

Evaluating Deep Learning Change Detection in Aerial Imagery: A New Multi-Metric Benchmarking Framework

Ahmed Alaa Abdelbaky Hassouna

Minia University

Mohamed Badr Ismail

Minia University

Amany Shaban Hassan

Minia University

Huthaifa Ashqar

`huthaifa.ashqar@aaup.edu`

Arab American University

Anas M.R. Alsobeh

Southern Illinois University Carbondale

Abdallah A. Hassan

Minia University

Mohammed Elhenawy

Queensland University of Technology

Method Article

Keywords: Change Detection (CD), Remote Sensing, Aerial Images, Deep Learning (DL), 25 Sustainable Development, Benchmarking, Robustness Analysis, Contour Analytics, Model 26 Evaluation.

Posted Date: May 20th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6486635/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Evaluating Deep Learning Change Detection in Aerial Imagery: A New Multi-Metric Benchmarking Framework

Ahmed Alaa Abdelbaky Hassouna ¹, Mohamed Badr Ismail ¹, Amany Shaban Hassan ¹, Huthaifa I Ashqar ^{2*}, Anas M.R. Alsobeh ³, Abdallah A. Hassan ¹, and Mohammed Elhenawy ⁴

¹ Faculty of Engineering, Computers and Systems Department, Minia University, Minia, Egypt; ahmed.alaahassouna@gmail.com; m.badr.ismail@gmail.com; amanyshabban216@gmail.com; abdallah@mu.edu.eg

² AI and Data Science Department, Arab American University, Ramallah, Palestine and AI program, Columbia University, New York, NY USA; huthaifa.ashqar@aaup.edu

³ Information Technology, School of Computing, Southern Illinois University Carbondale, Carbondale, IL USA; anas.alsobeh@siu.edu

⁴ CARRS-Q, Queensland University of Technology, Brisbane, Australia; mohammed.elhenawy@qut.edu.au

* Correspondence: huthaifa.ashqar@aaup.edu

Abstract: Change detection (CD) in aerial images is a critical task in various applications such as urban expansion monitoring, environmental assessment, and disaster response. However, the existing literature often lacks comprehensive and systematic evaluations of deep learning (DL)-based CD models, leaving gaps in understanding their generalizability, robustness, and performance trade-offs across diverse conditions. This study addresses these gaps by proposing a novel framework for benchmarking and assessing CD models, offering a detailed and quantitative evaluation of five state-of-the-art models: CSA-CDGAN, Changeformer, BIT, Tiny, and SNUNet. Our framework consists of three distinct evaluation pipelines: (1) cross-testing across diverse benchmark datasets to assess generalization, (2) sensitivity analysis to examine model performance with respect to change size and complexity, and (3) robustness analysis to evaluate resilience against image corruptions and noise. Key results demonstrate the utility of our framework in revealing the strengths and weaknesses of the evaluated models. CSA-CDGAN excels in handling high noise levels, which showed the highest precision and F1 score, and maintained strong recall across a wide noise spectrum. Changeformer outperforms others in moderately noisy conditions (30-31 dB), while Tiny excels in detecting smaller changes under severe noise (29.35-29.5 dB). Additionally, the framework highlights the challenges faced by BIT, particularly its lower performance in both precision and recall, making it less suited for high-noise environments. This comprehensive benchmarking framework provides critical insights for selecting suitable CD models based on real-world application needs, considering factors like noise levels, change sizes, and dataset variability. The results also lay the groundwork for future research, guiding the development of more robust and versatile CD models. The study establishes a new standard for model evaluation, offering a systematic approach to improve the reliability and applicability of CD models in practical scenarios.

Keywords: Change Detection (CD); Remote Sensing; Aerial Images; Deep Learning (DL); Sustainable Development; Benchmarking; Robustness Analysis; Contour Analytics; Model Evaluation.

Received:

Revised:

Accepted:

Published:

Citation: Lastname, F.; Lastname, F.; Lastname, F. Evaluating Deep Learning Change Detection in Aerial Imagery: A New Multi-Metric Benchmarking Framework. *Journal Not Specified* **2025**, *1*, 0. <https://doi.org/>

Copyright: © 2025 by the authors.

Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Change Detection (CD) is an essential process widely utilized in several fields such as remote sensing, signal processing, and machine learning, crucial for interpreting the

dynamics of our environment. This technique analyzes temporal differences in images from the same location, supporting applications such as land cover change detection (LCCD) [1], disaster assessment [2], urban growth analysis [3], environmental monitoring [4], and infrastructure evaluation [5].

Despite its significance, CD in remote sensing poses notable challenges due to the expanding volume of data, varying atmospheric conditions, and sensor differences [27]. The primary aim is to detect changes precisely to derive actionable insights. Specifically, LCCD plays a pivotal role in monitoring environmental changes over time, facilitating crucial decisions for sustainable development and urban planning [52]. LCCD insights are particularly influential in shaping policies and strategies for urban development, infrastructure resilience, disaster preparedness, and environmental conservation [23]. Such information underpins efforts to sustainably manage urbanization, respond to environmental changes, and develop resilient infrastructures [50].

The methodologies for CD in remote sensing are broadly categorized into traditional and deep learning-based techniques (Figure 1). Traditional approaches encompass algebra-based, transformation-based, and classification-based methods, which are recognized for their efficiency in feature representation and expeditious extraction of changes. However, the advent and integration of deep learning (DL) methods have revolutionized the field by offering robust hierarchical feature representation, outperforming traditional methods in many aspects.

Deep learning-based CD methods include architectures such as Autoencoders (AEs) and Convolutional Neural Networks (CNNs), which are adept at extracting complex features and detecting subtle changes in imagery. These methods demonstrate superior performance due to their intricate feature analysis capabilities.

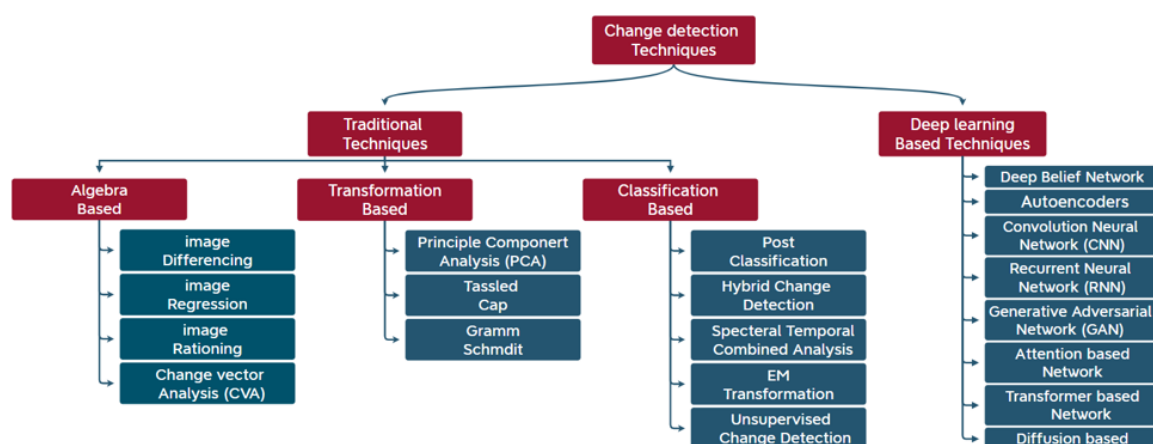


Figure 1. Change Detection (CD) methods.

In Figure 1, the CD methods in remote sensing can be divided into two categories: traditional and deep learning-based techniques. Traditional CD methods include three techniques. Each technique includes a group of models; for instance: the Algebra-based method such as image differencing [6], image regression [7], image rationing [8], and Change Vector Analysis (CVA) [9]. The Transformation-based methods, which include principal component analysis (PCA) [10], Tassled Cap [11], and Gram Schmidt [12], and The Classification-based methods such as post classification [13], hybrid CD [14], spectral temporal combined analysis [15], EM transformation [16], and unsupervised CD [17]. Hence traditional CD methods have been widely applied in CD research for

their advantages in succinct feature representation and rapid change extraction. With the development and growing popularity of deep-learning methods within computer vision, it is natural to apply them to the problem of CD in remote sensing. DL-based methods can represent complex and hierarchical features within the data, which makes them have superior performance over traditional methods.

In terms of DL-based methods, there are seven major types. Autoencoder (AE)-Based Methods, which are widely used in CD tasks as the feature extractor. The commonly used AE-based models are stacked AEs [18], stacked denoising AEs [19], stacked fisher AEs [20], and sparse AEs [21]. Convolution Neural Network (CNN) based methods, enable the accurate extraction of rich and relevant features and capture intricate details, allowing these DL models to effectively analyze and detect changes such as [22] [23] [24] [25].

Nonetheless, the task of CD in aerial images plays a crucial role in various fields, such as urban development, environmental monitoring, and disaster recovery. However, the growing reliance on DL models for CD has highlighted a significant gap in the literature, which is the lack of comprehensive evaluation frameworks that systematically assess the performance, generalizability, and robustness of these models across diverse real-world conditions. Traditional comparisons between models often fail to provide insights into their practical applicability in varying environmental settings, noise levels, and change complexities. Motivated by this gap, our study introduces a novel and extendable benchmarking framework designed to address these challenges. This framework employs three distinct evaluation pipelines including cross-testing on diverse datasets, sensitivity analysis to assess the impact of change size, and robustness analysis against image corruptions. The significance of this framework lies in its ability to rigorously evaluate and compare CD models under a wide range of conditions, providing crucial insights into their strengths, weaknesses, and generalizability. By offering a comprehensive assessment, our framework enables the informed selection of the most suitable models for specific applications, laying the foundation for more reliable and effective CD solutions in practice.

2. Background and Literature Review

LCCD using remote sensing imagery is a critical task with far-reaching implications. By analyzing changes in land cover features like vegetation, water bodies, and built-up areas over time, LCCD supports a wide range of applications [60]. Environmental monitoring is facilitated by tracking deforestation, urbanization patterns, wetland losses, and ecological shifts. Disaster response efforts rely on LCCD to rapidly assess damage from events like floods, earthquakes and wildfires [57] [58]. Urban planners leverage LCCD to guide sustainable development, zoning policies, and infrastructure planning aligned with landscape dynamics. Moreover, LCCD plays a vital role in areas like agriculture by monitoring crop rotations and yield patterns, enabling optimized resource allocation and food security measures. In the realm of climate change studies, LCCD provides crucial data on phenomena like desertification, coastal erosion, and glacier retreats, informing mitigation strategies and impact assessments [51].

However, accurate LCCD remains challenging due to the complexities involved in discriminating true land cover transformations from irrelevant factors such as atmospheric conditions, seasonal variations, sensor viewpoint changes, and data quality issues. Developing robust CD models that can generalize well across diverse scenarios while exhibiting resilience to noise corruptions is of paramount importance for reliable large-scale monitoring and informed decision-making.

This underscores the need for comprehensive benchmarking frameworks that can systematically evaluate the real-world performance of LCCD models, uncover their strengths and limitations, and guide the selection of suitable approaches tailored to specific appli-

cation requirements and operational constraints. Such frameworks are crucial for driving future research towards developing more accurate, robust, and versatile CD solutions aligned with practical needs [52].

2.1. Deep Learning for CD

Deep learning (DL) for CD has been effectively applied across a wide range of infrastructures [23], [5], objects [36], [16], and topographies [1], [50] demonstrating its versatility and robustness. This highlights the adaptability of these models to different conditions and helping to maintain high accuracy across varied terrains [26].

Recurrent Neural Network (RNN) based methods, are capable of learning crucial information and effectively establishing the relationship between multiple sequential remote sensing images, enabling them to detect changes. Many models have been developed based on RNN such as [26], [27], [28]. Generative Adversarial Network (GAN) based methods are used for LCCD leveraging their ability to generate realistic images and discriminate between real and generated samples such as in [29], [30], [31]. Attention-based methods have been proposed to capture spatial and temporal dependencies within image pairs for CD tasks. These models leverage an attention mechanism to focus on crucial areas, enabling them to better identify subtle changes in the scene and distinguish them from usual scene variability such as in [32], [33], [34], and [35].

Transformers-based methods, which were originally developed for natural language processing have gained significant interest in computer vision applications [37]. In contrast to CNNs, transformers have demonstrated a remarkable capacity to capture global dependencies and mitigate the loss of long-range information [38], [36], [37], [39]. Lastly, diffusion-based methods, have been a notable proliferation of proposed CD tasks utilizing the concept of diffusion processes to detect changes in data distributions over time such as in [40], [41], [42].

In the evolving landscape of CD in remote sensing, recent literature has highlighted both the advancements and persisting challenges within the field. Cheng et al. (2023) offer a comprehensive review that navigates through the decade's journey of CD methodologies, particularly spotlighting the transformative role of deep learning. Their discourse unfurls a taxonomy of existing algorithms, shedding light on the nuanced strengths and limitations of various approaches [61]. Despite the breadth of their analysis, the study does not systematically benchmark deep learning-based CD models, leaving a gap in understanding the comparative efficacy of these models under varying real-world scenarios [53].

Similarly, Parelius, E.J. (2023), delves into the deep learning methodologies tailored for multispectral remote sensing images, traversing through algebra-based to transformation-based and deep learning-based methods. The study underscores the rising prominence of deep learning in CD, highlighting the hurdles such as the scarcity of large, annotated datasets and the challenges inherent in model performance evaluation. A critical limitation identified is the labor-intensive process of creating vast annotated datasets for CD and the difficulty in achieving consistency across these datasets, which complicates the comparison and evaluation of deep learning networks [54].

Barkur et al. (2023) introduce RSCDNet, a deep learning architecture designed for CD from bi-temporal high-resolution remote sensing images. Their model, incorporating Modified Self-Attention (MSA) and Gated Linear Atrous Spatial Pyramid Pooling (GL-ASPP) blocks, marks a significant stride in enhancing CD performance. RSCDNet's architecture is celebrated for its efficiency and robustness against various perturbations. However, the discussion on the adaptability of this model across diverse environmental conditions and its performance under different noise levels and image quality variations remains limited [55].

Josephina Paul (2022) explores CD through the lens of deep learning models, focusing on transfer learning and leveraging a Residual Network with 18 layers (ResNet-18) architecture for enhanced detection accuracy. Their innovative approach to batch denoising using CNNs for speckle noise reduction in remote sensing images underscores the potential of deep learning models in CD [25]. Nonetheless, the exploration of these methods' scalability and their comparative analysis in a broader spectrum of CD scenarios are areas left uncharted [56].

Another two related studies are [60], [61]. The first one explores the effectiveness of multi-branch deep learning architectures in satellite imagery classification, which may be applicable to our change detection framework, especially when handling diverse land cover types. The use of multi-branch networks in this context provides a valuable comparison for our model's ability to generalize across different types of change. Similarly, the second one highlights the importance of integrating multi-scale and context-aware features for improving segmentation accuracy in challenging image environments. This work resonates with our approach of leveraging multi-scale analysis for more accurate change detection, particularly when working with images affected by varying noise levels and environmental conditions. These references underscore the relevance of advanced DL and multi-scale techniques, supporting the foundational principles of our proposed framework in remote sensing applications.

2.2. Study Contributions

Unlike the existing literature, which either overlooks the need for a comprehensive benchmarking framework or highlights methodological advancements without extensive comparative analysis, our framework is designed to rigorously evaluate CD models across a spectrum of real-world conditions. It encompasses cross-testing models on a diverse array of benchmark datasets, robustness analysis against image corruptions, and contour level analytics to critically examine the models' detection capabilities [59]. This approach not only provides a panoramic view of each model's strengths and weaknesses but also offers a detailed comparison under varying conditions of image quality, noise levels, and environmental heterogeneity [57]. By addressing the limitations highlighted in the preceding studies, our framework presents a pioneering step towards a holistic evaluation of CD models, paving the way for informed model selection and targeted advancements in the field of remote sensing CD.

There is a pressing need for a thorough benchmarking framework that can evaluate DL-based CD models across varied and realistic conditions reflective of real-world scenarios. While numerous studies propose new models, they often lack comparative evaluation across diverse datasets, which is essential for assessing model generalizability. Additionally, most CD model evaluations do not address robustness, meaning their performance under different noise conditions, such as image corruption from environmental disturbances or lighting variations, remains unknown. This lack of robustness analysis limits our understanding of each model's reliability in practical settings. Furthermore, prior studies rarely examine models' sensitivity to change characteristics, such as the size and complexity of detected changes, despite the importance of these factors in applications like disaster response, urban planning, and environmental monitoring. This oversight makes it challenging to determine which models can effectively detect nuanced or complex changes within specific contexts.

