for model optimization and potentially larger training datasets to generalize well across representative and heterogeneous satellite image

datasets (see Fig. 30).

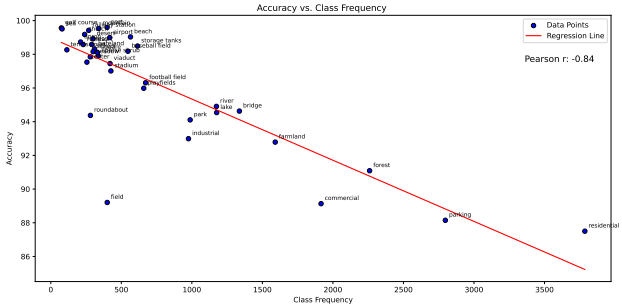


Figure 23: The merged dataset Accuracy Scores-Class Frequency Graphics

5.6. Conclusion

This paper offers a comprehensive evaluation of the performance of the MiniCPM-V model on satellite image recognition for a range of datasets. The results show that MiniCPM-V performs remarkably well in the situation of clear and well-separated classes for the RSSCN7 and RSICD datasets with high accuracy and well-balanced precision and recall. However, the model's performance deteriorates for the more complex and heterogeneous datasets, such as the MAI and the merged dataset, where multi-label classification and class variation are severe challenges.

The results highlight the necessity for further improving MiniCPM-V so that it can be more capable in terms of generalization in different and multi-labeled settings. Future work must focus on applying sophisticated techniques such as data augmentation, transfer learning, and utilizing balancing class imbalance techniques. Additionally, model architecture fine-tuning for it to perform better on multi-label classification and with greater training data could make it more robust and accurate across various satellite image databases.

In conclusion, while MiniCPM-V is good for some scenarios, persistent and high performance across all datasets tested will need to be tackled through targeted improvement. Addressing the identified limitations will pave the way for MiniCPM-V to be a more versatile and reliable tool in the field of remote sensing and satellite image analysis.

Table 16. Results for the Merged Dataset

Class	Accuracy	Precision	Recall	F1
Airport	0.9899	0.7366	0.9234	0.8195
Apron	0.9873	0.4805	0.1754	0.2569
Bare Land	0.9832	0.7077	0.1484	0.2453
Baseball Field	0.9918	0.6790	0.8355	0.7492
Beach	0.9903	0.9052	0.7947	0.8464
Bridge	0.9463	0.8859	0.3717	0.5237
Center	0.9753	0.2130	0.2305	0.2214
Church	0.9918	0.7429	0.6500	0.6933
Commercial	0.8913	0.6239	0.1143	0.1932
Desert	0.9892	0.7489	0.5886	0.6592
Farmland	0.9279	0.8689	0.2791	0.4225
Field	0.8921	0.1694	0.9075	0.2855
Football Field	0.9631	0.5314	0.6429	0.5818
Forest	0.9109	0.8287	0.4239	0.5609
Golf Course	0.9957	0.5114	0.6000	0.5521
Industrial	0.9299	0.4258	0.5994	0.4979
Lake	0.9454	0.6904	0.3963	0.5035
Meadow	0.9785	0.0000	0.0000	0.0000
Mountain	0.9953	0.9888	0.7765	0.8699
Park	0.9411	0.4941	0.1700	0.2530
Parking	0.8815	0.7970	0.3847	0.5189
Playfields	0.9598	0.3396	0.0273	0.0506
Port	0.9960	0.9299	0.8972	0.9133
Railway Station	0.9942	0.8981	0.7212	0.8000
Residential	0.8750	0.6929	0.7979	0.7417
Resort	0.9859	0.7766	0.2517	0.3802
River	0.9491	0.7300	0.4284	0.5400
Roundabout	0.9437	0.1267	0.4021	0.1927
Runway	0.9858	0.4702	0.3087	0.3727
School	0.9816	0.4583	0.1833	0.2619
Sea	0.9951	0.3846	0.0625	0.1075
Sparse Shrub	0.9979	0.2500	0.0238	0.0435
Square	0.9811	0.7727	0.0515	0.0966
Stadium	0.9701	0.4275	0.5329	0.4744
Storage Tanks	0.9849	0.8399	0.7252	0.7784
Tennis Court	0.9827	0.1993	0.5175	0.2878
Viaduct	0.9745	0.2000	0.7100	0.0138

6. Experiment 5: Multispectral Caption Data Unification Using Diffusion and Cycle GAN Models

The great breakthroughs in geo-spacial field in recent years have been driven by computer vision application on RGB images [SRF+24; VNS23; YLL+23; ZYZ+24], while share amount of labeled datasets on multispectral data like Sentinel-2 is very limited. The geo-spatial field have unrelated dataset like object detection or segmentation on RGB images, or captions of RGB images, and largest open-source satellite multispectral data sources like Sentinel-2 [DDC+12] is remain unannotated for the most cases.

Generative models, particularly diffusion-based approaches and generative adversarial networks (GANs), have demonstrated remarkable capabilities in image synthesis and transformation tasks [LGZ+24]. However, generating high-quality multispectral satellite imagery from textual descriptions remains a significant challenge due to the complex spectral characteristics inherent in remote sensing data [ZZ22].

Existing research has explored the use of CycleGAN for image-to-image translation tasks in remote sensing and geospatial applications [RLC+24]. Additionally, diffusion models have recently gained prominence in hyperspectral image synthesis, demonstrating their potential to generate high-fidelity data with improved spatial and spectral consistency [LCC+23]. However, the integration of diffusion models with CycleGAN for multispectral caption-based image generation remains underexplored.

We addressing the problem of unifying existing existing datasets of captions, RGB images and Sentinel-2 like multispectral data (later we will call it triplet). In this study, we propose a novel methodology, which aims to integrate caption-based synthetic image generation with multispectral image translation techniques. By combining Stable Diffusion models with CycleGAN-based models, we create an pipeline that enables the generation of realistic synthetic of missing data of the triple. This approach allows to generate missing data of the triple and propose unified process to combine datasets which till now was uncombinable.

This experiment contributes to the field by:

- Proposing a pipeline for unifying datasets of captions, RGB images and multispectral Sentinel-2 images.
- Delivering the Stable Diffusion 2-1 Base model to improve text-to-image generation, as well as creating CycleGAN [ZPI+17] to convert RGB images into multispectral Sentinel-2 images, and vice versa.
- Creating new an artifical multispectral satellite image-caption dataset.

By unifying these techniques, our study aims to enhance the quality and applicability of generated remote sensing imagery. This approach has allowing to expand accessible data and successfully improve remote sensing applications like land cover classification, disaster monitoring, and environmental change detection [LJG+24].

6.1. Related Work

Generative models have significantly contributed to computer vision [GPM+14; HJA20] and remote sensing[LZX+20; XYJ+23], particularly in image synthesis and transformation. Among these, CycleGAN has demonstrated effectiveness in unpaired image-to-image translation tasks [ZPI+17]. It has been widely used in geospatial analysis and remote sensing, where it has been leveraged for domain adaptation and multispectral image synthesis [RLC+24]. Recent studies have improved CycleGAN's ability to preserve spectral details and enhance image translation accuracy by incorporating additional geospatial derivatives such as NDVI and digital surface models [RZD+19].

CycleGAN also used to generate seasonal changes [RZT+20]. Similarly, style transfer between cartographic maps and satellite images has been attempted to translate a city's street map into a pseudo-satellite image of that city, and vice versa [AHT+22]. The result is a synthetic image that looks like a Sentinel or Google Earth view of a city given only the map. Extensions of CycleGAN, such as AttentionGAN [TLX+21] also been tried to addressing focusing on important regions or preserving edges. While diffusion models and CycleGAN have been widely explored individually, their combined potential for multispectral image synthesis remains underexplored. Studies such as [BHF+19] have applied GAN-based approaches to synthesize missing or corrupted multispectral images, while [ZWC+24] proposed TransCycleGAN, a novel CycleGAN-based architecture integrated with transformers for super-resolving remote sensing images.

Recently, large vision language models have been explored for captioning. One approach is to use a pre-trained vision encoder like CLIP [RKH+21] visual backbone or a ViT [DBK+20] trained on ImageNet and connect it to a pre-trained language model. Recently, Geochat model was proposed [KDN+24] as vision language model for remote sensing. However, our experiments shows that Qwen2-VL-2B-Instruct model [WBT+24b] generated more detailed captions for satellite images.

6.2. Dataset

In this study, we utilize multiple datasets to facilitate the training and evaluation of our proposed methodology. We used two datasets and we created our dataset: the SkyScript dataset [ZPI+17], the Eurosat dataset [HBD+19] [HBD+18], and our synthetic dataset. Each dataset serves a distinct purpose in training the Stable Diffusion and CycleGAN models.

