ChangeDA: depth-augmented multi-task network for remote sensing change detection via differential analysis.

MENG, J., XU, X., ZHANG, Z., LI, P., XIE, G., REN, J. and ZHENG, Y.

2025

ChangeDA: Depth-Augmented Multi-task Network for Remote Sensing Change Detection via Differential Analysis

Jiangtao Meng, Xinying Xu, *Member, IEEE,* Zhe Zhang, Pengyue Li, Gang Xie, Jinchang Ren, *Senior Member, IEEE,* Yuxuan Zheng

***Abstract*—In the field of Remote Sensing Change Detection (RSCD), accurately identifying significant changes between bi-temporal images is essential for environmental monitoring, urban planning, and disaster assessment. In recent years, advancements in deep learning for computer vision have transformed RSCD, significantly enhancing its effectiveness. However, existing methods often overlook the importance of depth information, focusing primarily on two-dimensional information. This limits their ability to capture subtle changes and structural details in three-dimensional space. To address these limitations, we introduce ChangeDA—a depth-augmented multi-task network designed to enhance the effectiveness of RSCD. ChangeDA introduces a depth encoder module to extract implicit depth information from optical images, enabling the utilization of 3D structural information without reliance on external data sources. Through the Depth Infusion Module (DIM), depth information is integrated into the dual-temporal feature maps, significantly enhancing the network's ability to perceive changes in three-dimensional spatial structures. Additionally, ChangeDA includes a Differential Feature Extractor (DFE) tailored to pinpoint differential features between sequential images, and an Adaptive All feature Fusion (AAFF) strategy that significantly improves recognition accuracy and generalization capability through cross-level feature integration. Performance evaluations on four prominent single-modal datasets—LEVIR-CD, S2-Looking, WHU-CD, and SYSU-CD—yielded state-of-the-art F1-scores of 92.27%, 66.42%, 94.12%, and 82.74%, respectively. Furthermore, ChangeDA also achieved outstanding results on the multi-modal 3DCD dataset, with an F1 score of 63.52% in 2D CD and an RMSE of 1.20 in the 3D CD task. These results demonstrate ChangeDA's robust adaptability across diverse targets and real-world scenarios.**

***Index Terms*—Change detection, Multi-task, Multimodal, Depth estimation, Optical flow, Remote sensing.**

This study was jointly supported by the National Natural Science Foundation of China (No.62203319), Natural Science Foundation of Shanxi Province (No.202203021212220), Open Project Fund of the State Key Laboratory of Virtual Reality Technology and Systems (No.VRLAB2023A06), Shanxi Science and Technology Cooperation and Exchange Project (No.202104041101030), Key Research and Development Plan of Shanxi Province (No.202102020101002), Key Research and Development Plan Project (International Cooperation) of Shanxi Province (No.201803D421039), and in part by the Guangdong Province Key Construction Discipline Scientific Research Ability Promotion Project (No.2022ZDJS015, No.2021ZDJS025), Special Projects in Key Fields of Ordinary Universities of Guangdong Province (No.2021ZDZX1087), and the Guangzhou Science and Technology Plan Project (No.2024B03J1361, No.2023B03J1327). (Corresponding author: Xinying Xu.)

Jiangtao Meng, Xinying Xu, Zhe Zhang, Pengyue Li and Yuxuan Zheng are with the College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China (e-mail: mengjiangtao0354@link.tyut.edu.cn, xuxinying@tyut.edu.cn, zhangzhe@tyut.edu.cn, lipengyue@tyut.edu.cn, zhengyuxuan0484@link.tyut.edu.cn)

Gang Xie is with the College of Electronic and Information Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China (e-mail:xiegang@tyust.edu.cn)

Jinchang Ren is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China. He is also with the School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen AB10 7AQ, U.K. (e-mail:jinchang.ren@ieee.org)

I. Introduction

Remote Sensing Change Detection (RSCD) aims to identify significant changes in target objects from dual-temporal remote sensing images of the same area. RSCD finds applications in a variety of domains, including land cover change, urbanization processes, and geological disaster evolution [1]. In recent years, deep learning, with its exceptional feature learning capabilities, has made significant advancements in most areas of computer vision (CV), such as semantic segmentation, depth estimation, optical flow estimation, and land use and cover classification [2], [3]. As it continues to evolve, deep learning based RSCD has been increasingly attracting academic attention.

In the field of RSCD, Siamese neural networks serve as a prominent baseline method [4]. These networks use two identical sub-networks with shared weights to simultaneously extract information from bi-temporal images and process the differential information for change detection. Various methodologies have been proposed, leveraging Convolutional Neural Networks (CNNs) or Transformer architectures [5]. However, existing methods often overlook the importance of depth information, focusing primarily on information from two-dimensional images.

Similar to the human visual systems, three-dimensional (3D) information helps capture subtle differences in temporal changes [6]. Depth information reflects the geometric characteristics of the 3D world and provides more precise descriptions of man-made objects, such as buildings and their changes, offering additional discriminative features [7]. These features are crucial for tasks like change detection, as they provide the 3D structure of the scene, enhancing the overall understanding of changes. Therefore, exploring the depth information to improve change detection is essential.