Our proposed framework introduces a systematic and flexible approach to CD model evaluation, addressing the aforementioned gaps in three key ways. First, our cross-dataset generalizability benchmarking pipeline tests models on a range of benchmark datasets, allowing us to measure how each model performs across different data distributions and

real-world scenarios. This cross-testing provides insights into each model's adaptability, establishing a standardized way to assess model generalizability, a crucial need previously unaddressed in CD research. The second pipeline focuses on robustness by examining each model's ability to withstand various image distortions, such as noise, blur, and compression artifacts. By simulating these realistic image conditions, this pipeline fills the existing gap in robustness evaluation, helping practitioners understand which models maintain performance in challenging imaging conditions. Lastly, our third pipeline analyzes models' sensitivity to different change sizes and complexities through contour-level analytics. This sensitivity analysis reveals the limitations and proficiencies of each model in detecting subtle versus extensive changes—particularly relevant in high-stakes scenarios like tracking urban expansion or assessing post-disaster damage. This addition directly addresses the lack of nuanced evaluations on models' detection capabilities across diverse change characteristics.

By integrating these three pipelines, our framework goes beyond traditional model comparison by providing a comprehensive evaluation toolkit. This framework sets a standard for comparing both established and newly developed CD models, enabling objective and reproducible evaluations that foster consistency and progress in the field. The insights it offers into each model's generalizability, robustness, and sensitivity to change enable researchers and practitioners to make more informed decisions, ensuring the suitability of selected models for specific applications and enhancing deployment outcomes. Moreover, the framework is designed to be extendable, allowing new evaluation pipelines to explore additional factors such as specific change types, variations in lighting and weather conditions, and even speed-performance trade-offs. Its adaptability supports the evolving needs of the field, aligning closely with the practical demands of computer vision and remote sensing applications.

3. Cross-Testing and Robustness Analysis of Discrete-Point Models Framework

The proposed Framework, depicted in Figure 2, comprises three parallel pipelines designed to conduct comparative experiments on state-of-the-art CD models. These pipelines aim to both challenge and uncover the capabilities and limitations of these models through rigorous evaluation and benchmarking based on performed experiments. The three pipelines encompass distinct aspects: the first pipeline involves cross-testing of five CD models, the second pipeline examines performance sensitivity analysis with a focus on minimum detectable change size, and the third pipeline focuses on robustness analysis.

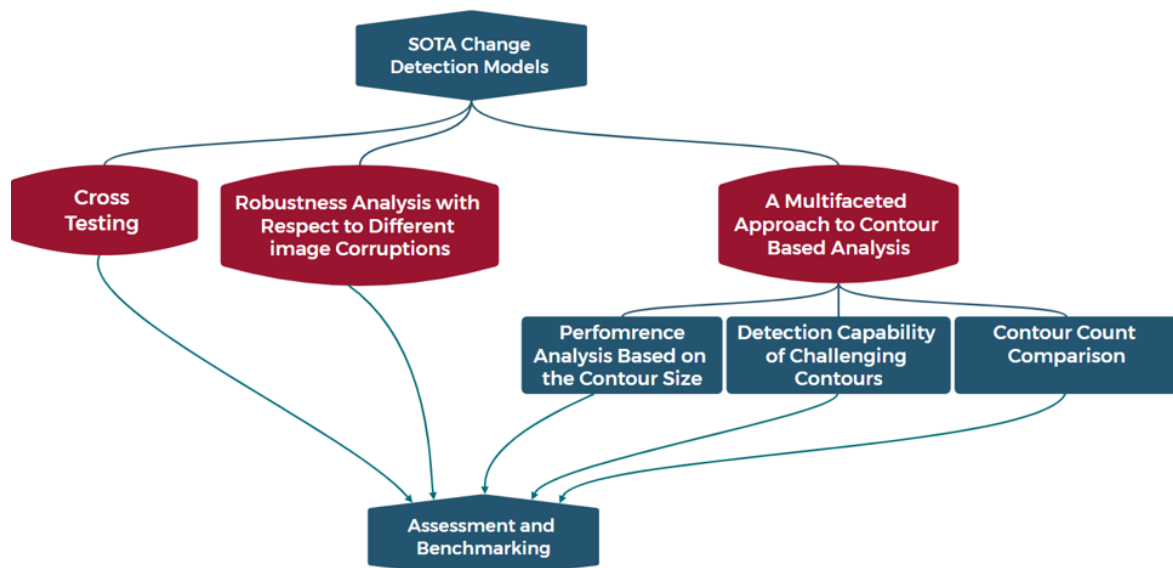


Figure 2. The flow chart of the proposed framework.

The proposed Framework exhibits a range of significant attributes that position it as a valuable instrument for evaluating, comparing, and assessing diverse CD models across various applications and scenarios. The proposed Framework offers a comprehensive approach to evaluating CD models in computer vision applications. It serves multiple purposes: firstly, it assesses the generalizability of CD models by subjecting them to diverse datasets, ensuring their effectiveness in various scenarios. Secondly, it acts as a benchmark, enabling researchers to establish performance baselines and compare new models to existing ones, fostering progress in the field. Additionally, it helps identify strengths and weaknesses in different models through comparative analysis, specifies performance limits regarding change size sensitivity, and evaluates model robustness against noise. Finally, the Framework's results aid in selecting the most suitable CD model based on various criteria, enhancing accuracy and efficiency in real-world applications. The Framework's components are described as following Algorithm 1.

The Framework is thoughtfully designed to be flexible and extendable, Within our Framework, novel pipelines can be incorporated, such as evaluating the performance of CD models in identifying specific types of changes like urban transformations or alterations in natural environments. Furthermore, these pipelines could encompass the assessment of model performance under varying forms of noise or data perturbations and conducting sensitivity analyses regarding distinct lighting or weather conditions. Its adaptability spans diverse computer vision domains, including Aerial Image Segmentation and Object Detection, the focus revolves around identifying key performance factors impacting model performance in this context. Designing experiments or pipelines that evaluate model performance under these factors becomes the key. Similar pipelines to those used in our work can be applied to these tasks. Additionally, pipelines conducting sensitivity analyses concerning scale variations of the object being segmented or detected and the robustness of models against varying lighting intensities or weather conditions can be instrumental. Also analyzing Object detection models based on detection speed becomes equally relevant. Sensitivity analyses for factors like scale, lighting, and weather conditions are essential. Evaluating object detection models for speed is also crucial. The Framework's utility extends to domains like Medical Imaging and Augmented Reality, offering systematic performance evaluation for informed decision-making.

Algorithm 1 Cross-testing CD framework.

Require:

$D = \{D_1, D_2, \dots, D_n\}$: A set of datasets where each D_i is partitioned into training (D_i^{train}) and testing (D_i^{test}) segments.

$M = \{M_1, M_2, \dots, M_m\}$: A set of CD models.

Θ : A set of initialization parameters for the models.

P : The performance metrics used to evaluate the models (e.g., Precision, Recall, F1 Score, PRD).

Dataset Preparation:

Collect a set of diverse datasets D with variability in attributes like geography, land cover types, and imaging conditions.

Partition each dataset $D_i \in D$ into a training segment D_i^{train} and a testing segment D_i^{test} .

Model Training:

for each CD model $M_j \in M$ **do**

 Initialize M_j with parameters Θ_j .

 Train M_j on D_i^{train} for each $D_i \in D$.

end for

Cross-Testing:

for each trained CD model $M_j \in M$ **do**

for each dataset $D_i \in D$ **do**

 Test M_j on D_i^{test} .

 Calculate performance metrics $P(M_j, D_i^{\text{test}})$.

end for

end for

Evaluating Generalization Capacity:

Analyze performance variation $\Delta P(M_j, D_i^{\text{test}})$ for each M_j across all D_i .

Detect patterns of biases or limitations in M_j based on performance variations.

Benchmarking and Selection:

Define benchmarks B based on performance metrics P .

Select a model M_j^* for practical deployment based on its adherence to B .

3.1. The Proposed Framework's Significance

The significance of the proposed framework is highlighted by its comprehensive approach to evaluating the performance of CD models across a variety of computer vision applications. **Generalizability Assessment:** The framework plays a crucial role in assessing the generalizability of CD models. By testing models with varied datasets characterized by distinct features, we can determine their effectiveness in detecting changes across different scenarios, ranging from urban areas to natural environments.

Benchmarking: Acting as a benchmark, the framework facilitates the evaluation of state-of-the-art methods in CD. This capability allows us to establish a performance baseline and compare emerging models or techniques against established ones, promoting advancements in the field.

Identification of Strengths and Weaknesses: Through comparative analysis, the inherent strengths and weaknesses of different CD models are identified. Performance comparisons based on specific criteria or various contexts help us pinpoint each model's limitations and areas for improvement.

Detection Capability Analysis: The framework provides a performance sensitivity analysis of CD models based on the change size, shedding light on the models' sensitivity and performance limits.

Robustness Analysis: By evaluating model performance under various noise types and levels, the framework identifies the most robust models for specific tasks, leading to more reliable outcomes.

Model Selection: Results from framework experiments enable us to identify the most suitable CD model for a given task. This selection is based not only on performance metrics regarding the testing dataset but also on the model's overall generalizability, detection capability, and resilience against noise in the targeted environment and application scenario.

Model Generalizability Pipeline: This pipeline rigorously tests CD models across various datasets with different characteristics to evaluate model adaptability to different environmental conditions, seasonal variations, and terrains. Cross-dataset evaluations gauge models' performance maintenance outside their training conditions, highlighting their robust applicability across different real-world scenarios without extensive retraining or customization.

Detection Capacity Pipeline: This pipeline focuses on the models' ability to detect changes of various sizes and complexities, critical for applications where detection granularity has far-reaching implications.

Noise Resilience Pipeline: It assesses model robustness against image corruptions such as noise from varying weather conditions, sensor inaccuracies, or transmission errors, vital for ensuring CD model consistency and dependability in real-world situations.

These pipelines collectively provide a comprehensive overview of each model's strengths and weaknesses in scenarios that mimic operational environments, allowing for informed model selection based on adaptability, discernment, and resilience—qualities indispensable for practical, real-world CD tasks.

3.2. Cross-Testing Pipeline

Figure 3 describes a collection of diverse datasets are used to assess the adaptability and accuracy of various CD algorithms. Each algorithm is subjected to a battery of tests across different datasets that have been partitioned into training and testing segments. This ensures that the performance of each model can be scrutinized under varied conditions, highlighting their strengths and exposing any weaknesses. The comprehensive cross-testing aims to establish benchmarks for model reliability and effectiveness, ensuring that the selected models can robustly handle real-world scenarios and data variability.

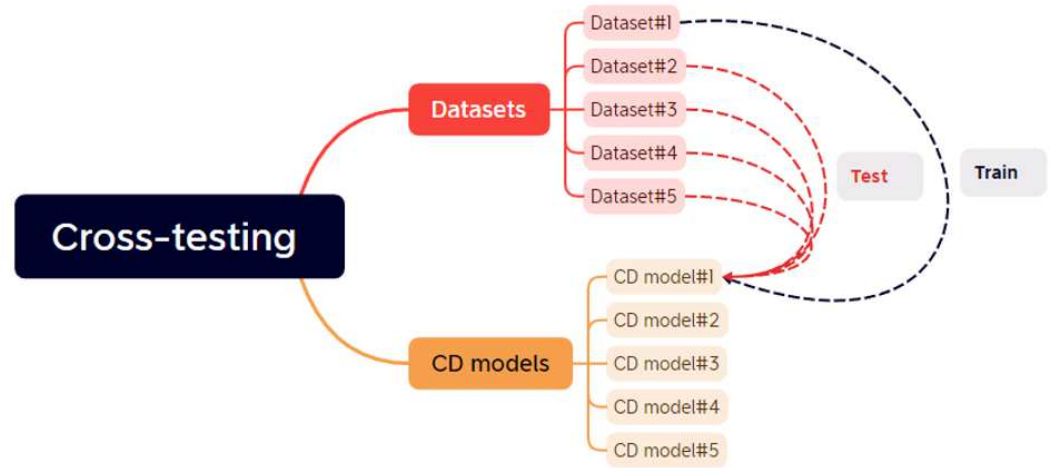


Figure 3. Illustration of the cross-testing pipeline.

A significant challenge for deep learning CD models is their ability to generalize across diverse and complex datasets that differ substantially from their training data. Our cross-testing pipeline evaluates CD models' generalization capacity across benchmark datasets and the impact of testing them on data distinct from their training set. This approach determines whether CD models can effectively identify changes in various real-world scenarios and adapt to unseen contexts, aiming to uncover any potential biases or limitations within the models' training that could affect their performance in the real world. The chosen testing datasets exhibit variations in resolution, change classes, locations, and camera setups, including building change datasets and general change datasets, encompassing diverse change types.

3.3. A Multifaceted Approach to Contour-Based Analysis

The current approach includes three related pipelines focused on contour-level analysis, where the term "contour" refers to interconnected curves or lines that delineate the boundaries of changed objects sharing the same intensity. Conducting experiments at the contour level rather than at the image level offers valuable insights and provides fine-grained information. The investigated phases are as follows:

Performance Sensitivity Analysis Based on the Change Size: A critical factor influencing CD model performance is their sensitivity to changes of various sizes. We have established a comprehensive pipeline for the comparative assessment of these models, specifically evaluating their ability to detect changes across different dimensions. This evaluation requires a dataset that includes well-separable contours of different sizes in its ground truth images. We then assess our CD models on each contour separately, establishing a relationship between the size of each contour in the ground truth and the performance metrics for the four models.

For each contour C_i in the ground truth dataset, let $S(C_i)$ denote the size of the contour. The performance metric $P(C_i, M_j)$ for model M_j on contour C_i can be defined as a function of the contour size:

$$P(C_i, M_j) = f(S(C_i)),$$

where f is a function that maps contour size to the model's performance metric (e.g., accuracy, precision).

Detection Capability of Challenging Contours: Identifying scenarios where CD models fail to detect changes is crucial for improving model performance. In this pipeline, we examine each contour within the ground truth dataset and assess its Intersection over

Union (IoU) with the predictions from the four CD models. An IoU of zero indicates that the contour is particularly challenging and remains undetected by the models, allowing us to compare the models' capabilities in detecting these challenging contours.

The Intersection over Union (IoU) for a contour C_i with respect to model M_j 's prediction can be defined as:

$$IoU(C_i, M_j) = \frac{\text{area}(C_i \cap \text{prediction}(M_j))}{\text{area}(C_i \cup \text{prediction}(M_j))},$$

where $\text{prediction}(M_j)$ is the set of contours predicted by model M_j .

Contour Count Comparison: Analyzing the disparity between the counts of predicted and ground truth contours provides valuable insights into the model's treatment of contours, especially in terms of whether it merges them or separates them. This analysis reveals the model's behavior towards closely connected or neighboring building outlines, indicating its tendency to either merge adjacent changes or identify subtle differences within closely situated building contours.

Let N_{GT} be the number of contours in the ground truth and $N_P(M_j)$ be the number predicted by model M_j . The disparity can be calculated as:

$$Disparity(M_j) = |N_{GT} - N_P(M_j)|$$

Robustness Analysis with Respect to Different Image Corruptions: One of the foremost challenges for CD models in aerial imagery is the presence of noise, which can significantly affect the accuracy of LCCD models. Therefore, analyzing and comparing CD models based on their resilience to various forms of noise offers insights into how noise levels impact these models' performance, contributing to the development of strategies for enhancing their robustness in noisy conditions.

For a given noise level L and type T , the robustness metric $R(M_j, L, T)$ can be defined as the performance of model M_j under the specified noise conditions:

$$R(M_j, L, T) = \text{performance_under_noise}(M_j, L, T)$$

Algorithm 2 shows the structure of the framework for analyzing CD models at the contour level, focusing on sensitivity to change sizes, detection capabilities for challenging contours, contour count disparities, and robustness against image corruptions. We clearly define the inputs required for the entire analysis and the expected outputs. This gives readers an overview of what data is needed and what results they can anticipate from the analysis. This phase focuses on determining the size of each contour in the ground truth and evaluating the performance of each model on these contours. The inputs are the set of models and the ground truth contours, and the outputs are the sizes of the contours and the performance metrics (e.g., accuracy, precision) for each model and contour. After evaluating individual contours' sizes and performance, we compare the number of contours detected by each model against the ground truth. This step identifies discrepancies in model predictions. The inputs are the predicted contours from each model and the ground truth contours, with the output being the disparity in contour counts for each model. The models' robustness against various types of image corruption by applying different noise levels and types. This phase evaluates how well each model performs under less-than-ideal conditions. The inputs are the models, noise levels, and noise types, and the outputs are the performance metrics of models under these conditions.

Algorithm 2 Comprehensive contour analysis for CD models.

Inputs: Set of models $\{M_j\}$, ground truth contours $\{C_i\}$, noise levels, and noise types.

Outputs: Performance metrics, robustness analysis, contour count disparity.

Contour Size and Performance Sensitivity Analysis

Inputs: Ground truth contours $\{C_i\}$, models $\{M_j\}$.

Outputs: Contour sizes, performance metrics for each model and contour.

```

for each model  $M_j$  in models do
  for each contour  $C_i$  in ground_truth_contours do
    size  $\leftarrow$  compute_size( $C_i$ )
    performance  $\leftarrow$  evaluate_performance( $M_j, C_i$ )
    record(size, performance)
  end for
end for

```

Contour Count Comparison

Inputs: Predicted contours from models $\{M_j\}$, ground truth contours.