6.2.1. SkyScript Dataset

The SkyScript dataset [ZPI+17] comprises 5.2M satellite images accompanied by textual captions. We use this dataset to generate captions using the Qwen2-VL-2B-Instruct model [WBT+24b] and compare them with the original captions using approximately 675,000 images. These original image-caption pairs serve as the primary training data for fine-tuning our Stable Diffusion 2-1 base model. By leveraging this dataset, we aim to enhance the ability of our model to generate realistic satellite imagery from textual descriptions while maintaining semantic consistency and we used our generated captions to generate images after the Stable Diffusion 2-1 base model training.

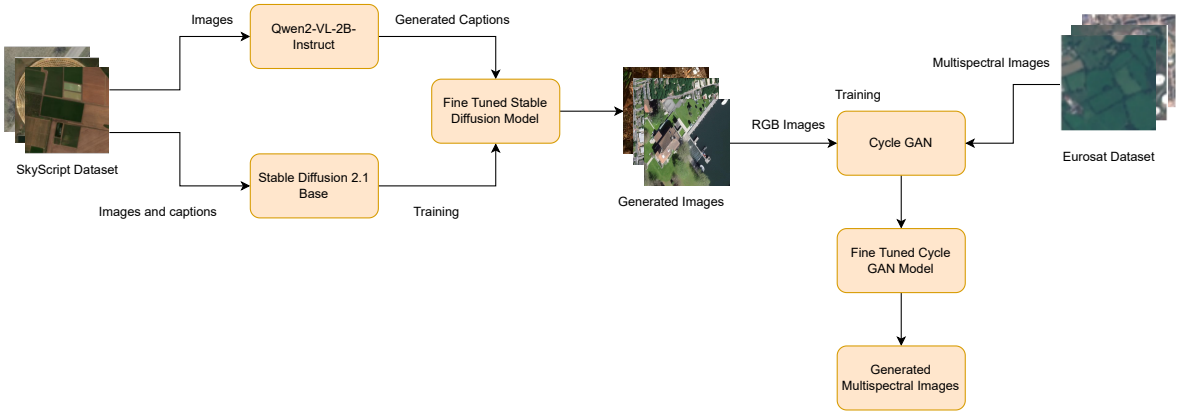


Figure 24: Proposed workflow pipeline of our methodology.

6.2.2. Eurosat Dataset

To train our CycleGAN model, we used 27,000 multispectral images from the Eurosat dataset [HBD+19] [HBD+18]. This dataset contains RGB images and their corresponding 13-band Sentinel-2 multispectral representations, which are essential for training the CycleGAN model to perform accurate image-to-image translation. We used only multispectral images in the Eurosat dataset, because we had unlimited amount of RGB images from our generated images. The Eurosat dataset is widely used in remote sensing applications and provides high-quality multispectral data for various geospatial tasks [HBD+19] [HBD+18]. By utilizing this dataset, we ensure that our model learns to transform RGB images into realistic multispectral representations while preserving spectral integrity and we used our generated RGB images to create multispectral images after the CycleGAN training.

6.2.3. Synthetic Dataset

A novel dataset called SkyScript created in our experiments. This dataset contain 120,000 generated RGB and Sentinel-2 multispectral images, along with their corresponding textual captions. The novel synthetic dataset serves multiple purposes: first it allowed to have extended dataset of all three modalities; secondly, by creating this dataset, we establishing a new benchmark for generated multispectral imagery and provide a new valuable resource in the remote sensing field.

6.3. Methodology

6.3.1. General Overview

The proposed methodology aims to unify caption, RGB and multispectral data with ability to translate all three types of data. Our workflow consists of two main stages: text-to-image generation by fine-tuning Stable Diffusion model and image-to-multispectral conversion using a CycleGAN model see Fig. 24. This proposed pipeline allows for the creation of realistic synthetic satellite imagery that can be transformed into multispectral Sentinel-2 representations. Experiments were implemented using Google Colab Pro+ account and A100 GPU were used during this study.

6.3.2. Caption Generation

The Qwen2-VL-2B-Instruct model [WBT+24a] was used in our experiments. Its performance on large-scale datasets made it particularly well-suited for generating captions for our SkyScript dataset. These generated captions were then compared against the original captions provided with the dataset. Comparison was performed using both automated text similarity metrics like widely used evaluation metrics, including BLEU [PRW+02], METEOR [LA07], ROUGE-L [Lin04], CIDEr-D [VLP15], BERT-F1 [ZKW+19], and CLIPScore [HHF+21], were employed to assess the quality and semantic accuracy of the generated captions. We also apply manual inspection to ensure semantic consistency and overall quality.

6.3.3. Fine-tuning the Stable Diffusion Model

The Stable Diffusion 2-1 Base model [RBL+22] was used in our experiments. Chosen due to its high image generation quality for text-to-image synthesis [LXH+24; SKD+24]. It's pre-trained weights made as an ideal candidate for fine-tuning on our the domain-specific SkyScript dataset. The fine-tuning process involved using image-caption pairs derived from the SkyScript dataset. During training, key hyperparameters were carefully optimized. For instance, a batch size of 4 and a learning rate of 1×10^{-5} were selected. The number of training epochs was determined through experimental evaluation to ensure optimal performance while avoiding overfitting. Post fine-tuning, the model's performance was assessed by generating approximately 120,000 synthetic images.

6.3.4. CycleGAN Model Training

CycleGAN when compared to alternative models like Pix2Pix [IZZ+17] or StarGAN [CCK+18], CycleGAN's advantage lies in its capability to handle transformations without the need for strictly paired training data, which is critical for converting RGB images to multispectral formats [Tad22], [BG23], [ZXS+22]. The CycleGAN model was trained using approximately 27,000 multispectral images from the Eurosat dataset. In this setup, syntetic RGB images served as inputs, and Eurosat 13-band Sentinel-2 multispectral images acted as target outputs. The training process converts input RGB images (3-channel) into output multispectral images consisting of 13 spectral bands. This transformation is designed to accurately map the RGB domain into the multispectral domain, preserving the critical spectral characteristics necessary for remote sensing applications.

The CycleGAN model employed in this study consists of a generator and a discriminator network, designed to translate RGB images into multispectral representations while preserving critical spectral information (Figure. 25). The generator network G takes input RGB images (3 channels, domain A) and maps them into a target multispectral space (13 channels, domain B) through a deep convolutional architecture enhanced with residual learning and channel attention mechanisms. Generator F have the opposite task.

Traditional CycleGAN models typically enforce cycle consistency using pixel-based losses such as L1 or L2 norms. However, in the context of converting RGB images to multispectral representations, preserving the inherent spectral relationships between bands is crucial. To address

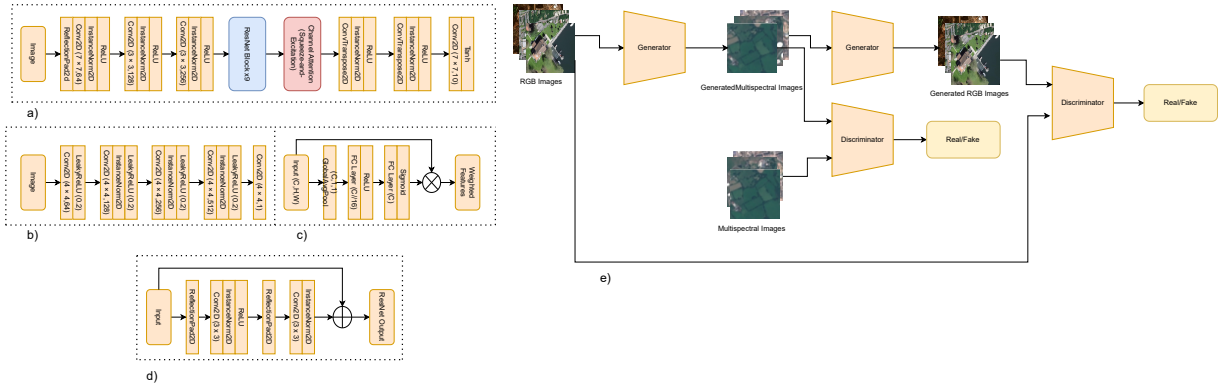


Figure 25: a) Generator Arcitecture, b) Discriminator Architecture, c) Channel Attention (Squeeze & Excitation) Block, d) ResNet Block, e) Cycle GAN diagram

this, we integrated the Spectral Angle Mapper (SAM) [YBG92] loss into our model. SAM loss measures the angular difference between spectral vectors, providing a robust metric for spectral similarity that is less sensitive to variations in illumination intensity. This approach ensures that the generated multispectral images maintain the critical spectral characteristics necessary for accurate remote sensing analysis. The equation for SAM Loss:

$$\mathcal{L}_{SAM}(x, y) = \arccos\left(\frac{\langle x, y\rangle}{\|x\|_2 \|y\|_2}\right) = \arccos\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}\right)$$