Some multimodal networks have achieved promising results by utilizing the multimodal data for RSCD, demonstrating the importance and additional value of the depth information over the optical images. For example, Liu et al. [8] combined the Digital Surface Model(DSM) data from the earlier phase and digital aerial imagery from the later phase, creating a DSM-optical multi-modal dataset that used cross-attention learning

to achieve significant results in semantic and height change detection. Xie et al. [9] used the height difference (HDiff) between bi-temporal DSMs as input, designing a co-learning framework for bi-temporal images and HDiff, which demonstrated the advantages of incorporating the height information in remote sensing building change detection. However, the diverse imaging mechanisms of different sensors and the variability in satellite imaging increase the complexity of feature extraction and processing [10]. Additionally, the high cost of data acquisition and the difficulty of annotation severely limit the development and application of these algorithms [11], [12].

Therefore, an interesting research direction in RSCD is to introduce the depth information by retrieving 3D information from the minimal 2D data, typically from the optical images [13]. Monocular Depth Estimation (MDE) is a low-cost and efficient method that addresses this challenge. MDE estimates the depth information from single-camera images, making it economically viable in terms of equipment and computational resources. The generated depth maps can be easily integrated into the change detection tasks, improving the detection accuracy and robustness [14].

Under the multi-task learning (MTL) framework, the combination of MDE and RSCD shows significant advantages, enabling them to work collaboratively. Depth estimation, as an auxiliary task, shares the feature maps with change detection and realizes interaction through a Depth Injection Module (DIM). This combination fully utilizes the prior knowledge and ground truth (GT) of change detection, effectively improving the accuracy of depth estimation. At the same time, the introduction of the depth information greatly enhances the granularity of change detection and provides a more detailed perspective for understanding the spatial changes, significantly improving the generalization ability and inference efficiency of the model, making it surpass conventional 2D semantic change detection methods.

In this paper, we propose a novel multi-task learning framework for RSCD, named ChangeDA. ChangeDA leverages a multi-task learning approach, where depth estimation and change detection share the same underlying feature extraction layers to better understand the three-dimensional structure of scenes. The depth encoder module automatically extracts the depth information from single-modal optical images, addressing the limitations of single-modal data and enhancing change perception. The Depth Injection Module (DIM) employs a hierarchical fusion strategy, combining direct depth addition with depth attention mechanisms to integrate depth information into bi-temporal feature maps, thus improving the detection of three-dimensional structural changes. Additionally, we introduce a Difference Feature Extractor (DFE) to capture the differences between bi-temporal feature maps. To further enhance the accuracy and efficiency of change detection, we use an Adaptive All-Feature Fusion (AAFF) strategy to integrate these cross-level difference features.

The major contributions of our work can be highlighted as follows.

(1)To the best of our knowledge, ChangeDA pioneers in integrating depth estimation into the RSCD domain and adopting a depth-augmented multi-task learning paradigm for end-to-end RSCD tasks.

(2) We propose a novel Differential Feature Extractor, which introduces the additional optical flow consistency assessment to extract the change features, thereby capturing the robust change characteristics from bi-temporal feature maps.

(3) We innovatively introduce Adaptive All Feature Fusion, which adaptively integrates the depth cues, semantic information, and edge contours by fusing differential feature pyramids across various levels.

(4) The proposed ChangeDA achieves state-of-the-art performance not only on the single-modal LEVIR-CD, S2Looking, WHU-CD, and SYSU-CD datasets but also on the multimodal 3DCD dataset, yet maintaining a modest computational complexity.

The rest of this article is organized as follows. In Section II, we provide a brief review of the related work. Section III details the proposed ChangeDA framework. Section IV presents the experimental results on various datasets to validate the efficacy of our proposed ChangeDA. Finally, Section V summarizes the contributions and concludes the work.

II. Related works

A. Depth Estimation

Monocular Depth Estimation (MDE), as a prototypical ill-posed problem [15], revolves around recovering three-dimensional depth information from two-dimensional images. Early approaches [16], [17], [18] were heavily reliant on handcrafted features and conventional computer vision techniques. These methods, constrained by their need for explicit depth cues, struggled with complex scenes involving occlusions and texture-less regions, where factors such as lack of cues, scale ambiguity, transparency, or reflective materials could significantly increase the uncertainty of estimation results [19].

Deep learning-based methods revolutionized MDE by effectively learning depth representations from meticulously annotated datasets, thereby enabling reasonable depth map prediction from a single RGB input [20], [21]. Eigen et al. [22] pioneered the use of a multi-scale fusion network for depth regression, inspiring in-depth research into deep learning methodologies for depth estimation. For instance, CLIFFNet [23] employs a multi-scale CNN-based framework to yield high-quality depth predictions, while TransDepth [24] enhances the accuracy and robustness in complex scenes with multiple objects through a Transformer-based dual relation graph learning framework that integrates structural and semantic information.