Outputs: Disparity in contour counts for each model.

```

for each model  $M_j$  in models do
   $N_{GT} \leftarrow$  count(ground_truth_contours)
   $N_P \leftarrow$  count(predicted_contours( $M_j$ ))
  disparity  $\leftarrow$  abs( $N_{GT} - N_P$ )
  record(disparity)
end for

```

Robustness Analysis with Respect to Different Image Corruptions

Inputs: Models $\{M_j\}$, noise levels, noise types.

Outputs: Performance of models under various noise conditions.

```

for each noise level  $L$  in noise levels do
  for each noise type  $T$  in noise types do
    for each model  $M_j$  in models do
      apply_noise( $T, L$ )
      performance  $\leftarrow$  evaluate_performance( $M_j$ )
      record_noise_impact( $L, T, M_j$ , performance)
    end for
  end for
end for

```

4. Materials and Methods

4.1. Data Collection: Datasets

We selected five benchmark land cover CD datasets that are publicly available. Three of them are primarily focused on building CD datasets: LEVIR-CD, WHU-CD, and S2Looking. The other two datasets, CDD and CLCD, cover multiple types of changes, hence can be referred to as general CD datasets.

- LEVIR-CD [43]: A large-scale dataset in the CD field, consisting of 637 sets of Google Earth images with a resolution of 0.5 m px^{-1} and a size of 1024×1024 . Each set contains the image before and after the building changes and a corresponding label. The original images were divided into 256×256 size images without overlap. The dataset was divided into a training set of 7120 images, 1024 for validation, and 2048 for testing.
- WHU-CD [44]: A public building CD dataset consisting of a pair of aerial images of size 32507×15354 and a high resolution of 0.075 m px^{-1} . A default cropping of 256×256 was applied to it without overlap, obtaining a training set of 6096 images, 762 in the validation set, and a test set of 762.
- S2Looking [45]: A building CD dataset that contains large-scale side-looking satellite images captured at varying off-nadir angles. It includes 5,000 bi-temporal image pairs with a size of 1024×1024 and a resolution ranging from 0.5 m px^{-1} to 0.8 m px^{-1} of rural locations over the world, with more than 65,920 annotated change instances.
- CDD [46]: A widely used dataset for CD, containing 11 pairs of remote sensing images obtained via Google Earth in different seasons, with a spatial resolution ranging from 3 cm px^{-1} to 100 cm px^{-1} . After cropping the original image pairs into the same size of 256×256 pixels, 10000 image pairs were generated for training, 3000 for validation, and 3000 for testing.
- Cropland Change Detection (CLCD) Dataset [47]: Consists of 600 pairs of 512×512 bi-temporal images collected by Gaofen-2 in Guangdong Province, China, with a spatial resolution of 0.5 m to 2 m, each group of samples comprises two images of 512×512 and a corresponding binary label of cropland change. A default cropping of 256×256 was applied to it without overlap.

The datasets selected for this study were chosen to ensure comprehensive coverage of real-world conditions, thus enhancing the robustness and generalizability of the proposed framework. We used five publicly available benchmark land cover CD datasets, which encompass a wide variety of scenarios and environmental conditions. Three of the datasets—LEVIR-CD, WHU-CD, and S2Looking—are focused specifically on building CD, offering high-resolution images of urban areas where building structures change over time. These datasets include images captured at varying resolutions, from 0.075 meters per pixel in WHU-CD to 0.5 meters per pixel in LEVIR-CD and S2Looking, ensuring that the framework can evaluate the performance of CD models under different levels of detail and noise.

In addition to these building-centric datasets, we incorporated two general CD datasets—CDD and CLCD—that cover a broader range of changes, such as those occurring in rural areas and agricultural landscapes. The CDD dataset spans a range of seasonal changes with varying spatial resolutions, from 3 cm per pixel to 1 meter per pixel, allowing for an evaluation of models under diverse seasonal and environmental conditions. The CLCD dataset, with images captured by the Gaofen-2 satellite, provides additional coverage of land-use changes, specifically in cropland areas, offering a different type of environmental context. By selecting these diverse datasets, the framework is exposed to various real-world conditions, such as urban, rural, and agricultural environments, as well

as different image resolutions, seasonal variations, and noise levels. This wide coverage ensures that the evaluation process is robust and that the proposed framework can assess models across a variety of change types and environmental conditions, making it a valuable tool for real-world applications where such variations are common.

4.2. State-of-the-Art (SOTA) CD Models

To validate the effectiveness of the proposed Framework in comparing and analyzing CD models, we applied it to five of the SOTA land CD models which achieved high-performance metrics with benchmark datasets such as LEVIR, WHU, and CDD.

- BIT [48]: A transformer-based feature fusion method that combines CNN with a transformer encoder-decoder structure, capturing effective and meaningful global contextual relationships over time and space.
- SNUNet [49]: A multilevel feature concatenation method that combines NestedUNet with a Siamese network, using an Ensemble Channel Attention Module for deep supervision.
- Changeformer [50]: A transformer-based CD method leveraging a hierarchically structured transformer encoder and multilayer perception (MLP) decoder in a Siamese network architecture to efficiently render multiscale long-range details required for accurate CD.
- Tiny [51]: A lightweight and effective CD model using a Siamese Unet to exploit low-level features globally temporally and locally spatially. It adopts a novel space-semantic attention mechanism called MIX and Attention Mask Block (MAMB).
- CSA-CDGAN [44]: A CD network using a Generative Adversarial Network for detecting changes and a channel self-attention module to improve network performance.

The framework includes five pipelines: the first performs cross-testing over five CD models, the next three perform contour-based analysis, and the last pipeline performs robustness analysis of the CD models by testing them in three noisy versions of the LEVIR dataset. The second and third pipelines are applied only to the CD models trained on the LEVIR dataset.

4.3. Cross-Testing Pipeline

In this experiment, we carefully chose five benchmark datasets. We categorize them into mainly building change datasets, such as LEVIR, WHU, and S2Looking, whose masks contain changes in buildings, and general change datasets such as CDD and Cropland, which contain different types of change such as buildings, roads, lakes, cars, and natural objects. We conducted our experiment with five high-performing CD models: SNUNet, Changeformer, Tiny, BIT, and CSA-CDGAN. The SNUNET model was trained on the CDD dataset, and the other four CD models were trained on the LEVIR dataset. Each CD model was tested with the five CD datasets.

4.4. A Multifaceted Approach to Contour-Based Analysis

We conduct an extensive evaluation of our CD models using the LEVIR dataset, which has high-quality annotation and its ground truth images encompass a wide range of contours with diverse sizes and shapes. Each dataset instance contains the ground truth and the corresponding predictions of our four CD models, as our analysis mainly depends on them. This approach includes three key pipelines, each serving a distinct purpose.

4.4.1. Performance Sensitivity Analysis Based on the Changed Contour Size

Figure 4 shows the evaluation pipeline for assessing CD models in LEVIR instances involves first creating a high-quality dataset where the outlines of changes are easily

distinguishable from the background in both actual images and model predictions. This initial step guarantees that the following analysis concentrates solely on situations where detecting modifications is achievable and clear-cut.

Create a dataset containing N pairs of pre-change (I_i^0) and post-change images (I_i^1) along with their respective object masks, (M_i^0, M_i^1), sourced from Ground Truth (GT) data. Ensure precisely defined bounding boxes minimizing overlap and preventing false positives caused by nearby objects interfering with contour measurement.

Our evaluation pipeline begins by carefully curating a dataset of LEVIR instances where the contours representing changes are clearly separable from their surroundings in both the ground truth images and model predictions. This selective approach ensures we focus our analysis on scenarios where CD is feasible and unambiguous.

Apply edge detection algorithms and morphological operations (erosion, dilation) to extract unique contours $C_j = \{c_j^k\}$ present in both original and altered imagery. Quantify contour size using $|C_j|$, denoting the quantity of constituent pixels composing each contour c_j^k .

We then employ contour separation techniques to isolate individual contours within each image instance. The size of these separated contours is calculated, allowing us to study how contour dimensions impact model performance. Through a contour matching algorithm, we establish correspondence between the ground truth contours and those predicted by each CD model under evaluation.

Estimate geometric properties describing each contour, encompassing area ($A_j^k \propto |C_j^k|$), eccentricity (ϵ_j^k), solidity (σ_j^k), Euler numbers (χ_j^k), and compactness (γ_j^k). Incorporate these traits to characterize shapes presenting challenges for detection.

With the contours matched, we proceed to comprehensively evaluate each model's performance using metrics such as precision, recall, F1 score, and pixel-level relative difference (PRD). These evaluations are conducted on a per-contour basis, considering only the best matched contour prediction for a given ground truth instance. This iterative process continues until all contours and instances have been analyzed.

To uncover deeper insights, we model the intricate relationships between contour size and the various performance metrics calculated during evaluation. This modeling step reveals how factors like contour area or complexity influence metrics like precision or recall for each CD model.

Finally, armed with a granular understanding of size-metric relationships, we conduct a thorough analysis and comparison across all evaluated models. This multi-faceted assessment highlights each model's strengths, weaknesses, and biases, identifying trends that can guide optimal model selection and inform strategies for further performance optimization tailored to specific application requirements.

4.4.2. Match True Positives, False Negatives, and False Positives

Determine true positive (TP), false negative (FN), and false positive (FP) contours according to proximity and IOU thresholds. Set up binary variables for TP, FN, and FP correspondences:

$$y_{ijk}^{TP} = \begin{cases} 1, & \text{if } c_{ik}^0 \in GT, \exists c_{jk}^1 : IOU(c_{ik}^0, c_{jk}^1) > T \\ 0, & \text{otherwise} \end{cases}$$

$$y_{ijk}^{FN} = \begin{cases} 1, & \text{if } c_{ik}^0 \notin GT, \nexists c_{jk}^1 : IOU(c_{ik}^0, c_{jk}^1) > T \\ 0, & \text{otherwise} \end{cases}$$

$$y_{ijk}^{FP} = \begin{cases} 1, & \text{if } c_{ij}^1 \notin GT, \exists c_{ik}^0 : IOU(c_{ij}^1, c_{ik}^0) > T \\ 0, & \text{otherwise} \end{cases}$$

For each contour correspondence category, calculate conventional evaluation indices:
True Positive Rate (Recall):

$$R = \frac{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{TP}}{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{TP} + \sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{FN}}$$

False Discovery Rate (FDR):

$$FDR = \frac{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{FP}}{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{TP} + \sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{FP}}$$

Precision:

$$P = \frac{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{TP}}{\sum_{i=1}^N \sum_{j=1}^{|C_j^1|} y_{ijk}^{TP}}$$

Model complex relationships between contour characteristics and error rates by fitting regression functions coupling estimated evaluation indices (Step 5) with computed contour attributes (Step 3). Examples include fitting a linear model associating Recall with Area: $R = \beta_0 + \beta_1 \cdot A$, followed by interpretation of coefficients β_0, β_1 .

Visualize Receiver Operating Characteristics (ROC) illustrating balances between TP rate and FP rate throughout varying threshold settings for numerous models. Determine overall model discernibility levels by computing areas below ROC curves (AUC). Report mean scores for Precision, Recall, and additional pertinent indices. Execute hypothesis testing to ascertain substantial disparities in model performances.

Identify prevailing themes related to strong or weak model abilities contingent upon targeted applications and issue contexts. Suggest ideal model selections, tunings, and prospective advancements supported by discoveries acquired during the evaluation pipeline.

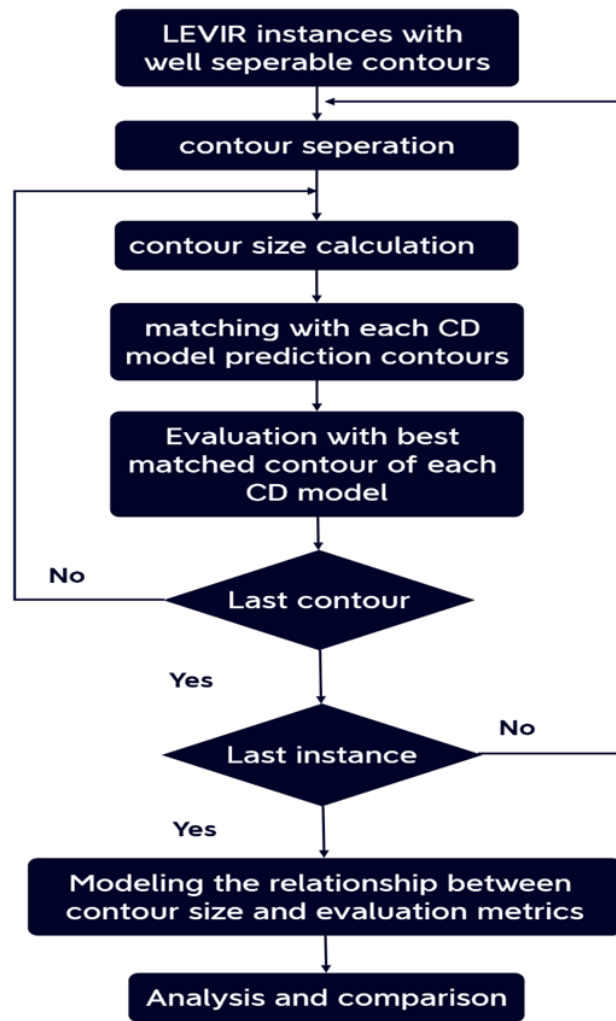


Figure 4. The pipeline of the performance sensitivity analysis based on the change size.

4.4.3. Detection Capability of Challenging Contours

A critical aspect of our evaluation pipeline focuses on assessing the detection capabilities of our CD models, specifically concerning challenging contours. These challenging contours are defined as those present in the ground truth data that at least one CD model fails to detect entirely. To identify such contours, we calculate the Intersection over Union (IoU) between each ground truth contour and the corresponding predicted contours from the CD models. If the IoU equals zero with any model's prediction, we classify that particular contour as challenging. Subsequently, we conduct a comparative analysis of our CD models based on their effectiveness in detecting the maximum number of these challenging contours. We apply this analysis to each non-black ground truth image in the LEVIR testing dataset as shown in Algorithm 3.

Within this pipeline, Algorithm 3, for each instance in the LEVIR dataset, we perform two key tasks: first, we count the number of contours present in the ground truth image, and second, we tally the number of contours in the corresponding prediction from each CD model. Subsequently, we establish a mathematical relationship between these counts through regression analysis, producing a regression line for each model. Our analysis then extends to the comparison of the resulting regression lines generated by our four CD models. We contrast them with the 'identity line,' which forms a 45-degree angle, signifying a scenario where the number of contours in the ground truth perfectly matches the predicted contours. This comparison allows us to evaluate the performance of each

model in terms of its ability to accurately predict the correct number of contours present in the input data. 572
573

Algorithm 3 Comprehensive contour analysis for CD models.

Let C_{gt} be the set of ground truth contours in the LEVIR dataset.
 Let C_p be the set of predicted contours by CD models.
 Initialize $IoU_{scores} = \{\}$ to store IoU scores.
 Initialize $challenging_contours = \{\}$ to store challenging contours.
 Initialize $counts_{actual} = \{\}$ and $counts_{predicted} = \{\}$ for contour counts.
 Initialize $slopes = \{\}$ to store slopes of regression lines.
for each image I in the LEVIR dataset **do**
 $M_{GTRUTH} \leftarrow$ Mask of I with ground truth contours.
 $PREDICTION \leftarrow$ Mask of I with predicted contours.
 $C_{gt} \leftarrow$ Find contours in M_{GTRUTH} .
 $C_p \leftarrow$ Find contours in $PREDICTION$.
 for each contour C_{gt}^i in C_{gt} **do**
 for each contour C_p^j in C_p **do**
 $IoU \leftarrow$ Calculate $IoU(C_{gt}^i, C_p^j)$.
 $IoU_{scores} \leftarrow IoU_{scores} \cup \{IoU\}$.
 if $IoU = 0$ **then**
 $challenging_contours \leftarrow challenging_contours \cup \{C_{gt}^i\}$.
 end if
 end for
 $n_{actual} \leftarrow$ Count of C_{gt} .
 $n_{predicted} \leftarrow$ Count of C_p .
 $counts_{actual} \leftarrow counts_{actual} \cup \{n_{actual}\}$.
 $counts_{predicted} \leftarrow counts_{predicted} \cup \{n_{predicted}\}$.
 end for
 $slope \leftarrow$ Slope from linear regression of $counts_{actual}$ vs $counts_{predicted}$.
 $slopes \leftarrow slopes \cup \{slope\}$.
end for
 Visualize $counts_{actual}$ vs $counts_{predicted}$ and compare to the identity line.
 Rank CD models based on $challenging_contours$ and $slopes$.