The final loss function of generator:

$$\begin{aligned}
\mathcal{L}_G &= \mathcal{L}_{GAN}(G, D_B) + \mathcal{L}_{GAN}(F, D_A) + \lambda_{cycle}\left(\mathcal{L}_{cycle}^{SAM}(A) + \mathcal{L}_{cycle}^{SAM}(B)\right) \\
&= \underbrace{\mathbb{E}_{a\sim p_A}\left[(D_B(G'(a)) - 1)^2\right]}_{\text{loss_GAN_G}} + \underbrace{\mathbb{E}_{b\sim p_B}\left[(D_A(F(b)) - 1)^2\right]}_{\text{loss_GAN_F}} \\
&\quad + \lambda_{cycle}\left(\underbrace{\mathbb{E}_{a\sim p_A}\left[SAM(F(G'(a)), a)\right]}_{\text{loss_cycle_A}} + \underbrace{\mathbb{E}_{b\sim p_B}\left[SAM(G'(F(b)), b)\right]}_{\text{loss_cycle_B}}\right)
\end{aligned}$$

For the discriminators:

$$\begin{aligned}
\mathcal{L}_{D_A} &= \frac{1}{2}\,\mathbb{E}_{a\sim p_A}\left[(D_A(a) - 1)^2\right] + \frac{1}{2}\,\mathbb{E}_{b\sim p_B}\left[(D_A(F(b)))^2\right], \\
\mathcal{L}_{D_B} &= \frac{1}{2}\,\mathbb{E}_{b\sim p_B}\left[(D_B(b) - 1)^2\right] + \frac{1}{2}\,\mathbb{E}_{a\sim p_A}\left[(D_B(G(a)))^2\right].
\end{aligned}$$

with $\lambda_{cycle} = 10$, mean square loss and SAM loss.

To assess the realism of the generated multispectral images, a PatchGAN-based [IZZ+17] discriminator is utilized. This discriminator employs four convolutional layers with progressively increasing feature map depth and decreasing spatial resolution. To stabilize training, LeakyReLU activations and instance normalization are applied. The final layer outputs a scalar value indicating whether the input image is real or synthetic.

Moreover, since the images in the Eurosat dataset are 64x64, we trained the model by randomly cropping a 64x64 region from each generated image. During inference, we tested our model on 512x512 images using a 64x64 sliding window.

Additionally, a specialized preprocessing pipeline was implemented for Sentinel-2 multispectral images. The SentinelToTensor transformation converts raw multispectral arrays into PyTorch tensors with a channel-first ordering. The SentinelResize transformation ensures uniform image dimensions using bilinear interpolation, facilitating consistent model training. These preprocessing steps, combined with the CycleGAN architecture, optimize the model's ability to generate realistic multispectral images from unpaired RGB inputs.

6.4. Experiments

6.4.1. Comparison of Original and Generated Captions

We evaluated the effectiveness of our proposed methodology by conducting multiple experiments. We first assess the performance of the Qwen2-VL-2B-Instruct model in generating 675,000 textual captions for satellite images by comparing the generated captions with the original captions from the SkyScript dataset. The captions are measured as presented in Table 17.

Table 17. Qwen2-VL-2B-Instruct Caption Similarities

Metric	Value	References
BLEU-1	0.1567	[PRW+02]
BLEU-2	0.0526	[PRW+02]
BLEU-3	0.0298	[PRW+02]
BLEU-4	0.0183	[PRW+02]
METEOR	0.1353	[LA07]
ROUGE-L	0.1812	[Lin04]
CIDEr-D	0.0132	[VLP15]
BERT-F1	0.4636	[ZKW+19]
CLIPScore	**0.6822**	[HHF+21]

- BLEU: The BLEU-1 score of 0.1567 indicates some word-level overlap between the generated and reference captions. However, the sharp drop to a BLEU-4 score of 0.0183 reveals that the model has difficulty generating longer, coherent n-gram phrases that match the human references.
- METEOR: With a score of 0.1353, METEOR which factors in stemming and synonym matching suggests that there is only limited semantic overlap between the generated captions and the references.
- ROUGE-L: A ROUGE-L score of 0.1812 reflects a low overlap in the longest common subsequences between the generated and reference captions, indicating differences in sentence structure and phrasing.

- CIDEr-D: The extremely low CIDEr-D score of 0.0132 implies that the generated captions diverge significantly from the human references in terms of content emphasis and term frequency weighting.

- BERT-F1: Achieving a score of 0.4636, the BERT-F1 metric shows a moderate level of semantic similarity between the generated and reference captions by leveraging contextualized word embeddings.

- CLIPScore: The highest value, 0.6822, obtained via CLIPScore, indicates a strong alignment between the generated captions and the visual content of the satellite images. This suggests that despite low overlap in traditional text metrics, the captions capture the key visual elements effectively.

Overall, these results demonstrate that while the Qwen2-VL-2B-Instruct model successfully captures key visual elements as evidenced by its high CLIPScore it faces challenges in generating longer, coherent phrases that fully align with human-authored captions. Traditional metrics like BLEU, METEOR, ROUGE-L, and CIDEr-D indicate that the lexical and structural similarities remain modest, even though the moderate BERT-F1 score confirms some preserved semantic content. This balance strong visual-text alignment but limited contextual and phrase-level detail highlights the model's potential for our text-to-image generation tasks using the fine-tuned Stable Diffusion model, while also pointing to clear avenues for further improvement.

6.4.2. Comparison of Original and Generated Images

Table 18. Similarities of Original and Generated Images

Metric	Value	References
Inception Score	6.4737 ± 0.1200	[BS18]
FID	34.2663	[HRU+17]
KID	0.0180 ± 0.0010	[BSA+18]

To evaluate the fidelity of images produced by our fine-tuned Stable Diffusion 2-1 model, we benchmark them against authentic satellite imagery from the SkyScript collection. We employ three standard metrics for generative image quality assessment—Inception Score (IS), Fréchet Inception Distance (FID), and Kernel Inception Distance (KID)—and also calculate CLIPScore to quantify how well each generated image aligns with its corresponding caption. The outcomes of these comparisons are summarized in Table 18.

Higher Inception Score (IS) values suggest more diverse and realistic images, while lower values indicate mode collapse or repetitive patterns in the generated dataset [BS18]. The for our generated images IS is 6.47, indicating good image diversity.

Fréchet Inception Distance (FID) calculates the Wasserstein-2 distance between the feature distributions of generated and real images using a pre-trained Inception network [HRU+17]. Lower

values correspond to higher similarity, with real-world images distribution, typically achieving values close to zero. The FID score in experiment is 34.26, which justify the good similarity between the generated images and real-world satellite images. The FID metric.

The Kernel Inception Distance (KID) is an unbiased metric and provides a more stable estimate of similarity, particularly for smaller datasets [BSA+18]. A lower KID score indicates better similarity between generated and real image distributions. KID score is 0.018, reflecting the statistical difference between the generated and real images.

To further evaluate the semantic consistency of the generated images, we compute CLIPScore, which measures vision-language alignment. The CLIPScore between generated captions and generated images is 0.31, indicating a relatively weak correlation. This suggests that while the model generates images that are visually coherent, they do not always precisely align with the intended textual descriptions.

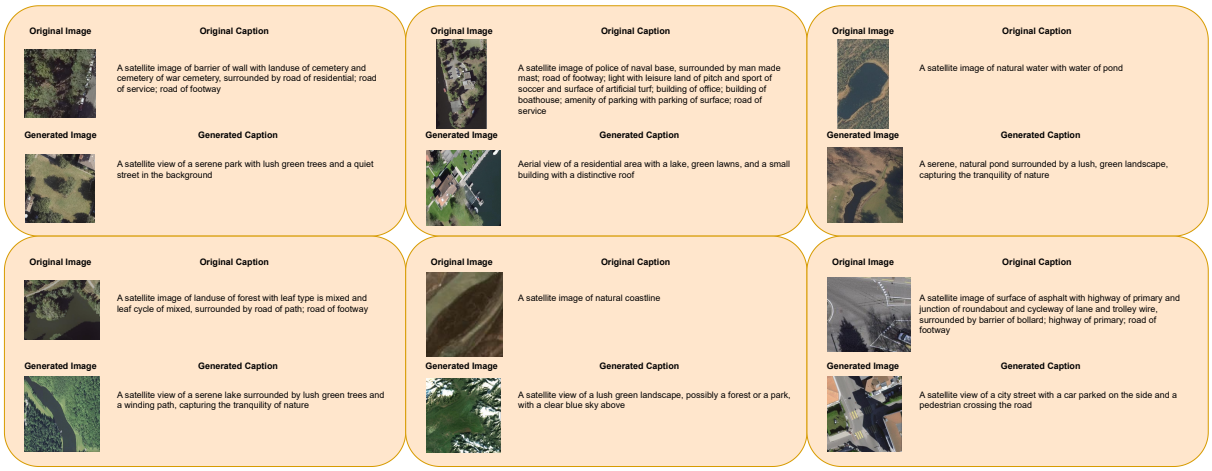


Figure 26: Images and Caption Examples

In contrast, the CLIPScore between original captions and generated captions is 0.68 (Table 17). This relatively high score suggests that the Qwen2-VL-2B-Instruct model successfully generates textual descriptions that are semantically similar to the original captions. However, the much lower 0.31 CLIPScore between generated images and generated captions highlights a disconnect between the text-to-image generation step, suggesting that further fine-tuning is required to ensure the model accurately captures the spatial and contextual details described in the captions. Some image and caption examples are presented in Fig. 26.

6.5. Results

The proposed methodology demonstrates a versatile pipeline capable of generating and transforming different modalities of satellite imagery and text descriptions. Specifically, the pipeline enables three distinct applications:

1. Caption-to-RGB and Multispectral Image Generation

2. RGB Image-to-Caption and Multispectral Image Translation

3. Multispectral Image-to-RGB and Caption Generation

Each of these transformations is facilitated by the integration of fine-tuned Stable Diffusion and CycleGAN models, enabling a bidirectional relationship between textual descriptions, RGB imagery, and multispectral data.

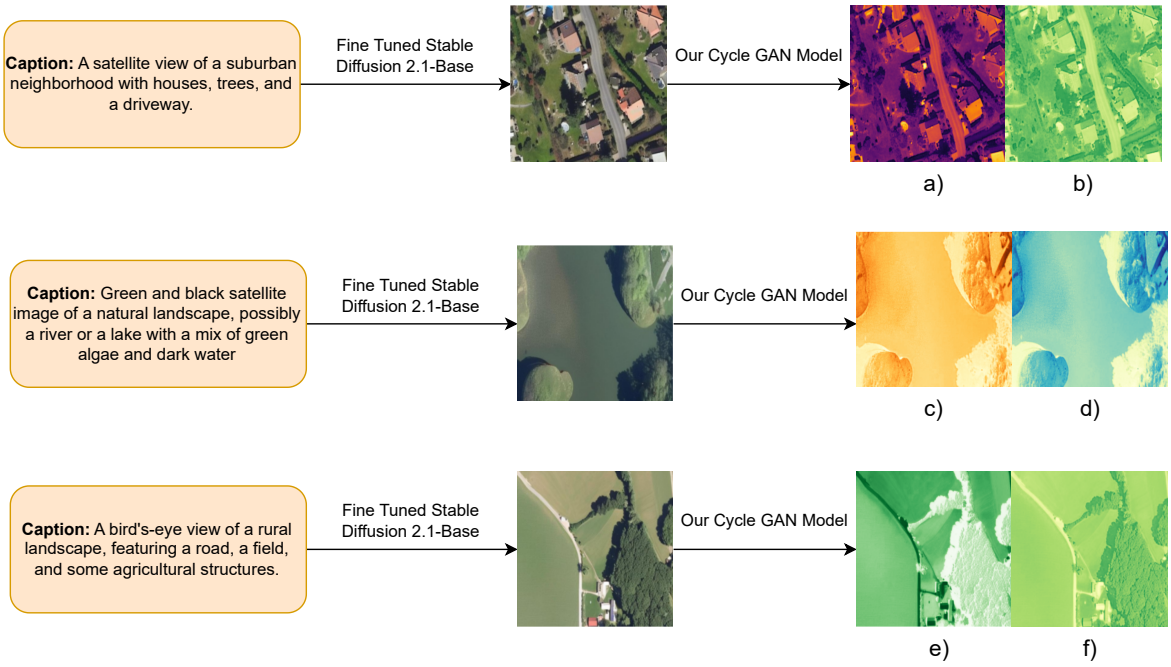


Figure 27: Examples of generated spectra: (a) Thermal Spectrum, (b) NDVI Spectrum, (c) Short Wave Infrared (SWIR) Spectrum, (d) Bathymetric Spectrum, (e) Agriculture Spectrum, (f) NDVI Spectrum.

When only a textual description (caption) is available, the fine-tuned Stable Diffusion 2-1 Base model is capable of generating a realistic synthetic RGB satellite image based on the input caption. The generated image maintains key structural elements described in the text, including features such as water bodies, vegetation, roads, and urban areas. However, due to the limitations of diffusion models, some fine-grained details and spectral characteristics may not be perfectly aligned with real-world satellite images.

Once the synthetic RGB image is generated, it is passed through the CycleGAN model, which translates it into a 13-band Sentinel-2 multispectral image. The CycleGAN model has been trained to map RGB textures to corresponding spectral responses, ensuring that the resulting multispectral image retains realistic spectral information. This approach provides a way to generate plausible multispectral satellite data from captions, which can be useful in scenarios where real multispectral observations are unavailable or incomplete (see Figure 27).

When an RGB satellite image is available without any associated metadata, the proposed pipeline can generate a corresponding textual caption and a multispectral version of the image.

Figure 28 presents the generator loss curve during training. The steady decrease in loss, with the SAM loss dropping to around 4.3 °, indicates the network's improving ability to synthesize multispectral images that preserve critical spectral characteristics.

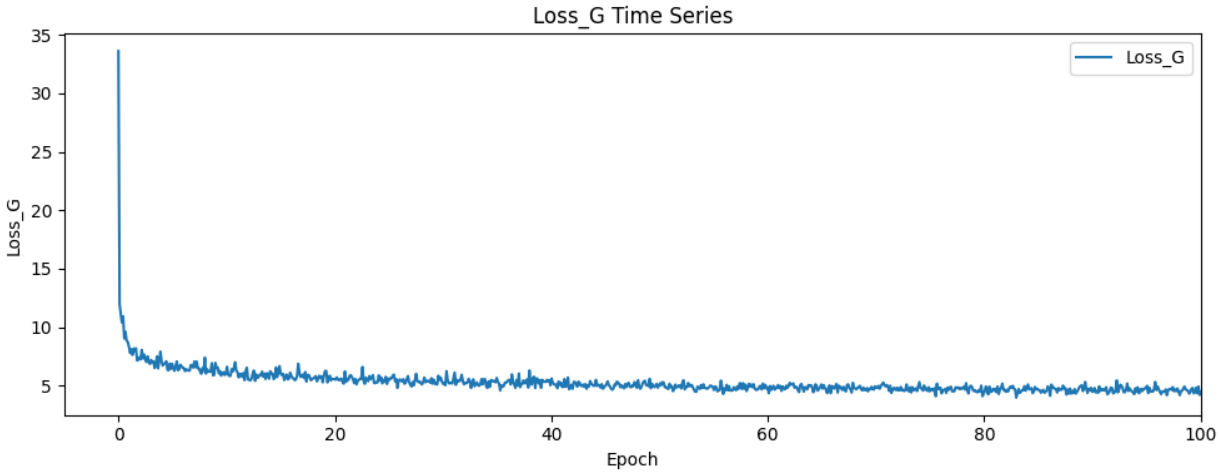


Figure 28: Generator Loss (SAM loss) during training. Notably, the SAM loss decreased steadily, reaching values as low as approximately 4.3.

During inference, the multispectral conversion performance was evaluated using several quantitative metrics. Testing was conducted on 512x512 images by randomly cropping a 64x64 patch from each image. The following table (Table 19) summarizes the results along with the corresponding references for each metric:

Table 19. CycleGAN Inference Metrics

Metric	**Value**	**References**
SAM (°)	10.203992	[YBG92]
SID	0.0726578	[Cha99]
ERGAS	22.931707	[DYK+07]
MAE	3.826507	N/A
MSE	18.279225	N/A

CycleGAN performance is gauged by a series of significant measures that assess different aspects of the quality of the generated images. Spectral Angle Mapper (SAM) gauges the angular difference between the spectral vectors of the generated and reference images. A value of approximately 10.20 ° indicates that the spectral fidelity is highly preserved, which suggests that the spectral information is well maintained across images.

Similarly, the Spectral Information Divergence (SID) value measures the divergence of generated multispectral images from the true ones. With an SID value of 0.07266, the lowest divergence reported ensures high spectral consistency between both sets of images.

The ERGAS measure, which is the average relative error between the reconstruction image and the reference image, is 22.93. This measure is an indicator of the total error, and lower values are generally better since they represent good overall reconstruction quality.