4.5. Robustness Analysis with Respect to Different Image Corruptions 574

In the proposed pipeline, we analyze and compare the robustness of four state-of-the-art CD models against three common types of noise found in aerial images: Gaussian noise, salt and pepper noise, and speckle noise. *Types of Noise Affecting CD Models:* 575
576
577

- **Gaussian Noise:** Characterized by a Gaussian distribution with mean $\mu = 0$ and standard deviation σ , representing the noise amount. It arises from sensor limitations, electrical interference, or atmospheric conditions. 578
579
580
- **Salt and Pepper Noise:** Manifests as random occurrences of white and black pixels due to abrupt signal disruptions, often caused by malfunctioning image sensors. 581
582
- **Speckle Noise:** A multiplicative noise that compromises image quality and precision, challenging the identification and analysis of features in remote-sensing images. 583
584
- **Peak Signal-to-Noise Ratio (PSNR)** 585
 The PSNR metric evaluates the quality of a corrupted image H compared to its original form I , calculated as: 586
587

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right), \quad (1)$$

where MAX is the maximum pixel value in the image, and MSE is the mean squared error between I and H :

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (I_{ij} - H_{ij})^2, \quad (2)$$

assuming I and H are of size $m \times n$. Higher PSNR values indicate lower levels of noise, whereas lower PSNR values suggest higher noise levels.

4.6. Experimental Setup

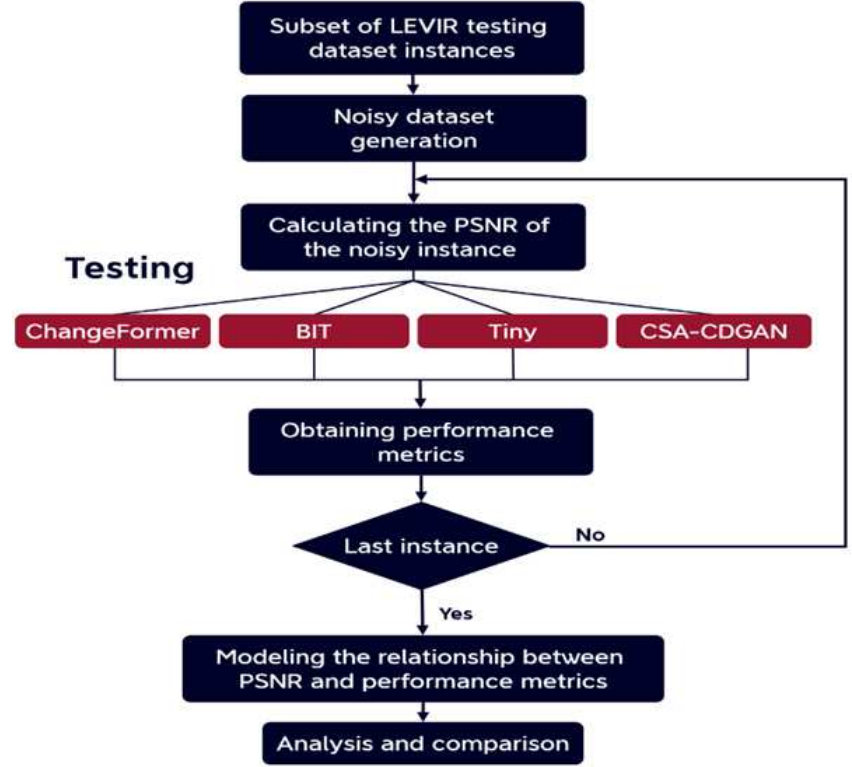


Figure 5. The overall experimental flow.

Figure 5 shows the overall experimental flow of analyzing the robustness of four state-of-the-art CD models to three common types of noise found in aerial images Gaussian noise, salt and pepper noise, and speckle noise. We initially selected randomly 200 instances from the LEVIR testing dataset and then created three distinct noisy datasets from the original LEVIR dataset: Gaussian LEVIR, salt and pepper LEVIR, and speckle LEVIR. The generation process involved selecting each instance from the original LEVIR dataset and applying various levels of noise exclusively to image $T1$ using a Python environment. This process resulted in the creation of new instances, each comprising noisy $T1$, $T2$, and ground truth images.

To diversify the noisy datasets, we introduced different levels of noise to each instance. This intentional variation aimed to expand the dataset by incorporating instances with a range of noise levels. The applied noise levels were carefully chosen to ensure that the resulting noisy datasets exhibit a PSNR range with practical significance.

For each noisy dataset, we selected each instance and calculated the PSNR value between the noisy and clear $T1$ images only because image $T2$ is clear. Subsequently, we obtained the performance metrics of our four CD models BIT, Changeformer, CSA-CDGANs, and Tiny on that instance. The performance metrics include F1, Precision, Recall,

and PRD. We then modeled the relationship between the PSNR and each performance metric based on the distributions of data between them. Finally, we analyzed and compared the CD models based on the modeled relationships.

For each instance in the noisy datasets, the performance metrics - F1, Precision, Recall, and PRD - were computed for the CD models. We modeled the relationship between the PSNR and each performance metric, analyzing and comparing the CD models based on these relationships.

5. Results and Discussion

In the field of CD models, specifically those that use aerial and satellite imagery for analyzing land cover, performance metrics are crucial for assessing the effectiveness of different algorithms. The metrics used, such as precision, recall, F1 score, and Precision-Recall Distance (PRD), provide quantitative measures that indicate the accuracy and reliability of the CD models.

As mentioned above, accuracy is a measure of how close a value is to the true value, it provides an answer to the question, "Out of all the detected changes, how many were actually changes?" A model that exhibits high precision will result in fewer false alarms, demonstrating its strong capability to accurately identify true changes with minimal error. Recall, addresses the question of how many changes the model was able to detect among the actual changes in the data. Having a high recall means that the model is skilled at detecting the majority of changes, with fewer instances of missing actual changes. The F1 score represents a balanced measure of precision and recall, its score is especially useful when one aims to achieve a balance between precision and recall, which is frequently necessary in real-world scenarios where both false alarms and missed detections can have significant consequences. The PRD provides a geometric perspective on how well a model performs in the precision-recall space. The calculation involves measuring the distance between the precision and recall values of the model and the ideal point (100,100) using the Euclidean distance formula. The PRD captures the difference between a model's performance and perfection. A lower PRD suggests that the model is closer to achieving ideal performance. Together, these metrics offer a comprehensive evaluation of CD models, highlighting strengths and trade-offs in detecting true changes (recall) and ensuring the accuracy of detected changes (precision).

Figure 6 demonstrates that CD models exhibit optimal performance metrics when the distribution of the testing dataset matches that of the training dataset. Specifically, we observe that BIT, Changeformer, Tiny, and CSA-CDGAN attain the highest performance metrics on the LEVIR dataset, while SNUNet excels with the CDD dataset. When evaluating CD models on the WHU dataset, a notable decline in performance metrics is observed for SNUNet. Additionally, based on F1 scores, precision, and PRD (Predicted Relevance Density), Tiny models achieve the highest values, followed by Changeformer, BIT, and CSA-CDGAN, respectively. Moreover, CSA-CDGAN secures the top spot for Recall values, with BIT and Changeformer registering the lowest.

Testing on the CDD dataset reveals that SNUNet achieves superior performance metrics, highlighting a significant decrease in performance for models initially trained on the LEVIR dataset when applied to the CDD dataset. Despite the distribution differences between the two datasets, the BIT model exhibits relatively higher performance metrics compared to other CD models trained on the LEVIR dataset, which experienced a marked performance reduction. The performance metrics for all five CD models plummet when tested on the S2looking dataset, characterized by images collected at varying off-nadir angles. This outcome underscores the significant influence of camera setup and imagery

collection angles on CD model performance metrics, with our five CD model architectures failing to generalize effectively to such dataset characteristics.

On the cropland dataset, SNUNet displayed suboptimal performance metrics, despite being trained on a similar dataset. The F1, Recall, and PRD metrics for other CD models witnessed a complete decline. Interestingly, CSA-CDGAN outperformed others in terms of precision, followed by Tiny, whereas SNUNet recorded the lowest metrics.

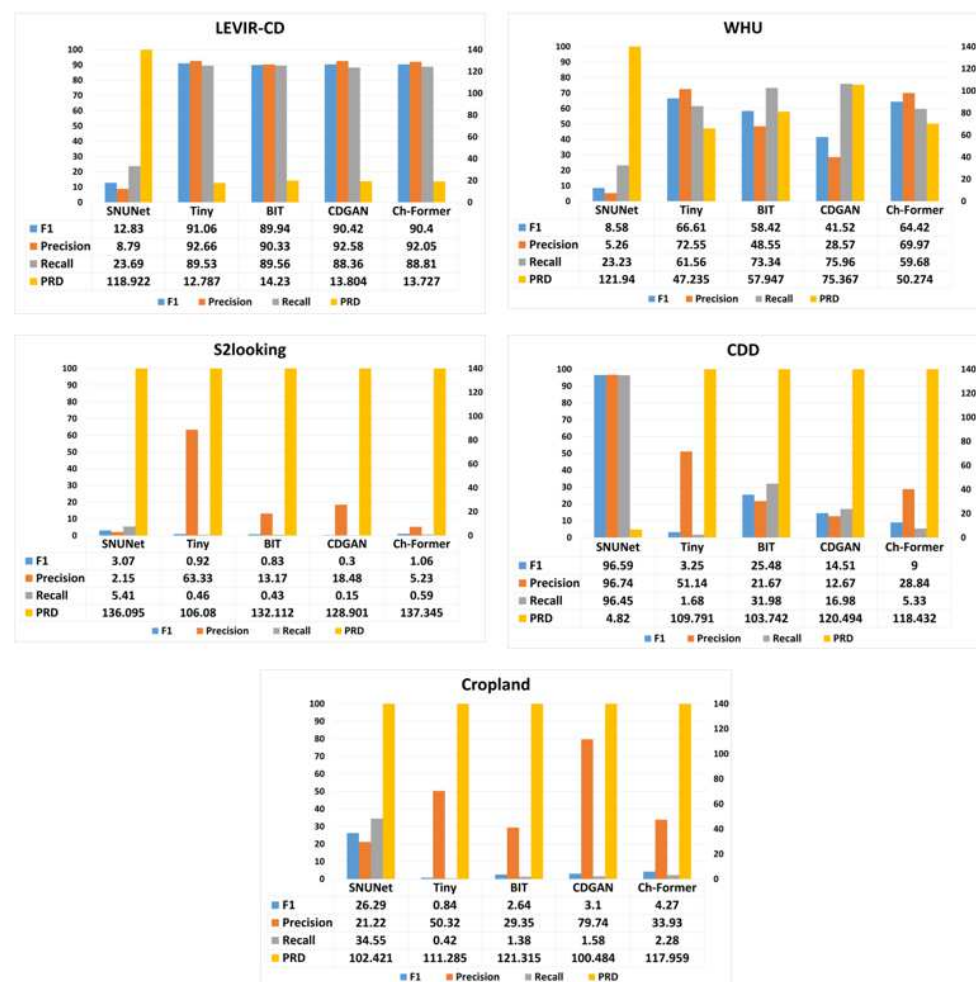


Figure 6. The CD models performance on the benchmark datasets, note that Ch-Former is a brief of Changeformer and CDGAN is a brief of CSA-CDGAN.

5.1. Contour-based Analysis Results

5.1.1. Performance Sensitivity Analysis Based on the Changed Contour Size

The majority of the contour sizes in the testing dataset ground truths fall within the range of [0 – 2100] pixels. Consequently, outlier sizes outside this range were excluded to concentrate the analysis on relevant data and enhance the statistical significance of our findings. Table 1 presents a descriptive summary of the testing dataset after outlier removal.

Table 1. Descriptive statistics of the testing dataset.

Parameter	Count	Mean	Std	Min	25%	50%	75%	Max
Size of contour (pixels)	1682	769	435	2	479	756	1039	2086

The analysis reveals significant variability in the sizes of changed contours, indicating the diverse complexity and structure within the test dataset instances. This diversity

impacts the performance metrics of our four CD models. Detailed insights are provided in Appendix 1.

Figure 7 visualizes the correlation between changed contour size and performance metrics of each CD model, illustrating variations in Precision, Recall, F1 Score, and Precision-Recall Distance (PRD) across different contour sizes.

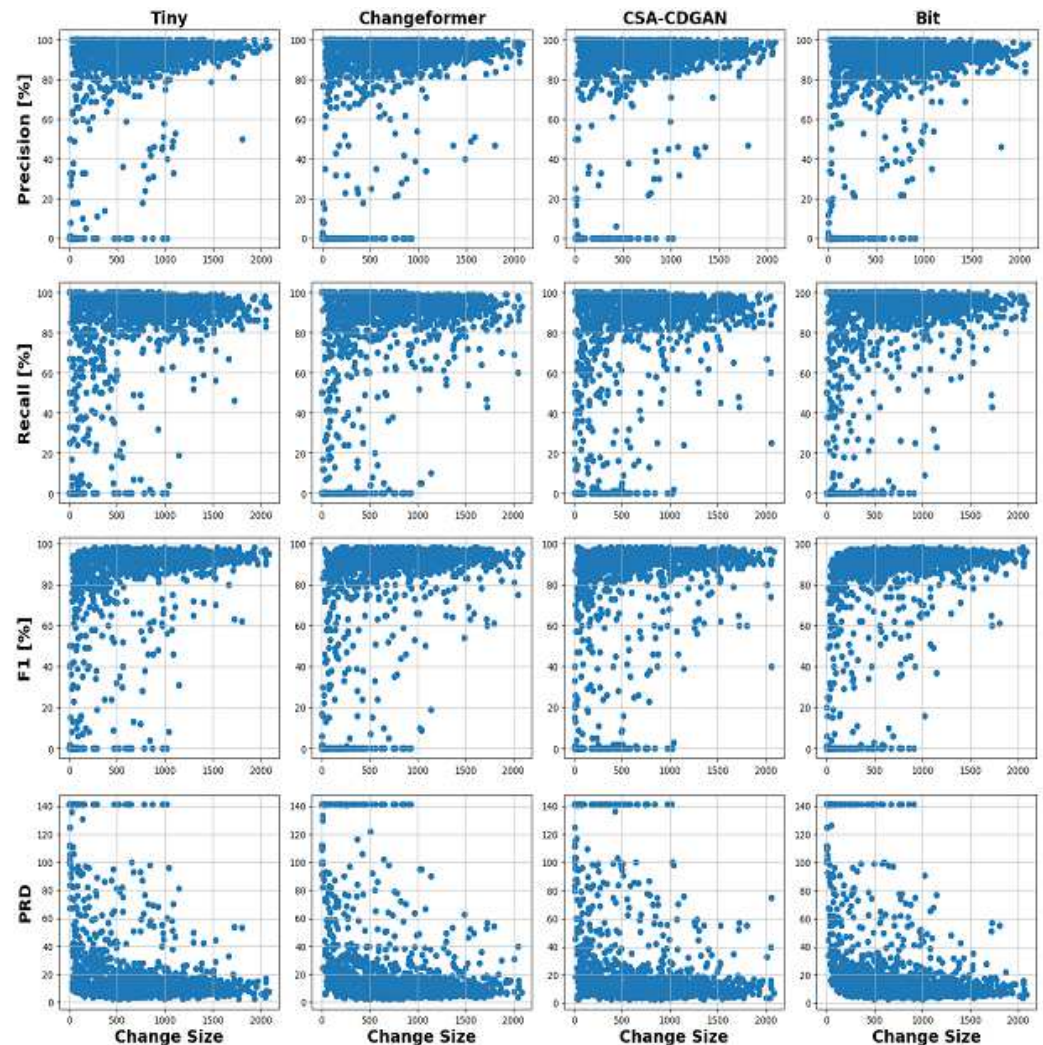


Figure 7. Scatter plots showing the relationships between the average change size and performance metrics of the CD models.

The scatter plots emphasize the relationship across the entire range of average change sizes, from 0 to 2100 pixels. For a more nuanced analysis, we divided the change sizes into three categories, as detailed in Table 2, based on data distribution.

Table 2. Categorization of change sizes based on the number of pixels of the changed contour.

Changed Contour Size	Number of Pixels
Small	0 - 600
Medium	600 - 1200
Large	1200 - 2100

The analysis across these size ranges reveals that precision tends to be higher for smaller changes, suggesting better model performance in less complex scenes. Conversely, recall increases with change size, indicating improved detection of larger changes. The F1

Score and PRD metrics also reflect this trend, with overall performance generally improving with larger change sizes.

The performance metrics generally improve with the increase in change size for all models. Tiny appears to provide the best balance between detecting changes accurately (precision) and detecting most changes (recall), as evidenced by the high F1 Scores. Changeformer demonstrates a reliable detection capability across different change sizes, with particularly strong recall for smaller changes. CSA-CDGAN and BIT show more variability in performance, which may indicate a dependency on certain characteristics of the change or the image conditions that were not consistently present across all change sizes. There is a clear trend that all models perform better in detecting larger changes compared to smaller ones. Understanding these patterns is crucial for improving CD models, particularly for applications where detecting small changes is vital, such as early detection of environmental or urban changes.

Subsequent linear regression models were fitted for each range to model the relationship between average change size and performance metrics (Precision, Recall, F1 Score, PRD), as depicted in Figure 8.

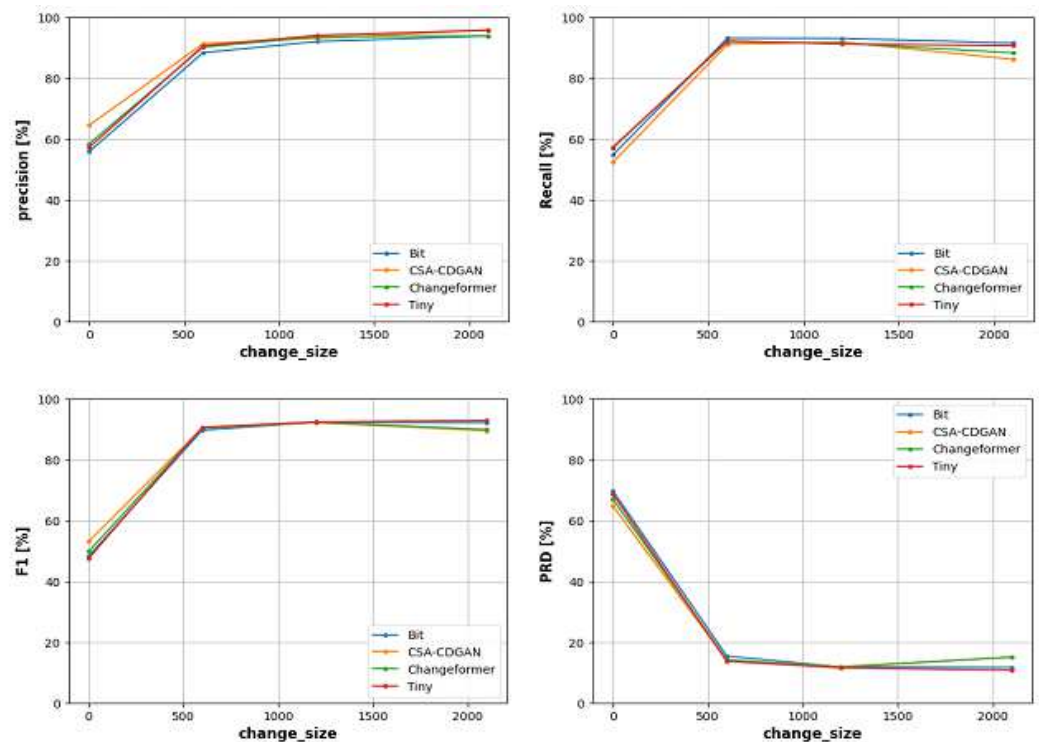


Figure 8. Concatenated linear regression models across the three change size ranges.

The precision improves significantly as the change size increases up to around 500 pixels, beyond which it plateaus. Similarly, recall shows improvement with increasing change size. The F1 Score and PRD values corroborate these findings, indicating that all models effectively detect larger changes. The initial variability for smaller changes highlights the challenges CD models face with nuanced changes.

6. Analysis and Comparison

When assessing the precision of CD models in relation to the size of identified changes, different models exhibit varied strengths. For small-sized changes, the CSA-CDGAN model shows the highest precision, highlighting its ability to accurately detect minimal changes. Changeformer ranks second in performance for this category, while the Bit model

exhibits the lowest precision, indicating potential challenges in accurately identifying smaller changes.

As change sizes grow to medium and large, the Tiny model demonstrates superior precision over its counterparts, suggesting its proficiency in maintaining accuracy as the complexity or extent of changes increases. CSA-CDGAN maintains a strong position with the second-highest precision for larger changes, while Bit continues to underperform, suggesting its lesser suitability for tasks requiring high precision across various change sizes.

6.1. Detection Capability of Challenging Contours Results

The analysis focuses on the models' ability to discern challenging contours—boundaries of changed areas within an image as shown in Figure 9, a crucial measure of effectiveness. The Tiny model leads with 209 detected contours, indicating its superior capacity for recognizing changes. Changeformer and Bit follow with 193 and 171 detections, respectively, with CSA-CDGAN identifying 170, positioning them in descending order of this metric.

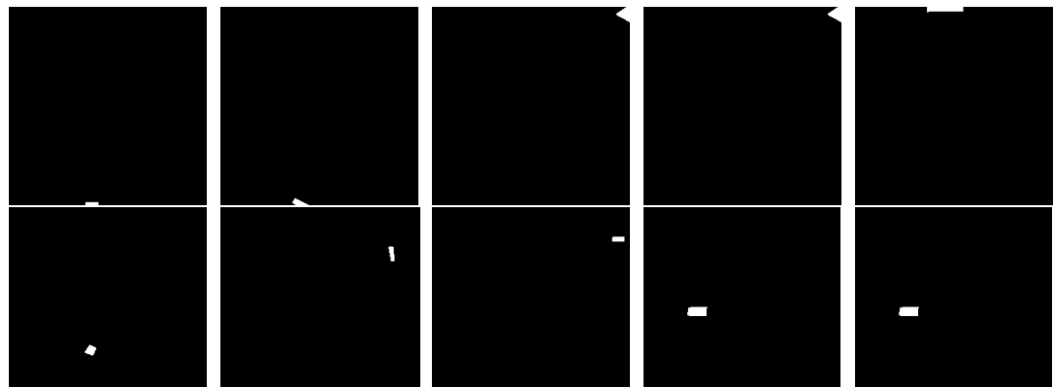


Figure 9. Samples of challenging contours in the LEVIR dataset.

6.2. Contour Count Comparison

Figure 10 illustrates the performance of various CD models in detecting contours compared to the ground truth. The x-axis represents the number of contours in the ground truth data, while the y-axis shows the number of contours predicted by the models. The 'Identity Line' symbolizes perfect alignment between ground truth and predictions. Models closer to this line predict contour numbers more accurately. Changeformer's predictions closely align with the Identity Line, indicating a high level of precision. Tiny shows slight deviation, suggesting it might merge or overlook separate instances of change. Bit and CSA-CDGAN exhibit more pronounced deviations, indicating tendencies to underestimate or overestimate the number of contours, respectively.

Changeformer emerges as the most accurate model in predicting the number of contours present, aligning closely with the actual numbers (ground truth), showcasing its effectiveness in CD tasks.

6.3. Robustness Analysis with Respect to Different Image Corruptions Results

Table 3 describes the characteristics of three noisy datasets used in our study. Each noisy dataset was constructed by considering the PSNR range. Also, to ensure the accuracy and comprehensiveness of our analysis, the number of instances in each noisy dataset varies, covering the respective PSNR range as extensively as possible.

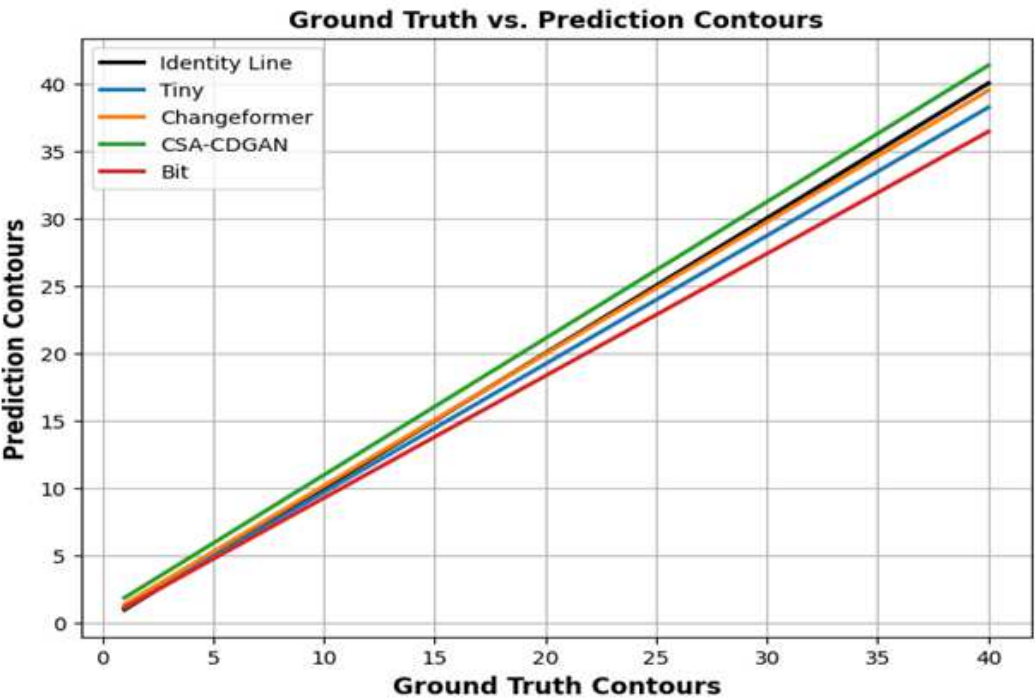


Figure 10. Ground Truth vs. Prediction Contours.

Table 3. Description of the three noisy datasets.

Noisy Dataset	Number of Instances	Range of PSNR (dB)
Gaussian LEVIR	2919	5 – 35
Salt–Pepper LEVIR	1849	30 – 55
Speckle LEVIR	6179	30 – 35

6.3.1. LEVIR with Gaussian Noise Dataset

The robustness analysis, particularly with Gaussian noise, tests the performance of CD models under image quality degradation. Gaussian noise, a common issue in image processing, can significantly affect algorithm performance.

Figure 11 displays examples from the LEVIR dataset corrupted with Gaussian noise at various intensity levels, marked by the PSNR range of [5 – 35]. High PSNR indicates better image quality (less noise).

Table 4 and Table 5 showcase the baseline results on the original LEVIR dataset and the performance on the Gaussian LEVIR dataset, respectively.

Table 4. Baseline results on the original LEVIR sample dataset.

CD model	F1	Precision	Recall	PRD
Tiny	85.17	89.06	82.79	20.67
Changeformer	85.36	90.18	82.54	20.28
CSA-CDGAN	78.45	84.00	76.29	30.30
BIT	85.14	87.17	84.06	20.78

Table 6 quantifies the drop in performance after adding Gaussian noise, highlighting how noise impacts each model to varying degrees. This further emphasizes that while all models are affected by noise, their ability to maintain performance varies. Changeformer generally shows the least impact on precision and F1 score, and Bit maintains recall robustly. Therefore, the analysis demonstrates that Changeformer has the best robustness



Figure 11. Samples from the Gaussian LEVIR dataset with added Gaussian noise.

Table 5. CD model’s performance on the Gaussian LEVIR dataset.

CD model	F1	Precision	Recall	PRD
Tiny	83.3	77.82	89.76	24.5
Changeformer	87.25	85.21	89.5	18.28
CSA-CDGAN	76.53	79.96	77.67	32.59
Bit	84.27	78.67	90.86	23.28

against noise in terms of maintaining high F1 and low PRD scores, while Bit exhibits the strongest recall. The variability in the models’ performance underlines the need for CD models capable of withstanding image quality issues commonly encountered in real-world scenarios.

Table 6. Drop in performance after adding Gaussian noise.

CD model	F1 Drop	Precision Drop	Recall Increase	PRD Increase
Tiny	-1.87	-11.24	6.97	3.83
Changeformer	1.89	-4.97	6.96	-2
CSA-CDGAN	-1.92	-4.04	1.38	2.29
BIT	-0.87	-8.5	6.8	2.5

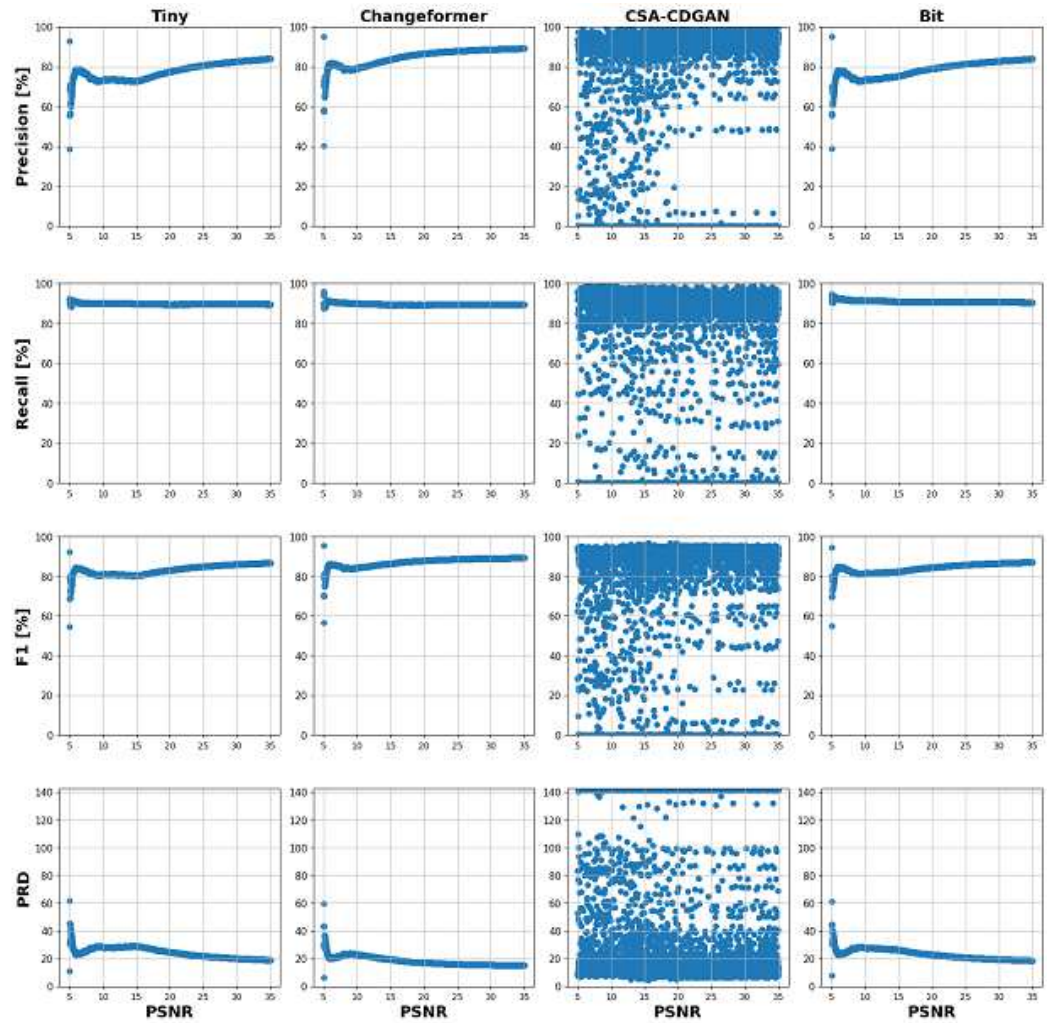


Figure 12. Scatter plots showing the relationship between PSNR and performance metrics of the CD models.

Figure 12 and Figure 19 illustrate the relationship between PSNR levels and CD model performance metrics, indicating variable performance across PSNR levels. Figure 12 depicts a wide distribution of results, particularly for the CSA-CDGAN model, indicating variable performance across different levels of PSNR. This variability suggests that the model's ability to accurately detect changes in the presence of noise is inconsistent.

Figure 13 offers a more aggregated view of performance across different PSNR levels. The line charts allow us to compare the overall trend of each model's performance as the noise level changes. Detailed observations based on the tables and figures are as follows: Across the noise intensity range of [5 - 10] dB, Changeformer consistently outperforms the other models, followed closely by Bit and Tiny, which show similar results. CSA-CDGAN struggles the most in this aspect. The Bit model consistently provides the best recall values, indicating its proficiency in identifying all relevant instances of change in the noisy images. Changeformer and Tiny show competitive performance, while CSA-CDGAN again lags behind. The F1 score, a harmonic mean of precision and recall, signifies a balance between the two. Changeformer leads in F1 score across the entire PSNR range, demonstrating an effective balance between precision and recall despite the noise. BIT also shows strong performance, while CSA-CDGAN records the lowest F1 scores. This metric represents the distance from the perfect score in the precision-recall space, where a lower PRD indicates better performance. Changeformer has the lowest (and thus best) PRD scores, especially in

higher noise scenarios, followed by Bit. CSA-CDGAN consistently has the highest (worst) PRD values.