Mean Absolute Error (MAE) and Mean Squared Error (MSE) are utilized to numerically score pixel-wise intensity differences, thereby highlighting reconstruction errors. The MAE value of 3.83 and MSE value of 18.28 illustrate the magnitude of such differences. It must be noted that because

the dataset is unpaired—i.e., generated and reference images belong to fundamentally different samples—the values of MAE and MSE are bound to be greater than they would otherwise be for paired cases.

Overall, the results demonstrate that the proposed pipeline can effectively generate and translate satellite imagery across modalities. While the quantitative metrics indicate promising performance, relatively high FID and LPIPS values suggest that further refinements—such as enhanced training strategies or additional data augmentation—could further improve the realism and spectral consistency of the generated multispectral images.

7. Results

- Achieved up to 97.61% confidence-interval prediction accuracy in targeted test cases, but accuracy dropped to 0% under noise and cloud cover in our semantic segmentation experiment.
- On the S2Looking dataset, threshold optimization yielded an F1 score of 0.969 and accuracy of 0.94 exceeding 0.90 across all datasets and closely matching traditional ML models in out zero shot classification experiment using CLIP Model.
- Generated captions improved interpretability; RoBERTa reached 82.88% validation accuracy on DSIFN, and SVM showed strong overall performance.
- On the RSSCN7 dataset, MiniCPM-V achieved 70.57%; on RSICD 62.19%; on MAI 7.01% (Top-5 Accuracy: 97.83%); and on the merged dataset 43.49%
- Qwen2-VL-2B-Instruct achieved a CLIPScore of 68.22%; constructed 120,000 synthetic RGB–Sentinel-2 pairs; Stable Diffusion (IS 6.47, FID 34.26, KID 0.0180) generated realistic images and CycleGAN generated sentinel-2 images using RGB images with 4.3° training SAM Loss, 10.2° testing SAM Loss.

8. Conclusion

This research investigated advanced methods for spatial-temporal change detection in satellite imagery, focusing on semantic segmentation, zero-shot classification, and caption-based analysis. In addition object recognition, image generation and RGB to Sentinel-2 transformation topics investigated. The key findings are as follows Table 20:

Table 20. Comparison of Methodologies for Change Detection

Methodology	**Strengths**	**Limitations**	**Best Dataset Performance**
Semantic Segmentation	High pixel-level accuracy; robust for spatial change detection	Sensitive to cloud cover and noisy data; computationally intensive	The data which we collected in baltic region
Zero-Shot Classification	No need for labeled data; scalable to unseen classes	Slightly lower precision compared to supervised methods	S2Looking (F1: 0.9690, Accuracy: 0.94)
Caption-Based Analysis	Provides interpretable and descriptive outputs; strong generalization	Dependent on the quality of generated captions	DSIFN (Accuracy: 82.88%)

- CLIP Model achieved high accuracy across diverse datasets for binary change classification task.
- MiniCPM-V Model achieved high accuracy across both single object labeled datasets and multi object labeled dataset for object recognition task. However, it showed worse performance on the merged dataset.
- Caption based change classification results achieved high accuracy using both traditional machine learning and transformer based algorithms.
- Synthetic RGB generation and RGB-to-multispectral translation produce realistic outputs, with ongoing improvements needed for semantic alignment.
- A satellite image captioning dataset created using publicly available datasets (LEVIR-CD, DSIFN, S2Looking, CLCD).
- An artificial multispectral satellite image and caption dataset created
- A Stable Diffusion 2.1 Base model fine tuned and a Cycle GAN model modified and trained for RGB to Sentinel-2 transformation

9. Results Approbation

KP24a K. Kömürcü and L. Petkevicius. Semantic segmentation for change detection in satellite imaging. open-series:57–64, 2024-05. DOI: 10.15388/LMITT.2024.8.

KP24b Kürşat Kömürcü and Linas Petkevičius. Change detection in satellite imagery using transformer models and machine learning techniques: a comprehensive captioning dataset. In DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024. Pp. 56–57. Vilniaus universiteto leidykla, 2024.

KP24c Kürşat Kömürcü and Linas Petkevičius. Zero shot classification for change detection in satellite imagery. In 2024 IEEE 11th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), pp. 1–6, 2024. DOI: 10.1109/AIEEE62837.2024.10586705.

KP25 Kürşat Kömürcü and Linas Petkevičius. Minicpm-v llama model for image recognition: a case study on satellite datasets. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18:7892–7903, 2025. DOI: 10.1109/JSTARS.2025.3547144.

Undergoing Review Kürşat Kömürcü and Linas Petkevičius. Multispectral Caption Data Unification Using Diffusion and Cycle GAN Models. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - ECML PKDD

Image Samples from MAI Dataset				
Ground Truths	residential, bridge, park, stadium	parking lot, residential, bridge, park	commercial, parking lot, residential	river, storage tanks
Predictions	apron, commercial, parking lot, port	residential, roundabout, works	parking lot, commercial, residential	apron, storage tanks
Image Samples from MAI Dataset				
Ground Truths	commercial, parking lot, residential, bridge	residential, park, roundabout	farmland, woodland	commercial, parking lot, residential
Predictions	apron, bridge, commercial, farmland, parking lot, residential, roundabout	apron, baseball field, parking lot	farmland, woodland	residential, commercial, parking lot, roundabout

Figure 29: Predictions for MAI Dataset. Red: Refers to wrong predictions. Blue: Refers to correct predictions which are in the image but not in the ground truth labels

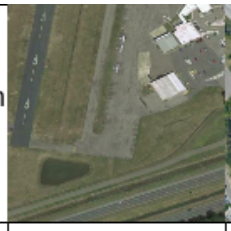
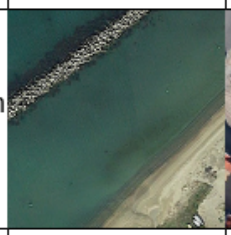
Image Samples from the Merged Dataset				
Ground Truths	apron, parking, lake	residential, river, bridge, lake	baseball field, parking, residential, lake, park	airport
Predictions	runway, apron, parking	residential, bridge, river	residential, roundabout, baseball field	airport
Image Samples from MAI Dataset				
Ground Truths	beach	church	desert	mountain
Predictions	bridge, beach, sea	church, parking, square	desert	mountain

Figure 30: Predictions for the Merged Dataset. Red: Refers to wrong predictions. Blue: Refers to correct predictions which are in the image but not in the ground truth labels

Appendix

References

[AA19] Anju Asokan and J. Anitha. Change detection techniques for remote sensing applications: a survey. *Springer Nature*, 12:143–160, 2019.

[AAA+24] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. Tech. rep. MSR-TR-2024-12, Microsoft, 2024-08. URL: https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/.

[AAE+11] M. Abdelrahman, Asem M. Ali, S. Elhabian, and A. Farag. Solving geometric co-registration problem of multi-spectral remote sensing imagery using sift-based features toward precise change detection, 2011. DOI: 10.1007/978-3-642-24031-7_61.

[AHT+22] Lydia Abady, János Horváth, Benedetta Tondi, Edward J Delp, and Mauro Barni. Manipulation and generation of synthetic satellite images using deep learning models. *Journal of Applied Remote Sensing*, 16(4):046504–046504, 2022.

[Ale20] Dosovitskiy Alexey. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[ALW16] Arabi Mohammed El Amin, Qingjie Liu, and Yunhong Wang. Convolutional neural network features based change detection in satellite images. In *Proceedings of the International Conference on High Performance Computing & Simulation*, 2016.

[AMT13] D. Argialas, S. Michailidou, and A. Tzotsos. Change detection of buildings in suburban areas from high resolution satellite data developed through object based image analysis. *Survey Review*, 45:441–450, 2013. DOI: 10.1179/1752270613Y.0000000058.

[AP18] A. Ambartsoumian and F. Popowich. Self-attention: a better building block for sentiment analysis neural network classifiers, 2018. DOI: 10.48550/arXiv.1812.07860. eprint: arXiv:1812.07860.

[AY+24] 01. AI, : Alex Young, Bei Chen, et al. Yi: open foundation models by 01.ai, 2024. arXiv: 2403.04652 [cs.CL].

[BBY+23] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: a frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[BG23] Anis Bourou and Auguste Genovesio. Unpaired image-to-image translation with limited data to reveal subtle phenotypes. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*:1–5, 2023. URL: https://api.semanticscholar.org/CorpusID:256901143.

[BH05] K. Beurs and G. Henebry. A statistical framework for the analysis of long image time series. *International Journal of Remote Sensing*, 26:1551–1573, 2005. DOI: 10.1080/01431160512331326657.