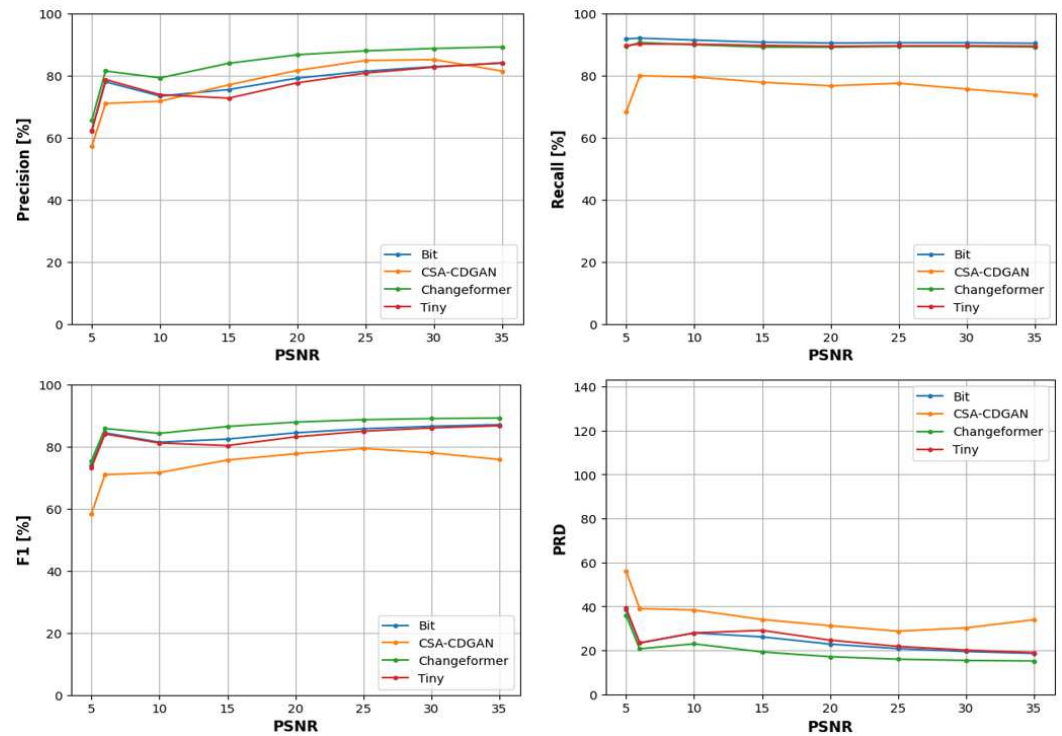


Figure 13. Performance of CD models across various PSNR ranges.

The analysis demonstrates the variability in models' robustness to noise, with Changeformer generally showing the least impact on precision and F1 score, and Bit maintaining robust recall. The performance trends across PSNR ranges provide insights into each model's efficacy under different noise conditions, emphasizing the importance of model selection based on specific application requirements.

6.3.2. LEVIR with Salt and Pepper Noise Dataset

Figure 14 showcases samples from the LEVIR dataset that have been modified by introducing 'salt and pepper' noise. This type of noise, characterized by random pixels being set to black or white, presents a significant challenge for algorithms by disrupting image content and complicating CD.

The PSNR range of [30 – 55] indicates the quality of these images, with higher values denoting better quality (less noise).

Table 7. CD model's performance on salt & pepper LEVIR.

CD model	F1	Precision	Recall	PRD
Tiny	78.08	82.2	74.75	31.04
Changeformer	84.76	80.29	90.02	22.27
CSA-CDGAN	75.73	79.37	77.03	33.57
Bit	82.91	76.01	91.45	25.55

In Table 7, we summarize the performance of four CD models on the salt and pepper noise-corrupted LEVIR dataset using key metrics such as F1 Score, Precision, Recall, and PRD. Changeformer demonstrates a well-balanced performance between precision and recall despite the noise, indicated by its high F1 score and PRD. The Bit model excels in recall, suggesting its effectiveness at identifying relevant instances but with more false



Figure 14. Samples of the LEVIR testing dataset with salt and pepper noise.

Table 8. Drop in performance after adding salt & pepper noise.

CD model	F1 Drop	Precision Drop	Recall Increase	PRD Increase
Tiny	-7.09	-6.86	-8.04	10.37
Changeformer	-0.6	-9.89	7.48	1.99
CSA-CDGAN	-2.72	-4.63	0.74	3.27
Bit	-2.23	-11.16	7.39	4.77

positives, as evidenced by its lower precision. Tiny, while scoring highest in precision, has the lowest recall, indicating a conservative prediction approach. CSA-CDGAN shows struggle across all metrics, indicating a greater challenge with this type of noise.

Table 8 highlights the drop in performance due to added noise. Changeformer and Bit exhibit stronger robustness by maintaining or even improving their recall. Tiny maintains the highest precision, potentially at the expense of missing some changes, while CSA-CDGAN is most affected, struggling to balance precision and recall under noisy conditions. These findings underline the importance of evaluating CD models under various noise conditions to ensure real-world applicability.

Figure 15 demonstrates the relationship between the PSNR and the performance of four CD models across various noise conditions. PSNR, a measure of image quality degradation due to compression or noise, indicates better image quality with higher values.

To analyze the relationship, we divided the PSNRs into five ranges: [30 – 35], [35 – 40], [40 – 45], [45 – 50], and [50 – 55] dB. For each range, linear regression models were fitted to understand the PSNR's influence on four performance metrics. Each row in the figure corresponds to a different performance metric for the CD models Tiny, Changeformer, CSA-CDGAN, and Bit, with each dot representing a different image in the noisy dataset. In the analysis of CD model performance across varying PSNR ranges, the F1 score, which combines precision and recall, shows that Changeformer consistently outperforms the other models between 30 to 40 dB, and continues to lead up to 55 dB. The Bit model ranks second in performance, indicating its robustness. However, Tiny struggles as noise increases, reflected by its lower F1 scores.

When focusing on precision alone, Tiny excels in the higher PSNR range of 30 to 35 dB, suggesting it can accurately predict changes when noise levels are lower. However, its precision advantage diminishes as noise increases. In contrast, CSA-CDGAN shows the least consistent precision, especially at lower PSNR levels (higher noise). Interestingly, the performance dynamics shift in the 40 to 45 dB range, where CSA-CDGAN steps up as the

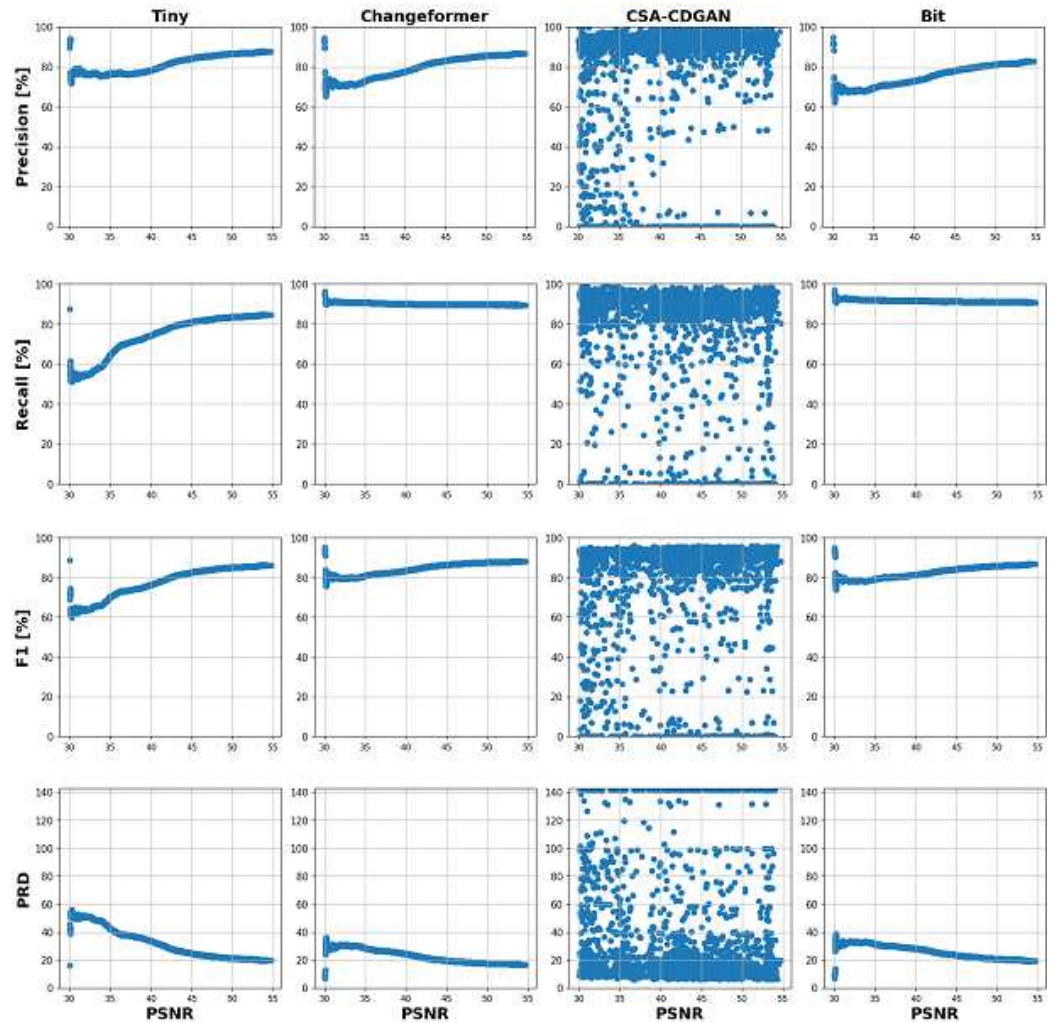


Figure 15. Scatter plots showing the relationship between PSNR and performance metrics of the CD models.

leading model, followed by Tiny. But as noise continues to increase beyond 45 dB, Tiny regains its top position with the least precision drop, maintaining it through to the 50 to 55 dB range.

For recall, the Bit model demonstrates superior performance in correctly identifying relevant instances of change from 30 to 45 dB, maintaining this lead as the PSNR range extends to 55 dB. Changeformer follows Bit closely, while CSA-CDGAN presents the lowest recall, indicating challenges in detecting true positives amidst higher noise levels.

The PRD metric, which measures the combined deviation of precision and recall from their ideal values, shows Changeformer achieving the closest proximity to ideal performance, followed by Bit in the 30 to 40 dB range. As noise becomes more prominent, Changeformer continues to maintain the lead in PRD, suggesting its predictions remain relatively unaffected by the increase in noise levels.

These observations underscore the importance of choosing the right CD model based on the noise characteristics of the dataset. Changeformer appears to be a robust choice across most noise conditions, while Bit is noteworthy for its recall. Tiny, although excelling in precision at lower noise levels, requires careful consideration due to its performance variability with increasing noise.

Figure 16 visualizes the performance of the CD models across different PSNR ranges, as determined by linear regression modeling.

- **Precision:** Across most models, precision improves slightly with higher PSNR (less noise), suggesting variable impact on the accuracy of positive predictions.
- **Recall:** Tends to increase with higher PSNR for most models, implying better detection of all relevant instances of change as image quality improves.
- **F1 Score:** Remains relatively stable across PSNR values for most models, indicating balanced performance despite varying noise levels.
- **PRD:** Lower PRD values near higher PSNR levels suggest models' predictions are closer to ideal precision and recall values with better image quality.

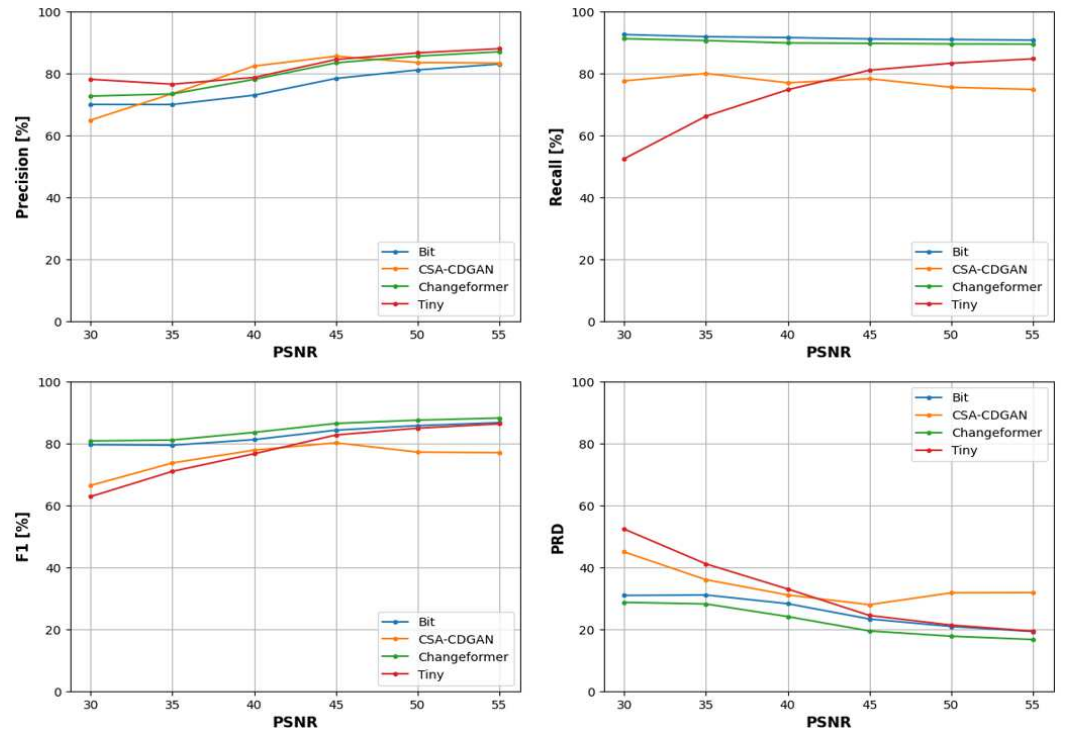


Figure 16. Performance of CD models across various PSNR ranges.

Changeformer consistently outperforms other models in F1 score from 30 to 55 dB, indicating robustness. Tiny excels in precision at lower noise levels but varies more at higher noise levels. Bit demonstrates superior recall performance, underscoring its effectiveness in identifying relevant changes even in noisier images. Changeformer is notably proficient in balancing precision and recall across different noise levels. These analyses reveal that Changeformer generally offers the best performance, particularly in terms of F1 and PRD scores. Tiny, while showing excellence in precision at lower noise levels, struggles more as noise increases. CSA-CDGAN maintains a uniform performance but generally underperforms compared to Changeformer and Bit. These insights are crucial for developing robust CD models that are resilient to the variations in image quality often found in real-world scenarios.

6.3.3. LEVIR with Speckle Noise Dataset

Figure 17 presents samples from the LEVIR dataset with speckle noise introduced, covering a PSNR range of 29.35 to 39 dB. This noise, characterized by granular interference, poses challenges for CD algorithms due to its disruptive speckled pattern.

The performance metrics of CD models on this speckle noise-affected dataset are summarized in Table 9. Following the introduction of speckle noise, the metrics reveal interesting shifts as shown in Table 10, indicating a tendency of CD models to maintain or



Figure 17. Samples of the LEVIR testing dataset with Speckle Noise.

improve their sensitivity to actual changes in noisier environments, despite a significant loss in precision.

Table 9. CD model’s performance metrics on the speckle LEVIR dataset.

CD model	F1	Precision	Recall	PRD
Tiny	74.46	63.85	89.5	37.66
Changeformer	75.54	66.14	88.26	35.86
CSA-CDGAN	72.12	70.3	82.83	37.32
Bit	74.6	63.52	90.56	37.69

Table 10. Drop in performance after adding speckle noise.

CD model	F1 Drop	Precision Drop	Recall Increase	PRD Increase
Tiny	-10.71	-25.21	6.71	16.99
Changeformer	-9.82	-24.04	5.72	15.58
CSA-CDGAN	-6.33	-13.7	6.54	7.02
Bit	-10.54	-23.65	6.5	16.91

These results highlight the resilience of these models to some extent, with BIT and Changeformer showing noteworthy robustness in recall, and CSA-CDGAN maintaining the highest precision despite the noise challenge. The trade-off between detecting changes (recall) and avoiding false alarms (precision) is crucial in noisy conditions, underlining the importance of evaluating CD models across various noise scenarios for real-world application readiness.

Figures 18 and Figure 19 depict scatter plots and line charts that illustrate the performance of various CD models across different PSNR ranges, quantifying the level of noise in images from the LEVIR dataset. Based on the data distribution, we divided the PSNRs into four ranges: [29.35 - 29.5], [29.5 - 30], [30 - 31], and [31 - 39] dB. For each PSNR range, four linear regression models were fitted to model the relationship between the PSNR and the four performance metrics, as demonstrated in Figure 19.

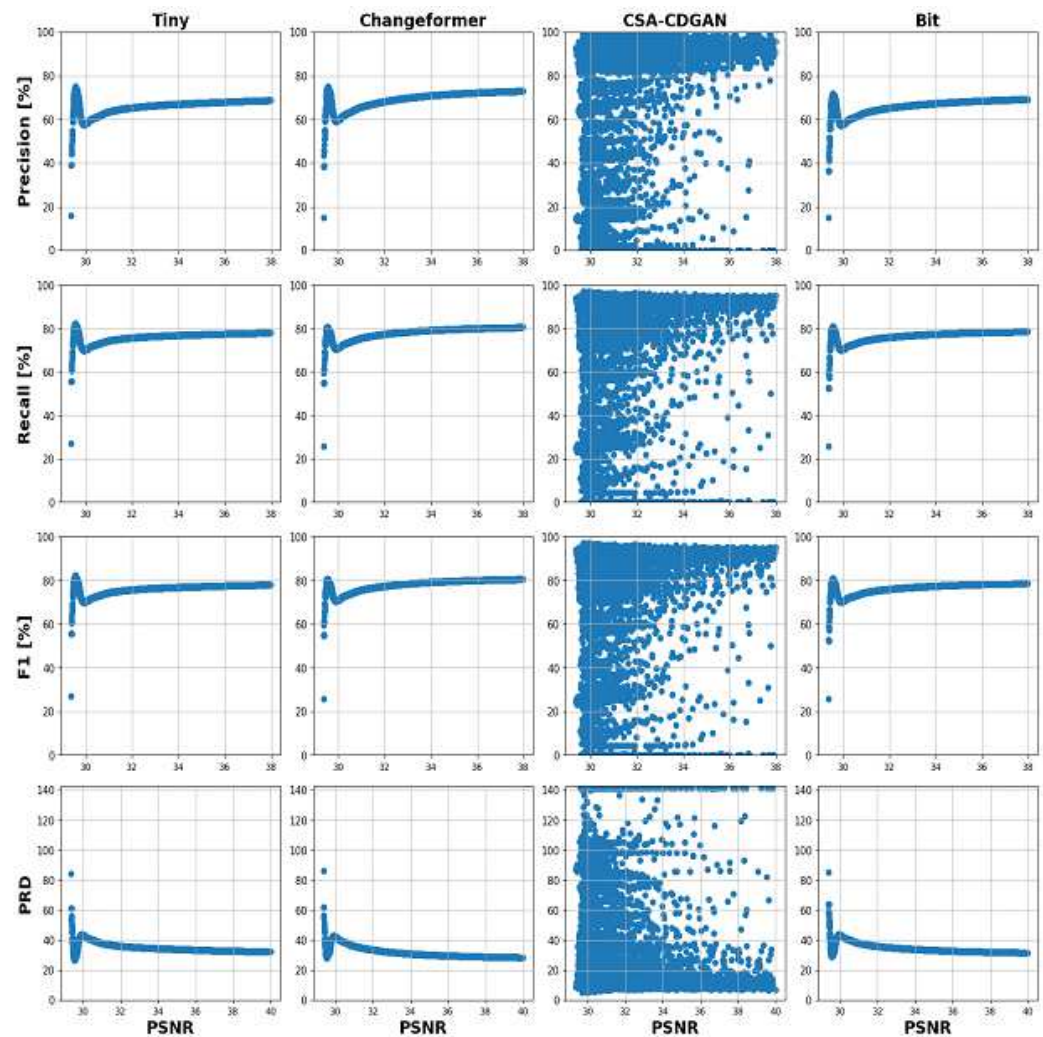


Figure 18. Scatter plots showing the relationship between PSNR and performance metrics of the CD models.

Analysis Across Specified PSNR Ranges

F1 Score

- At PSNRs between 29.35 and 29.5 dB, amidst the highest noise intensity, CSA-CDGAN leads with superior F1 scores, suggesting robustness against noise. Tiny is close behind, while BIT lags with the lowest scores.
- Within the PSNR range of 29.5 to 30 dB, Tiny excels, indicating effectiveness in moderately noisy environments. BIT and Changeformer are competitive, with CSA-CDGAN falling behind.
- For PSNRs from 30 to 31 dB, Changeformer surpasses all models, efficient at handling this specific noise level. BIT and Tiny remain close contenders.
- When PSNRs span 31 to 39 dB, CSA-CDGAN consistently achieves the highest F1 scores, denoting its capacity to maintain accuracy over a wide noise range. Tiny trails with lower F1 scores.

Precision

- In the highest noise bracket, CSA-CDGAN's precision is unrivaled, indicating fewer false positives under severe conditions. BIT exhibits the least precision.
- From PSNRs of 30 to 39 dB, CSA-CDGAN remains the precision leader, followed by Changeformer. Tiny and BIT show less precision.

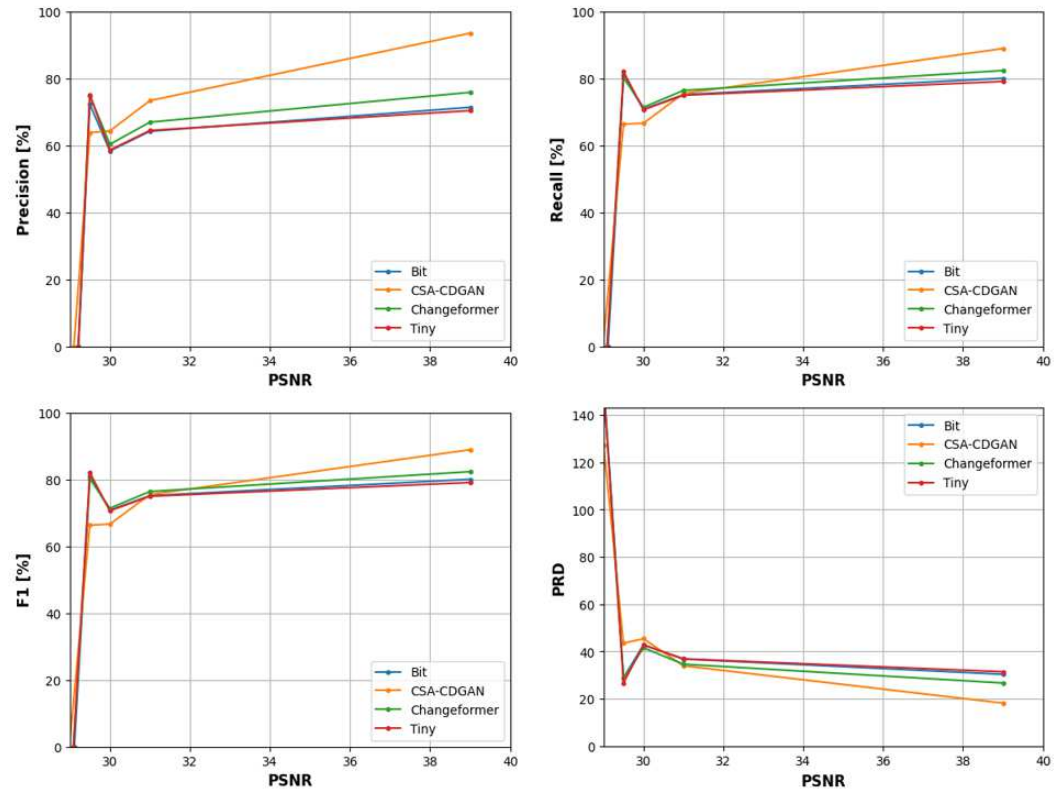


Figure 19. Performance of CD models across various PSNR ranges.

Recall

- In the range of 29.35 to 29.5 dB, Tiny and Changeformer exhibit the best recall, with BIT showing the lowest, potentially missing some true changes.
- From 31 to 39 dB, CSA-CDGAN surpasses others in recall, indicating consistent CD across this noise spectrum.

PRD (Precision-Recall Distance)

- At PSNRs of 29.35 to 29.5 dB, CSA-CDGAN showcases the best PRD, indicating a closer approximation to ideal values. BIT has the highest PRD, suggesting greater deviation.
- For PSNRs from 31 to 39 dB, CSA-CDGAN achieves the best PRD, with Tiny showing the highest PRD, indicating performance furthest from the ideal.

These analyses highlight that while noise levels universally impact CD model performance, effects are nuanced. CSA-CDGAN and Changeformer demonstrate robustness, especially in maintaining a balance between precision and recall, while Tiny and BIT experience more significant performance fluctuations. Understanding model behavior in noisy conditions is crucial for selecting and optimizing CD models for practical applications where image quality varies.

7. Discussion

To demonstrate the generalization performance of the models across significantly different datasets, we analyze how well the selected models—evaluated using the proposed framework—perform when tested on datasets that differ in terms of change types, environmental conditions, and image resolutions. The selected datasets span a variety of domains, including urban environments (building changes) and rural/agricultural landscapes (cropland and seasonal changes), and also vary widely in their image resolutions

and sensor characteristics. This section presents a detailed analysis of how the models performed across these diverse datasets.

With regards to the performance on urban and rural datasets, these three datasets LEVIR-CD, WHU-CD, and S2Looking primarily focus on detecting changes related to building construction or modification in urban areas. The high-resolution images in LEVIR-CD (0.5 meters per pixel) and WHU-CD (0.075 meters per pixel) are ideal for detecting fine-grained changes in building structures. Models that show high performance on these datasets need to effectively capture small, detailed changes. For example, the CSA-CDGAN model demonstrated consistent performance across these urban datasets, showcasing its robustness in handling high-resolution data and small-scale changes. This result highlights its ability to generalize to urban environments with varying building structures and sizes.

Rural and agricultural CD was tested using CDD and CLCD datasets that cover seasonal changes and cropland shifts and introduce different types of changes, such as vegetation transformations and agricultural modifications. The datasets vary from high-resolution imagery (CDD, 3 cm per pixel) to lower resolutions (CLCD, 2 meters per pixel). Models like Tiny and Changeformer, which performed well in more controlled, high-resolution urban environments, struggled in rural settings with coarser resolutions, where change detection is more complex due to larger, less defined features. However, CSA-CDGAN performed remarkably well across both urban and rural datasets, showing its flexibility in handling a range of real-world change types. This demonstrates its strong generalization capability across vastly different environments, from the detailed structures of urban areas to the broader landscapes of agricultural fields.

In terms of cross-dataset generalization, the framework's evaluation also tested the models across different image resolutions, from the fine details in the LEVIR-CD dataset (0.5 meters per pixel) to the relatively coarse details in CDD and CLCD (ranging from 0.03 meters per pixel to 2 meters per pixel). The performance of models like Changeformer, which showed higher adaptability to noise and resolution changes, demonstrated that it can generalize across datasets with different resolution qualities, making it useful in practical applications where satellite image resolutions vary.

Moreover, the framework's ability to evaluate models at multiple PSNR levels revealed significant differences in model performance. For instance, CSA-CDGAN outperformed others in lower-resolution datasets, especially in the CLCD dataset with a 2-meter resolution, showing its robustness to noise and resolution variations. In contrast, models like BIT and Tiny exhibited a steep decline in performance as the image resolution decreased, highlighting their reliance on high-resolution inputs. This behavior reinforces the importance of using a generalizable evaluation framework that can simulate real-world noise and resolution conditions across various datasets.

In the case of cross-seasonal and environmental CD, the CDD dataset's seasonal variability, with changes occurring across different environmental conditions, presents another challenge. For example, vegetation changes in one season might be more difficult to detect than construction in urban environments. The framework's sensitivity analysis, which examined how models responded to varying conditions (e.g., seasonal variations and image quality degradation), demonstrated that CSA-CDGAN and Changeformer handled these seasonal and environmental changes more robustly than others. This suggests that these models have been trained or designed to generalize better across diverse environmental contexts, an important factor for deployment in practical settings where conditions can change significantly.

Finally, with regards to the model robustness across diverse image corruptions, the robustness analysis incorporated into the framework tested how well each model handled image corruptions, such as noise and blurring, which commonly occur in real-world satel-

lite and aerial imagery. The performance of CSA-CDGAN was particularly impressive in this regard, consistently achieving high precision, recall, and F1 scores across datasets with varying levels of image quality degradation. This is a strong indicator of the model’s generalization ability, as it was able to maintain consistent performance even under challenging conditions like heavy noise or low-resolution inputs, common in operational scenarios.

8. Conclusions

This study explores the complex relationship between the effectiveness of CD models, the accuracy of aerial and satellite imagery, and the ability to withstand different sources of noise. Deep learning in the field of land cover CD has produced a variety of models, each claiming to have excellent performance metrics. Nevertheless, the increasing number of models poses a challenge when it comes to selecting the right one, especially due to the lack of widely available benchmarks for thorough performance analysis and comparison. We presented and explained a flexible, scalable framework designed for the systematic assessment and comparison of CD models. This innovative framework is constructed using three interconnected pipelines—testing across different datasets, evaluating under challenging conditions, and conducting a detailed analysis of performance sensitivity based on the magnitude of changes. The utilization of this framework on five modern CD models—Changeformer, BIT, Tiny, SNUNet, and CSA-CDGAN—highlighted its effectiveness in identifying the strengths and weaknesses of these models. The findings from this exercise are crucial for evaluating and comparing these models, providing valuable information about their individual strengths and weaknesses.

The evaluation framework’s ability to test models across a broad range of datasets—from urban to rural, from high to low resolution, and under varying environmental conditions—provides strong evidence of the generalization performance of the models. CSA-CDGAN and Changeformer stood out due to their robustness across these diverse conditions. These results demonstrate the importance of using a framework that rigorously tests models across a range of real-world scenarios, ensuring that the selected models will perform well in various practical applications, from urban change detection to agricultural monitoring. The framework’s adaptability and comprehensive evaluation allow for a clear understanding of model strengths and weaknesses across different domains, highlighting its significance for model selection in real-world CD tasks.

The proposed framework is a significant advancement in CD research, offering the potential to greatly increase the practicality of CD models. This work sets the stage for future research to build upon, offering a thorough comparison of existing CD models and establishing a standard for developing strong frameworks. Future avenues of research invite further investigation into the integration of these findings with wider computer vision tasks and the development of innovative models that embody the combined strengths identified and it has the potential to greatly advance the field of CD technologies.

Author Contributions: A.A.A.H, M.B.I., and A.S.H contributed to the conceptualization, methodology, and led the software development, validation, and formal analysis of the research. H.I.A. A.M.R.A., A.A.H, M.E prepared the original draft of the manuscript and also contributed to the writing, review, and editing of the manuscript. H.I.A., A.A.H, and M.E were in charge of supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Data will be made available to the editors of the journal for review or query upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CD	Change Detection	
DL	Deep Learning	
SOTA	State of the Art	
PSNR	Peak Signal-to-Noise Ratio	
CNN	Convolutional Neural Network	
RNN	Recurrent Neural Network	
GAN	Generative Adversarial Network	
AE	Autoencoder	
PCA	Principal Component Analysis	
CVA	Change Vector Analysis	1027
MSA	Modified Self-Attention	
GL-ASPP	Gated Linear Atrous Spatial Pyramid Pooling	
IoU	Intersection over Union	
PRD	Precision-Recall Distance	
LCCD	Land Cover Change Detection	
LEVIR-CD, WHU-CD, S2looking, CDD, CLCD	Names of datasets used for testing	
BIT, Tiny, SNUNet, CSA-CDGAN, Changeformer	Names of the CD models evaluated	

Appendix A Additional resources for performance sensitivity analysis based on the contour size

Across these metrics, Figure A1 illustrates the performance trade-offs of each model. In lower PSNR ranges, some models may maintain higher precision but at the cost of recall, or vice versa. The trend lines in the graphs also show how each model’s performance changes as the PSNR increases, providing insights into their robustness against noise. From the line graphs, it is apparent that:

- The Changeformer model generally shows a robust performance across all metrics, maintaining a high F1 score even as PSNR varies, suggesting that it achieves a good balance between precision and recall across different levels of image quality.
- The Tiny model excels in the precision metric at high PSNR levels, indicating its effectiveness in correctly labeling changed pixels in higher-quality images.
- The CSA-CDGAN model demonstrates varying performance, with notable dips in certain PSNR ranges, implying sensitivity to image quality for maintaining detection accuracy.
- The BIT model shows consistent recall across most PSNR levels but with variations in precision, which may suggest it is better at identifying all relevant changes at the cost of including more false positives.

The graphs indicate that the ideal choice of model depends on the specific requirements of the CD task, such as whether precision or recall is more critical, and the expected quality of the images being processed. Decision-makers can use these insights to select a model that aligns best with their operational context and the type of aerial or satellite imagery they will be working with.