[BHC15] Y. Byun, Youkyung Han, and T. Chae. Image fusion-based change detection for flood extent extraction using bi-temporal very high-resolution satellite images. *Remote. Sens.*, 7:10347–10363, 2015. DOI: 10.3390/rs70810347.

[BHF+19] J. Bermudez, P. Happ, R. Feitosa, and Dario Augusto Borges Oliveira. Synthesis of multispectral optical images from sar/optical multitemporal data using conditional generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 16:1220–1224, 2019. DOI: 10.1109/LGRS.2019.2894734.

[BHH+21] Eric D. Buduan, Lars Hein, Martin Herold, Johannes Reiche, Yaqing Gou, Maya Gabriela Q. Villaluz, and Ramon A. Razal. Intra-annual identification of local deforestation hotspots in the philippines using earth observation products. *Forests*, 12(8):1008, 2021. DOI: 10.3390/f12081008.

[BMB13] F. Bovolo, C. Marín, and L. Bruzzone. A hierarchical approach to change detection in very high resolution sar images for surveillance applications. *IEEE Transactions on Geoscience and Remote Sensing*, 51:2042–2054, 2013. DOI: 10.1109/TGRS.2012.2223219.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[BS18] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

[BSA+18] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

[CCL+20] Pengyu Chen, Bo Du Chen, Wei Li, and Haifeng Lu. A large-scale dataset for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[CD04] P. Cunningham and S. J. Delany. K-nearest neighbour classifiers: 2nd edition (with python examples), 2004. eprint: arXiv:2004.04523.

[Çel10] T. Çelik. Change detection in satellite images using a genetic algorithm approach. *IEEE Geoscience and Remote Sensing Letters*, 7:386–390, 2010. DOI: 10.1109/LGRS.2009.2037024.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016.

[Cha99] Chein-I. Chang. Spectral information divergence for hyperspectral image analysis. *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)*, 1:509–511 vol.1, 1999. URL: https://api.semanticscholar.org/CorpusID:62257196.

[CLB+19] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Oscd - onera satellite change detection, 2019. DOI: 10.21227/asqe-7s69. URL: https://dx.doi.org/10.21227/asqe-7s69.

[ÇM08] T. Çelik and K. Ma. Unsupervised change detection for satellite images using dual-tree complex wavelet transform. *IEEE Transactions on Geoscience and Remote Sensing*, 48:1199–1210, 2008. DOI: 10.1117/12.794363.

[CMR22] V. Coletta, V. Marsocci, and R. Ravanelli. 3dcd: a new dataset for 2d and 3d change detection using deep learning techniques. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2022:1349–1354, 2022. DOI: 10.5194/isprs-archives-XLIII-B3-2022-1349-2022. URL: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B3-2022/1349/2022/.

[Con23] OpenCompass Contributors. Opencompass: a universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

[Cox58] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242, 1958.

[CS20] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 2020. ISSN: 2072-4292. DOI: 10.3390/rs12101662. URL: https://www.mdpi.com/2072-4292/12/10/1662.

[CXH+20] Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[DCL+19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding, 2019. eprint: arXiv:1810.04805.

[DDC+12] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, et al. Sentinel-2: esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[DK23] Ayush Dabra and Vaibhav Kumar. Evaluating green cover and open spaces in informal settlements of mumbai using deep learning. *Neural Computing and Applications*:1–16, 2023.

[DLB+19] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2019.07.003. URL: http://www.sciencedirect.com/science/article/pii/S1077314219300992.

[dLC21] I. de Gélis, S. Lefèvre, and T. Corpetti. Change detection in urban point clouds: an experimental comparison with simulated 3d datasets. *Remote Sensing*, 13:2629, 2021.

[DMW+20] Huihui Dong, Wenping Ma, Yue Wu, Jun Zhang, and L. Jiao. Self-supervised representation learning for remote sensing image change detection based on temporal prediction. *Remote. Sens.*, 12:1868, 2020.

[DYK+07] Qian Du, Nicholas H. Younan, Roger King, and Vijay P. Shah. On the performance evaluation of pan-sharpening techniques. *IEEE Geoscience and Remote Sensing Letters*, 4(4):518–522, 2007. DOI: 10.1109/LGRS.2007.896328.

[FCS+23] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[Fis22] Sarah Fischer. Optimizing land cover segmentation accuracy in satellite images using jaccard loss. *International Journal of Advanced Remote Sensing*, 29(2):320–332, 2022.

[FKG+21] Tautvydas Fyleris, Andrius Kriščiūnas, Valentas Gružauskas, and Dalia Čalnerytė. Deep learning application for urban change detection from aerial images. In *GISTAM 2021: proceedings of the 7th international conference on geographical information systems theory, applications and management, April 23-25, 2021*, vol. 1, pp. 15–24. SciTePress, 2021.

[FKG+22] Tautvydas Fyleris, Andrius Kriščiūnas, Valentas Gružauskas, Dalia Čalnerytė, and Rimantas Barauskas. Urban change detection from aerial images using convolutional neural networks and transfer learning. *ISPRS International Journal of Geo-Information*, 11(4):246, 2022.

[FLB20] Lei Fang, Shutao Li, and Jon Atli Benediktsson. Dsifn: deeply supervised image fusion network. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020.

[Fri01] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*:1189–1232, 2001.

[GHD+17] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. DOI: 10.1016/j.rse.2017.06.031.

[GLK+21] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *ArXiv*, abs/2104.13921, 2021.

[GPM+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[HBD+18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207. IEEE, 2018.

[HBD+19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[HDO+98] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. DOI: 10.1109/5254.708428.

[HHF+21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: a reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[HLW+24] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

[HML+21] Yuansheng Hua, Lichao Mou, Jianzhe Lin, Konrad Heidler, and Xiao Xiang Zhu. Aerial scene understanding in the wild: multi-scene recognition via prototype-based memory networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:89–102, 2021.

[HNR88] Douglas Holtz-Eakin, Whitney Newey, and Harvey S Rosen. Estimating vector autoregressions with panel data. *Econometrica: Journal of the econometric society*:1371–1395, 1988.

[HRU+17] 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. *Advances in neural information processing systems*, 30, 2017.

[HSP08] M. Haklay, A. Singleton, and C. Parker. Web mapping 2.0: the neogeography of the geoweb. *Geography Compass*, 2(6):2011–2039, 2008. DOI: 10.1111/j.1749-8198.2008.00167.x.

[Hua18] Xiao Huang. Huber loss applications in satellite data analysis for surface elevation studies. *Journal of Atmospheric and Solar-Terrestrial Physics*, 174:50–59, 2018.

[Hua24] Tengjun Huang. Efficient remote sensing with harmonized transfer learning and modality alignment. *arXiv preprint arXiv:2404.18253*, 2024.

[IZZ+17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[JC13] John R. Jensen and David L. Cowen. Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering & Remote Sensing*, 72:577–600, 2013.

[Joh19] Emily Johnson. Application of cross-entropy loss in satellite image classification. *Remote Sensing Reviews*, 21(1):102–112, 2019.

[KDN+24] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[KP24a] K. Kömürcü and L. Petkevicius. Semantic segmentation for change detection in satellite imaging. *open-series*:57–64, 2024-05. DOI: 10.15388/LMITT.2024.8.

[KP24b] Kürşat Kömürcü and Linas Petkevičius. Change detection in satellite imagery using transformer models and machine learning techniques: a comprehensive captioning dataset. In *DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024*. Pp. 56–57. Vilniaus universiteto leidykla, 2024.

[KP24c] Kürşat Kömürcü and Linas Petkevičius. Zero shot classification for change detection in satellite imagery. In *2024 IEEE 11th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pp. 1–6, 2024. DOI: 10.1109/AIEEE62837.2024.10586705.

[KP25] Kürşat Kömürcü and Linas Petkevičius. Minicpm-v llama model for image recognition: a case study on satellite datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:7892–7903, 2025. DOI: 10.1109/JSTARS.2025.3547144.

[KRR+23] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

[LA07] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments:228–231, 2007-07.

[LBX+24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, et al. Mathvista: evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

[LCC+23] Liqin Liu, Bo-Ying Chen, Hao Chen, Zhengxia Zou, and Z. Shi. Diverse hyperspectral remote sensing image synthesis with diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. DOI: 10.1109/TGRS.2023.3335975.

[LCD+22a] M. Liu, Z. Chai, H. Deng, and R. Liu. A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4297–4306, 2022. DOI: 10.1109/JSTARS.2022.3177235.

[LCD+22b] Mengxi Liu, Zhuoqun Chai, Haojun Deng, and Rong Liu. A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4297–4306, 2022. DOI: 10.1109/JSTARS.2022.3177235.

[LCG+23] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: a vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023.

[LCG+24] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: a vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[LDC16] S. Li, S. Dragicevic, and F. A. Castro. Gis-based spatial modeling of land-use change: a case study in the florida keys. *Photogrammetric Engineering Remote Sensing*, 82(5):329–339, 2016. DOI: 10.14358/PERS.82.5.329.

[LDZ+24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, et al. Mmbench: is your multimodal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

[LGG+17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):298–313, 2017.

[LGH+18] Javier López-Fandiño, Alberto S. Garea, Dora B. Heras, and Francisco Argüello. Stacked autoencoders for multiclass change detection in hyperspectral images. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1906–1909, 2018. DOI: 10.1109/IGARSS.2018.8518338.

[LGZ+24] Siqi Lu, Junlin Guo, James Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, P. VanValkenburgh, Steven A. Wernke, and Yuankai Huo. Ai foundation models in remote sensing: a survey. *ArXiv*, abs/2408.03464, 2024. DOI: 10.48550/arXiv.2408.03464.

[LHY+23] Cheng Liao, Han Hu, Xuekun Yuan, Haifeng Li, Chao Liu, Chunyang Liu, Gui Fu, Yulin Ding, and Qing Zhu. Bce-net: reliable building footprints change extraction based on historical map and up-to-date images using contrastive learning. *arXiv preprint arXiv:2304.07076*, 2023. arXiv: 2304.07076 [cs.CV].

[Lin04] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL: https://api.semanticscholar.org/CorpusID:964287.

[Liu+21] C. Liu et al. S2looking: a satellite image dataset for large scale change detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2021.

[LJ22] Hongbing Ma Liangliang Li and Zhenhong Jia. Multiscale geometric analysis fusion-based unsupervised change detection in remote sensing images via flicm model. *Entropy*, 24(2):291, 2022. DOI: 10.3390/e24020291.

[LJG+24] Boyan Liu, Jiayi Ji, Liuqing Gu, and Ziheng Jiang. An integrated cyclegan-diffusion approach for realistic image generation. 13077:130770C - 130770C–10, 2024. DOI: 10.1117/12.3027121.

[LKC15] T. Lillesand, R. W. Kiefer, and J. Chipman. *Remote Sensing and Image Interpretation*. Wiley, 2015.

[LLL+23] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[LLL+24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024.

[LLW+17] Aoxue Li, Zhiwu Lu, Liwei Wang, T. Xiang, and Ji-Rong Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 55:4157–4167, 2017.

[LLW+23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, vol. 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

[LLZ+24] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[LMB+04] D. Lu, P. Mausel, E. Brondizio, and E. Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. DOI: 10.1080/0143116031000139863.

[LMB+21] Marrt Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. Self-supervised pre-training enhances change detection in sentinel-2 imagery. In *Proceedings of the Pattern Recognition and Remote Sensing (PRRS) workshop in International Conference on Pattern Recognition (ICPR)*, LNCS, volume 12667, 2021. DOI: 10.48550/arXiv.2101.08122. URL: https://doi.org/10.48550/arXiv.2101.08122. arXiv preprint arXiv:2101.08122.

[LOG+19] Y. Liu, M. Ott, N. Goyal, J. Du, et al. Roberta: a robustly optimized bert pretraining approach, 2019. eprint: arXiv:1907.11692.

[LPC+19] Yangyang Li, Cheng Peng, Yanqiao Chen, L. Jiao, Linhao Zhou, and Ronghua Shang. A deep learning method for change detection in synthetic aperture radar images. *IEEE Transactions on Geoscience and Remote Sensing*, 57:5751–5763, 2019. DOI: 10.1109/TGRS.2019.2901945.

[LPS+16] Meng Lu, E. Pebesma, Alber Sánchez, and J. Verbesselt. Spatio-temporal change detection from multidimensional arrays: detecting deforestation from modis time series. *Isprs Journal of Photogrammetry and Remote Sensing*, 117:227–236, 2016. DOI: 10.1016/J.ISPRSJPRS.2016.03.007.

[LTC+24] Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL: https://openreview.net/forum?id=dtvJF1Vy2i.

[LWH+23] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023.

[LWZ+] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195. DOI: 10.1109/TGRS.2017.2776321.

[LWZ+17] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[LXH+24] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

[LZC+22] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: a new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. DOI: 10.1109/TGRS.2022.3218921.

[LZW+24] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: mining the potential of multi-modality vision language models. *arXiv preprint*, arXiv:2403.18814, 2024. URL: https://arxiv.org/abs/2403.18814.

[LZX+20] Qingjie Liu, Huanyu Zhou, Qizhi Xu, Xiangyu Liu, and Yunhong Wang. Psgan: a generative adversarial network for remote sensing image pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10227–10242, 2020.

[LZZ+24] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: stronger llms supercharge multimodal capabilities in the wild. https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/, 2024.

[Min61] M. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961. DOI: 10.1109/JRPROC.1961.287775.

[MKJ21] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: a dataset for vqa on document images. In *WACV*, pp. 2200–2209, 2021.

[PPT+20] Chinmayee Pati, A. Panda, A. Tripathy, S. Pradhan, and S. Patnaik. A novel hybrid machine learning approach for change detection in remote sensing images. *Engineering Science and Technology, an International Journal*, 23:973–981, 2020.

[PRW+02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation, 2002-10. DOI: 10.3115/1073083.1073135.

[PVM+05] N. Pettorelli, J. O. Vik, A. Mysterud, J. M. Gaillard, C. J. Tucker, and N. C. Stenseth. Using the satellite-derived ndvi to assess ecological responses to environmental change. *Trends in Ecology Evolution*, 20(9):503–510, 2005. DOI: 10.1016/j.tree.2005.05.011.

[PWD+23] Chao Pang, Jiang Wu, Jian Ding, Can Song, and Gui-Song Xia. Detecting building changes with off-nadir aerial images. *SCIENCE CHINA Information Sciences*, 2023. DOI: 10.48550/arXiv.2301.10922. URL: https://doi.org/10.48550/arXiv.2301.10922. arXiv preprint arXiv:2301.10922v1.

[PZG19] Daifeng Peng, Yongjun Zhang, and H. Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote. Sens.*, 11:1382, 2019.

[Qui86] J. R. Quinlan. *Induction of Decision Trees*, vol. 1 of number 1. Kluwer Academic Publishers, 1986, pp. 81–106.

[RAA+05] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. DOI: 10.1109/TIP.2004.838698.

[RBE+23] Mohamad Mahmoud Al Rahhal, Y. Bazi, Hebah Elgibreen, and M. Zuair. Vision-language models for zero-shot classification of remote sensing images. *Applied Sciences*, 2023.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022-06.

[RHB+18] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/D18-1437. URL: https://aclanthology.org/D18-1437/.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[RLC+24] Papia F. Rozario, Junsu Lee, Yangguang Chen, Pavithra Devy Mohan, Matthew DeWitte, and Rahul Gomes. Analyzing the impact of geospatial derivatives on domain adaptation with cyclegan. *2024 IEEE International Conference on Electro Information Technology (eIT)*:710–715, 2024. DOI: 10.1109/eIT60633.2024.10609908.

[RMA13] D. Renza, E. Martínez, and A. Arquero. A new approach to change detection in multispectral images by means of ergas index. *IEEE Geoscience and Remote Sensing Letters*, 10:76–80, 2013. DOI: 10.1109/LGRS.2012.2193372.

[RXY+23] Mario Fuentes Reyes, Yuxing Xie, Xiangtian Yuan, Pablo d'Angelo, Franz Kurz, Daniele Cerra, and Jiaojiao Tian. A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:74–97, 2023.

ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2023.09.013. URL: https://www.sciencedirect.com/science/article/pii/S092427162300254X.

[RZD+19] C. Ren, A. Ziemann, A. Durieux, and J. Theiler. Cycle-consistent adversarial networks for realistic pervasive change generation in remote sensing imagery. *2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*:42–45, 2019. DOI: 10.1109/SSIAI49293.2020.9094603.

[RZT+20] Christopher X Ren, Amanda Ziemann, James Theiler, and Alice MS Durieux. Cycle-consistent adversarial networks for realistic pervasive change generation in remote sensing imagery. In *2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pp. 42–45. IEEE, 2020.

[Sah13] H. Sahbi. Interactive satellite image change detection. In *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, pp. 3471–3474, 2013. DOI: 10.1109/IGARSS.2013.6723576.

[SCK+22] Ayesha Shafique, Guo Cao, Zia Khan, Muhammad Asad, and Muhammad Aslam. Deep learning-based change detection in remote sensing images: a review. *Remote. Sens.*, 14:871, 2022.

[SDC+19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. eprint: arXiv:1910.01108.

[SEZ21] Sudipan Saha, Patrick Ebel, and Xiao Xiang Zhu. Self-supervised multisensor change detection. *ArXiv*, 2021.