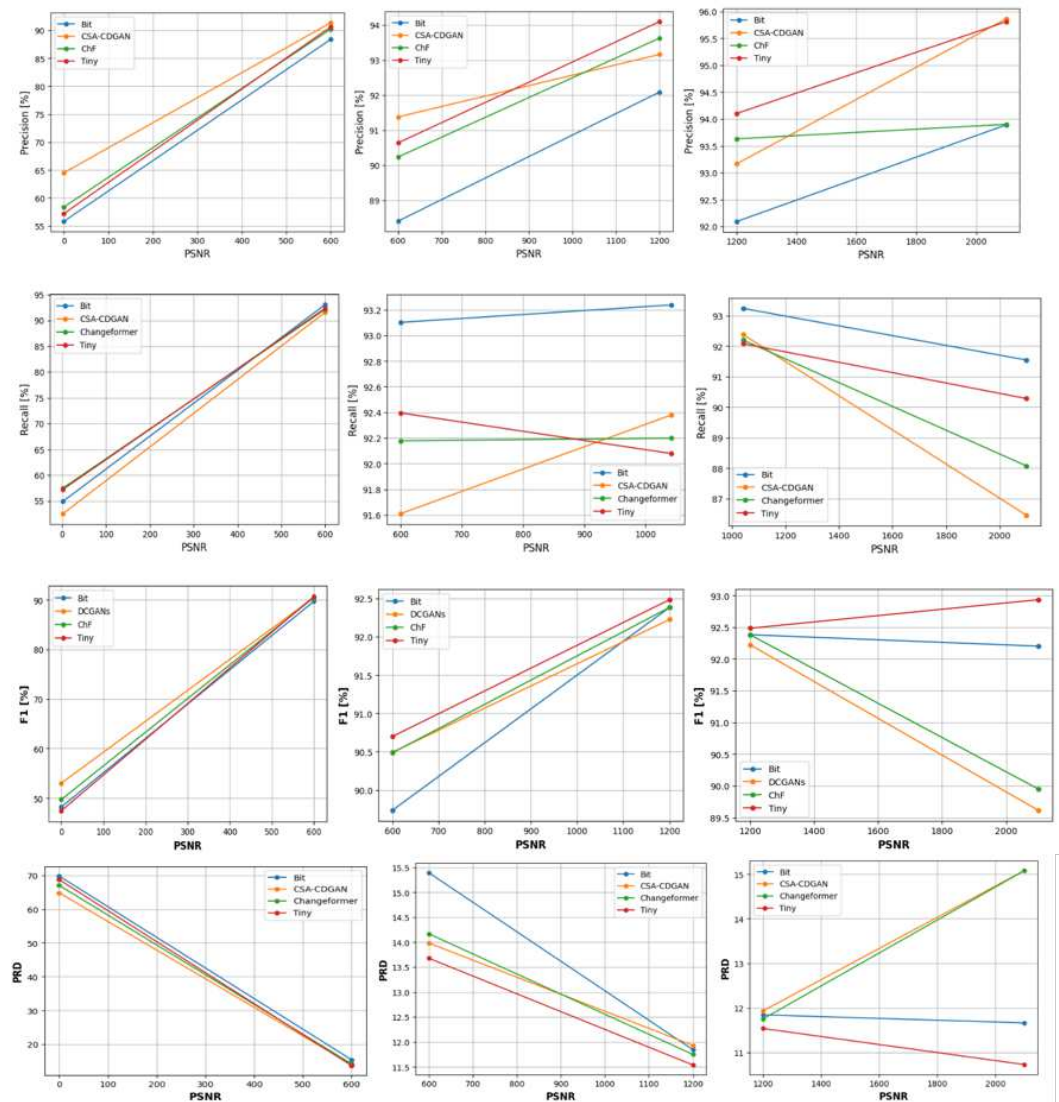


Figure A1. CD Models—BIT, CSA-CDGAN, Changeformer (ChF), and Tiny—are evaluated on their Precision, Recall, F1 Score, and PRD at varying PSNR levels.

References

1. Nafaa, S., Ashour, K., Mohamed, R., Essam, H., Emad, D., Elhenawy, M., Ashqar, H., Hassan, A. & Alhadidi, T. Automated Pavement Cracks Detection and Classification Using Deep Learning. *2024 IEEE 3rd International Conference On Computing And Machine Intelligence (ICMI)*. pp. 1-5 (2024)
2. Institute of Electrical and Electronics Engineers; IEEE Geoscience and Remote Sensing Society. *2010 IEEE International Geoscience & Remote Sensing Symposium: Proceedings, July 25-30, 2010, Honolulu, Hawaii, U.S.A.*; IEEE: 2010.
3. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens* **2020**, 12(10). doi: 10.3390/rs12101662.
4. El-Zeiny, A.M.; Effat, H.A. Environmental monitoring of spatiotemporal change in land use/land cover and its impact on land surface temperature in El-Fayoum governorate, Egypt. *Remote Sens Appl* **2017**, 8, 266–277. doi: 10.1016/j.rsase.2017.10.003.
5. Mohamed, R., Esam, H., Nafaa, S., Ashour, K., Emad, D., Elhenawy, M., Ashqar, H., Hassan, A. & Glaser, S. Deep Learning-Based pavement defect detection. *Australasian Road Safety Conference, 2023, Cairns, Queensland, Australia*. (2023)
6. Ke, L.; Lin, Y.; Zeng, Z.; Zhang, L.; Meng, L. Adaptive Change Detection with Significance Test. *IEEE Access* **2018**, 6, 27442–27450. doi: 10.1109/ACCESS.2018.2807380.
7. Luppino, L.T.; Bianchi, F.M.; Moser, G.; Anfinsen, S.N. Unsupervised Image Regression for Heterogeneous Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* **2019**. doi: 10.1109/TGRS.2019.2930348.
8. Liu, S.; Bruzzone, L.; Bovolo, F.; Zanetti, M.; Du, P. Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing* **2015**, 53(8), 4363–4378. doi: 10.1109/TGRS.2015.2396686.

9. Johnson, R.D.; Kasischke, E.S. Change vector analysis: A technique for the multispectral monitoring of land cover and condition. *International Journal of Remote Sensing* **1998**, *19*(3), 411–426. doi: 10.1080/014311698216062. 1070
10. Celik, T. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geoscience and Remote Sensing Letters* **2009**, *6*(4), 772–776. doi: 10.1109/LGRS.2009.2025059. 1071
11. Massarelli, C. Fast detection of significantly transformed areas due to illegal waste burial with a procedure applicable to landsat images. *International Journal of Remote Sensing* **2018**, *39*(3), 754–769. doi: 10.1080/01431161.2017.1390272. 1072
12. Vázquez-Jiménez, R.; Romero-Calcerrada, R.; Novillo, C.J.; Ramos-Bernal, R.N.; Arrogante-Funes, P. Applying the chi-square transformation and automatic secant thresholding to Landsat imagery as unsupervised change detection methods. *Journal of Applied Remote Sensing* **2017**, *11*(1), 016016. doi: 10.1117/1.JRS.11.016016. 1073
13. Raja, R.A.A.; Anand, V.; Kumar, A.S.; Maithani, S.; Kumar, V.A. Wavelet Based Post Classification Change Detection Technique for Urban Growth Monitoring. *Journal of the Indian Society of Remote Sensing* **2013**, *41*(1), 35–43. doi: 10.1007/s12524-011-0199-7. 1074
14. Luque, S.S. Evaluating temporal changes using Multi-Spectral Scanner and Thematic Mapper data on the landscape of a natural reserve: the New Jersey Pine Barrens, a case study. *International Journal of Remote Sensing* **2000**. Available online: <http://www.tandf.co.uk/journals> (accessed on [date]). 1075
15. Kristof, S.J.; Scholz, D.K.; Anuta, P.E.; Momin, S.A. R.A. WEISMILLER Change Detection in Coastal Zone Environments. Four techniques were used to analyze Landsat MSS temporal data in order to detect areas of change of the Matagorda Bay region of Texas. *Journal Name Not Specified* 1076
16. Ashqar, H., Elhenawy, M., Masoud, M., Rakotonirainy, A. & Rakha, H. Vulnerable road user detection using smartphone sensors and recurrence quantification analysis. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. pp. 1054-1059 (2019) 1077
17. Kalinicheva, E.; Ienco, D.; Sublime, J.; Trocan, M. Unsupervised Change Detection Analysis in Satellite Image Time Series Using Deep Learning Combined with Graph-Based Approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2020**, *13*, 1450–1466. doi: 10.1109/JSTARS.2020.2982631. 1078
18. De, S.; Pirrone, D.; Bovolo, F.; Bruzzone, L.; Bhattacharya, A. A NOVEL CHANGE DETECTION FRAMEWORK BASED ON DEEP LEARNING FOR THE ANALYSIS OF MULTI-TEMPORAL POLARIMETRIC SAR IMAGES. In Proceedings of the [Conference Name], [Location], [Date]; [Publisher]: [Location], [Year]. 1079
19. IEEE Computational Intelligence Society; International Neural Network Society; Institute of Electrical and Electronics Engineers. 2016 International Joint Conference on Neural Networks (IJCNN): 24-29 July 2016, Vancouver, Canada. 1080
20. Liu, G.; Li, L.; Jiao, L.; Dong, Y.; Li, X. Stacked Fisher autoencoder for SAR change detection. *Pattern Recognition* **2019**, *96*. doi: 10.1016/j.patcog.2019.106971. 1081
21. Fan, J.; Lin, K.; Han, M. A Novel Joint Change Detection Approach Based on Weight-Clustering Sparse Autoencoders. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2019**, *12*(2), 685–699. doi: 10.1109/JSTARS.2019.2892951. 1082
22. Liu, J.; Gong, M.; Qin, K.; Zhang, P. A Deep Convolutional Coupling Network for Change Detection Based on Heterogeneous Optical and Radar Images. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *29*(3), 545–559. doi: 10.1109/TNNLS.2016.2636227. 1083
23. Nafaa, S., Ashour, K., Mohamed, R., Essam, H., Emad, D., Elhenawy, M., Ashqar, H., Hassan, A. & Alhadidi, T. Advancing roadway sign detection with yolo models and transfer learning. *2024 IEEE 3rd International Conference On Computing And Machine Intelligence (ICMI)*. pp. 1-4 (2024) 1084
24. Gong, M.; Zhao, J.; Liu, J.; Miao, Q.; Jiao, L. Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* **2016**, *27*(1), 125–138. doi: 10.1109/TNNLS.2015.2435783. 1085
25. de Jong, K.L.; Bosman, A.S. Unsupervised Change Detection in Satellite Images Using Convolutional Neural Networks. 2018. Available online: <http://arxiv.org/abs/1812.05815> (accessed on [date]). 1086
26. Chen, H.; Wu, C.; Du, B.; Zhang, L.; Wang, L. Change Detection in Multisource VHR Images via Deep Siamese Convolutional Multiple-Layers Recurrent Neural Network. *IEEE Transactions on Geoscience and Remote Sensing* **2020**, *58*(4), 2848–2864. doi: 10.1109/TGRS.2019.2956756. 1087
27. Lyu, H.; Lu, H.; Mou, L. Learning a transferable change rule from a recurrent neural network for land cover change detection. *Remote Sensing* **2016**, *8*(6). doi: 10.3390/rs8060506. 1088
28. Institute of Electrical and Electronics Engineers; IEEE Geoscience and Remote Sensing Society. 2019 IEEE International Geoscience & Remote Sensing Symposium: Proceedings: July 28-August 2, 2019, Yokohama, Japan. 1089
29. Hou, B.; Liu, Q.; Wang, H.; Wang, Y. From W-Net to CDGAN: Bi-temporal Change Detection via Deep Learning Techniques. Mar. 2020. doi: 10.1109/TGRS.2019.2948659. 1090
30. Ren, C.; Wang, X.; Gao, J.; Chen, H. Unsupervised Change Detection in Satellite Images with Generative Adversarial Network. Sep. 2020. doi: 10.1109/TGRS.2020.3043766. 1091
31. Lebedev, M.A.; Vizilter, Y.V.; Vygolov, O.V.; Knyaz, V.A.; Rubis, A.Y. Change detection in remote sensing images using conditional adversarial networks. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS* 1092

- Archives*; International Society for Photogrammetry and Remote Sensing: May 2018; pp. 565–571. doi: 10.5194/isprs-archives-XLII-2-565-2018. 1124
32. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. ADS-Net: An Attention-Based Deeply Supervised Network for Remote Sensing Image Change Detection. *International Journal of Applied Earth Observation and Geoinformation* **2021**, *101*. doi: 10.1016/j.jag.2021.102348. 1125
33. Chen, J. et al. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 1194–1206. doi: 10.1109/JSTARS.2020.3037893. 1126
34. Jiang, H.; Hu, X.; Li, K.; Zhang, J.; Gong, J.; Zhang, M. PGA-SiamNet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. *Remote Sensing* **2020**, *12*(3). doi: 10.3390/rs12030484. 1127
35. Wang, Z.; Jiang, F.; Liu, T.; Xie, F.; Li, P. Attention-based spatial and spectral network with PCA-guided self-supervised feature extraction for change detection in hyperspectral images. *Remote Sensing* **2021**, *13*(23). doi: 10.3390/rs13234927. 1128
36. Elhenawy, M., Ashqar, H., Masoud, M., Almannaa, M., Rakotonirainy, A. & Rakha, H. Deep transfer learning for vulnerable road users detection using smartphone sensors data. *Remote Sensing*. **12**, 3508 (2020) 1129
37. Ashqar, H., Jaber, A., Alhadidi, T. & Elhenawy, M. Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing. *ArXiv Preprint ArXiv:2409.18286*. (2024) 1130
38. Abu Tami, M., Ashqar, H., Elhenawy, M., Glaser, S. & Rakotonirainy, A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. *Vehicles*. **6**, 1571-1590 (2024) 1131
39. Ashqar, H., Alhadidi, T., Elhenawy, M. & Khanfar, N. Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems. *Automation*. **5**, 508-526 (2024) 1132
40. Zhang, X.; Tian, S.; Wang, G.; Zhou, H.; Jiao, L. DiffUCD: Unsupervised Hyperspectral Image Change Detection with Semantic Correlation Diffusion Model. May 2023. Available online: <http://arxiv.org/abs/2305.12410> (accessed on [date]). 1133
41. Bandara, W.G.C.; Nair, N.G.; Patel, V.M. DDPM-CD: Remote Sensing Change Detection using Denoising Diffusion Probabilistic Models. Jun. 2022. Available online: <http://arxiv.org/abs/2206.11892> (accessed on [date]). 1134
42. Wen, Y.; Ma, X.; Zhang, X.; Pun, M.-O. GCD-DDPM: A Generative Change Detection Model Based on Difference-Feature Guided DDPM. Jun. 2023. Available online: <http://arxiv.org/abs/2306.03424> (accessed on [date]). 1135
43. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*(1), 574–586. doi: 10.1109/TGRS.2018.2858817. 1136
44. Shen, L. et al. S2Looking: A Satellite Side-Looking Dataset for Building Change Detection. Jul. 2021. doi: 10.3390/rs13245094. 1137
45. Liu, M.; Chai, Z.; Deng, H.; Liu, R. 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* **2022**, *15*, 4297–4306. doi: 10.1109/JSTARS.2022.3177235. 1138
46. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters* **2022**. doi: 10.1109/LGRS.2021.3056416. 1139
47. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. Jan. 2022. Available online: <http://arxiv.org/abs/2201.01293> (accessed on [date]). 1140
48. Codegoni, A.; Lombardi, G.; Ferrari, A. TINYCD: A (Not So) Deep Learning Model For Change Detection. Jul. 2022. Available online: <http://arxiv.org/abs/2207.13159> (accessed on [date]). 1141
49. Wang, Z.; Zhang, Y.; Luo, L.; Wang, N. CSA-CDGAN: channel self-attention-based generative adversarial network for change detection of remote sensing images. *Neural Computing and Applications* **2022**, *34*(24), 21999–22013. doi: 10.1007/s00521-022-07637-z. 1142
50. Kased, A., Rabee, R., Fahmy, A., Mohamed, H., Yacoub, M., Elhenawy, M., Ashqar, H., Hassan, A. & Glaser, S. Intersection detection using vehicle trajectories data: Deep Neural Network application. *Australasian Road Safety Conference, 2023, Cairns, Queensland, Australia*. (2023) 1143
51. Parelius, E.J. A review of deep-learning methods for change detection in multispectral remote sensing images. *Remote Sensing* **2023**, *15*(8), 2092. 1144
52. Barkur, R.; Suresh, D.; Lal, S.; Reddy, C.S.; Diwakar, P.G. Rscdnet: A robust deep learning architecture for change detection from bi-temporal high resolution remote sensing images. *IEEE Transactions on Emerging Topics in Computational Intelligence* **2022**, *7*(2), 537–551. 1145
53. Paul, J. Change Detection by Deep Learning Models. In Proceedings of the 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 323-326. IEEE, 2022. 1146
54. Walatkiewicz, J.; Darwish, O. A Survey on Drone Cybersecurity and the Application of Machine Learning on Threat Emergence. In *International Conference on Advances in Computing Research*, Springer Nature Switzerland, 2023; pp. 523-532. 1147
55. Abdelsalam, E.; Darwish, O.; Karajeh, O.; Almomani, F.; Darweesh, D.; Kisiwani, S.; Omar, A.; Alkisarawi, M. A classifier to detect best mode for Solar Chimney Power Plant system. *Renewable Energy* **2022**, *197*, 244-256. 1148

-
56. Darwish, O.; Al-Fuqaha, A.; Anan, M.; Nasser, N. The role of hierarchical entropy analysis in the detection and time-scale determination of covert timing channels. In *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*; IEEE: 2015; pp. 153-159. 1178
57. Alshattnawi, S.; AlSobeh, A.M.R. A cloud-based IoT smart water distribution framework utilising BIP component: Jordan as a model. *International Journal of Cloud Computing* **2024**, *13*(1), 25-41. 1179
58. AlSobeh, A. OSM: Leveraging Model Checking for Observing Dynamic behaviors in Aspect-Oriented Applications. *Online Journal of Communication and Media Technologies* **2023**, *13*(4), 1-18. 1180
59. Alsobeh, A.; Shatnawi, A. Integrating data-driven security, model checking, and self-adaptation for IoT systems using BIP components: A conceptual proposal model. In *International Conference on Advances in Computing Research*; Springer Nature Switzerland: 2023; pp. 533-549. 1181
60. Khan, S. & Basalamah, S. Multi-branch deep learning framework for land scene classification in satellite imagery. *Remote Sensing*. **15**, 3408 (2023) 1182
61. Khan, S. & Basalamah, S. Multi-scale and context-aware framework for flood segmentation in post-disaster high resolution aerial images. *Remote Sensing*. **15**, 2208 (2023) 1183
- 1184
- 1185
- 1186
- 1187
- 1188
- 1189
- 1190
- 1191