[Sin89] A. Singh. Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10:989–1003, 1989.

[SKD+24] Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: contextually-aware high-resolution satellite image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 460–470, 2024.

[SKH20] A. Song, Yongil Kim, and Youkyung Han. Uncertainty analysis for object-based change detection in very high-resolution satellite images using deep learning network. *Remote. Sens.*, 12:2345, 2020. DOI: 10.3390/rs12152345.

[SLL+21] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*:1–16, 2021. DOI: 10.1109/TGRS.2021.3085870.

[SLW19] Yilan Shao, Yanan Li, and Donghui Wang. Zero-shot detection with transferable object proposal mechanism. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3666–3670. IEEE, 2019.

[SLZ+21] Huiming Sun, Yuewei Lin, Qin Zou, Shaoyue Song, Jianwu Fang, and Hongkai Yu. Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 713–720, 2021.

[Smi20] John Smith. Advanced regression techniques for environmental parameter prediction from satellite data. *Journal of Environmental Monitoring*, 15(3):234–245, 2020.

[SNS+19] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.

[Spa72] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. DOI: 10.1108/eb026526.

[SRF+24] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, et al. Prithvi-eo-2.0: a versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024.

[SS18] Akansha Singh and Krishnavir Singh. Unsupervised change detection in remote sensing images using fusion of spectral and statistical indices. *The Egyptian Journal of Remote Sensing and Space Science*, 2018. DOI: 10.1016/J.EJRS.2018.01.006.

[Str+21] Peter Strobl et al. Digital elevation models: terminology and definitions. *Remote Sensing*, 13(18):3581, 2021. DOI: 10.3390/rs13183581.

[SWZ+20] Xiang Shen, Tongwen Wu, Zhengxia Zou, Xiaoqing Zhang, Zhiqiang Zhou, and Wen Wang. S2looking: a satellite side-looking dataset for building change detection. *Remote Sensing*, 12(16):2643, 2020.

[SYC+21] Li Shen, Lu Yu, Hai Chen, Wei Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: a satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. DOI: 10.3390/rs13245094.

[SZS15] M. Schmitt, X. X. Zhu, and U. Stilla. Fusion of sar and optical remote sensing data–challenges and recent trends. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:63–78, 2015. DOI: 10.1016/j.isprsjprs.2014.02.003.

[Tad22] Sai Pavan Tadem. Cyclegan with three different unpaired datasets. *arXiv preprint arXiv:2208.06526*, 2022.

[TBW+24] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, et al. Cambrian-1: a fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[Tho21] Mark Thompson. Enhancing satellite imagery segmentation with dice loss. *Journal of Geographic Information and Decision Analysis*, 25(4):456–468, 2021.

[TKW+22] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, et al. Dynamicearthnet: daily multi-spectral satellite dataset for semantic change segmentation. *arXiv preprint arXiv:2203.12560*, 2022. DOI: 10.48550/arXiv.2203.12560. arXiv: 2203.12560 [cs.CV]. URL: https://doi.org/10.48550/arXiv.2203.12560. Accepted to CVPR 2022.

[TLI+23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, et al. Llama: open and efficient foundation language models, 2023. eprint: arXiv:2302.13971.

[TLX+21] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 34(4):1972–1987, 2021.

[TMZ+20] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: a large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020.

[VHN+10] J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1):106–115, 2010.

[VLP15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

[VNS23] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

[VSP+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. DOI: 10.48550/arXiv.1706.03762. eprint: arXiv:1706.03762.

[WBT+24a] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, et al. Qwen2-vl: enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[WBT+24b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, et al. Qwen2-vl: enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[WCR+19] M. A. Wulder, N. C. Coops, D. P. Roy, J. C. White, and T. Hermosilla. Land cover 2.0. *International Journal of Remote Sensing*, 40(5-6):1967–1995, 2019. DOI: 10.1080/01431161.2019.1569331.

[WLY+24] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, et al. CogVLM: visual expert for pretrained language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL: https://openreview.net/forum?id=6dYBP3BIwx.

[WO23] Qiusheng Wu and Lucas Prado Osco. Samgeo: a python package for segmenting geospatial data with the segment anything model (sam). *Journal of Open Source Software*, 8(89):5663, 2023.

[Woo06] I. H. Woodhouse. *Introduction to microwave remote sensing*. CRC Press, 2006.

[WRK+16] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

[WWL+22] Yang Wu, Yuyao Wang, Yanheng Li, and Qizhi Xu. Optical satellite image change detection via transformer-based siamese network. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1436–1439. IEEE, 2022. DOI: 10.1109/IGARSS46834.2022.9884408.

[WYD+19] Qi Wang, Zhenghang Yuan, Qian Du, and Xuelong Li. Getnet: a general end-to-end 2-d cnn framework for hyperspectral image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):3–13, 2019. DOI: 10.1109/TGRS.2018.2849692.

[XTu23] XTuner Contributors. Xtuner: a toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. Accessed: 2025-01-30.

[XYJ+23] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Xianyu Jin, and Liangpei Zhang. Ediffsr: an efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2023.

[YBG92] Robert H. Yuhas, Joe W. Boardman, and Alexander F. H. Goetz. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, Pasadena, CA, USA. Jet Propulsion Laboratory (JPL), NASA, 1992-06. Conference Paper, Document ID: 19940023638.

[YDY+19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: generalized autoregressive pretraining for language understanding, 2019. eprint: arXiv:1906.08237.

[YLL+23] Zhiyuan Yan, Junxi Li, Xuexue Li, Ruixue Zhou, Wenkai Zhang, Yingchao Feng, Wenhui Diao, Kun Fu, and Xian Sun. Ringmo-sam: a foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[YN10] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*, 2010.

[YNZ+24] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, et al. Mmmu: a massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. DOI: 10.1109/CVPR52733.2024.00913.

[YQL+22] Bin Yang, Le Qin, Jianqiang Liu, and Xinxin Liu. Ircnn: an irregular-time-distanced recurrent convolutional neural network for change detection in satellite time series. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[YWH+22] Jining Yan, Lizhe Wang, Haixu He, Dong Liang, Weijing Song, and Wei Han. Large-area land-cover changes monitoring with time-series remote sensing images using transferable deep models. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

[YYZ+24a] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, et al. Minicpm-v: a gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[YYZ+24b] T. Yu, Y. Yao, H. Zhang, T. He, et al. Rlhf-v: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. eprint: arXiv:2312.00849.

[YYZ+24c] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, et al. Rlhf-v: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of CVPR*, pp. 13807–13816, 2024.

[YZF+21] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2021.

[YZF+22] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. DOI: 10.1109/TGRS.2021.3078451.

[ZKW+19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[ZNZ+15] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11):2321–2325, 2015.

[ZPI+17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[ZTM+17] X.X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5:8–36, 2017.

[ZV92] H. A. Zebker and J. Villasenor. Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, 30(5):950–959, 1992. DOI: 10.1109/36.175330.

[ZWC+24] Lujun Zhai, Yonghui Wang, Suxia Cui, and Yu Zhou. Transcyclegan: an approach for remote sensing image super-resolution. *2024 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*:61–64, 2024. DOI: 10.1109/SSIAI59505.2024.10508704.

[ZWD+23] Yanpeng Zhou, Jinjie Wang, Jianli Ding, Bohua Liu, Nan Weng, and Hongzhi Xiao. Signet: a siamese graph convolutional network for multi-class urban change detection. *Remote Sensing*, 15(9):2464, 2023. DOI: 10.3390/rs15092464. URL: https://doi.org/10.3390/rs15092464.

[ZXS+22] Yue Zi, Feng-ying Xie, Xuedong Roswell Song, Zhi-guo Jiang, and Haopeng Zhang. Thin cloud removal for remote sensing images using a physical-model-based cyclegan with unpaired data. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. URL: https://api.semanticscholar.org/CorpusID:245625717.

[ZYP+23] Xiaokang Zhang, Weikang Yu, Man-On Pun, and Wenzhong Shi. Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:1–17, 2023. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2023.01.018. URL: https://www.sciencedirect.com/science/article/pii/S0924271623000242.

[ZYT+20] Chen Zhang, Peng Yue, Deodato Tapete, Lijun Jiang, Bin Shangguan, Lei Huang, and Guixiang Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020.

[ZYZ+24] Yin Zhang, Mu Ye, Guiyi Zhu, Yong Liu, Pengyu Guo, and Junhua Yan. Ffca-yolo for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[ZZ22] Lefei Zhang and Liangpei Zhang. Artificial intelligence for remote sensing data analysis: a review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10:270–294, 2022. DOI: 10.1109/mgrs.2022.3145854.

[ZZG+24] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: a large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.