Despite marked improvements in depth prediction accuracy, these methods encounter several challenges in practice. They not only lead to slow convergence during training but also frequently settle into suboptimal solutions, a consequence of treating depth estimation as a conventional regression problem [25]. Hence, alternative strategies explored in studies like [25], [26] involve segmenting the continuous depth range into multiple discrete intervals, reframing the depth prediction task as per-pixel classification. This shift aims to expedite network convergence, evade local optima, potentially simplify

network architectures, and maintain or enhance depth estimation accuracy. For example, DORN [25] adopts the Spacing-Increasing Discretization (SID) strategy, transforming depth estimation into ordinal regression and, in conjunction with a multi-scale network architecture, efficiently produces high-resolution depth maps from single images. Building upon [25], SORD [26] bolsters the model's handling of ordinal data with soft labels, demonstrating enhanced generalization and adaptability, marking clear advancements in depth estimation with a more flexible and stable discretization setup.

To strike a balance between inference speed and prediction accuracy, some methods [15], [27] recast the problem as classification-regression, alleviating visual quality degradation and noticeable depth discontinuities induced by depth value discretization. Johnston et al. [27] leverage discrete disparity volumes to refine depth uncertainty estimation and harness self-attention mechanisms to capture global context features, effectively mitigating discontinuity issues in depth estimation. Bhat et al. [15] further advance this line of thought by dynamically adjusting depth bins based on image content, efficiently combining advantages of classification and regression through linear combinations of bin center values, significantly enhancing depth estimation precision. In view of this analysis, our work adopts a classification-regression approach to depth estimation, enabling real-time generation of accurate depth maps from single optical images. This approach balances accuracy and computational efficiency, leveraging the strengths of both classification and regression.

B. Multi-task RSCD

In the research progress of multi-task learning (MTL), numerous methods focus on simultaneously executing multiple tasks through a single network to enhance model performance [28]. Compared with traditional single-task RSCD, multi-task RSCD can combine the advantages of different tasks within the same framework and obtain more abundant context information. Currently, multi-task RSCD detection mainly concentrates on several types of task combinations, such as semantic segmentation and change detection. The research of Shen et al. [29] and Cui et al. [30] shows that by using semantic segmentation to assist change detection, the model can better understand the types of ground objects and thus accurately locate the change regions. Another type is building extraction and change detection. The research of Sun et al. [31] and Hong et al. [32] extracts the contour information of buildings to provide strong support for change detection in building-related change analysis. These multi-task architectures utilize other tasks as auxiliaries and enrich the performance of change detection.

However, the above algorithms fail to effectively incorporate depth information. In three-dimensional scenes, depth is crucial for analyzing the structure and distribution of ground objects. For example, in urban RSCD tasks, the height difference of buildings and the topographic undulation are key change features. The lack of integration of depth information makes it difficult for the network to understand three-dimensional space, resulting in impaired accuracy and completeness when handling related tasks.

Recently, methods of introducing depth information into change detection have emerged. For instance, Marsocci et al. [33] and Xiao et al. [34] added 3D prediction heads to the two-dimensional change detection network to enable the simultaneous execution of two-dimensional change detection (2D CD) and three-dimensional change detection (3D CD) tasks, thereby introducing depth information into RSCD. However, these methods have certain limitations. The depth information they obtain comes from the ground truth (GT) of 3D CD, so they cannot be applied to a wide range of 2D CD datasets without depth GT.

Based on this, we propose an innovative multi-task learning framework called ChangeDA. ChangeDA introduces depth estimation as an auxiliary task into the multi-task learning framework. By loading the pre-trained weights on the depth dataset, the network is endowed with initial depth estimation ability. Meanwhile, in terms of network design, depth estimation and change detection share some underlying feature extraction layers, making full use of the multi-task synergy advantage. With the help of the GT of RSCD, the network is gradually optimized, which can effectively improve the accuracy of depth estimation and further enhance the network's ability to perceive changes in the three-dimensional structure of the scene. In this way, even on 2D CD datasets without depth GT, ChangeDA can exert the multi-task advantage and achieve better results.

C. Differential Feature Learning

In RSCD, semantic segmentation is a common approach used to address change detection tasks. However, there are fundamental differences between change detection and semantic segmentation. Semantic segmentation focuses on enhancing object recognition by extracting global semantic information from a single image [35], [36], such as buildings or roads; in contrast, RSCD emphasizes the extraction of inter-temporal differences to accurately locate changed regions, regardless of their specific semantics. Due to the abstractness and uncertainty of change regions, RSCD presents greater challenges than semantic segmentation. Therefore, in RSCD tasks, it is crucial to effectively extract differences between images taken at different times and analyze these differences to precisely identify changed areas.

Direct extraction of change features can be achieved through unsupervised means using computational strategies like feature subtraction or similarity measures [37]. Within supervised learning frameworks, constructing differences between two temporal feature representations is a primary step. Common approaches involve combining feature maps from different times through subtraction, addition, or direct concatenation to generate a feature map that reflects temporal changes. However, these conventional fusion methods sometimes fail to comprehensively and finely express all differences between the two feature maps, limiting their representational power.

To address the challenges in RSCD, researchers have proposed a series of innovative solutions. For instance, Lee et al. [38] introduced a local similarity attention module for learning difference features. LSS-Net uses cosine similarity

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3532468

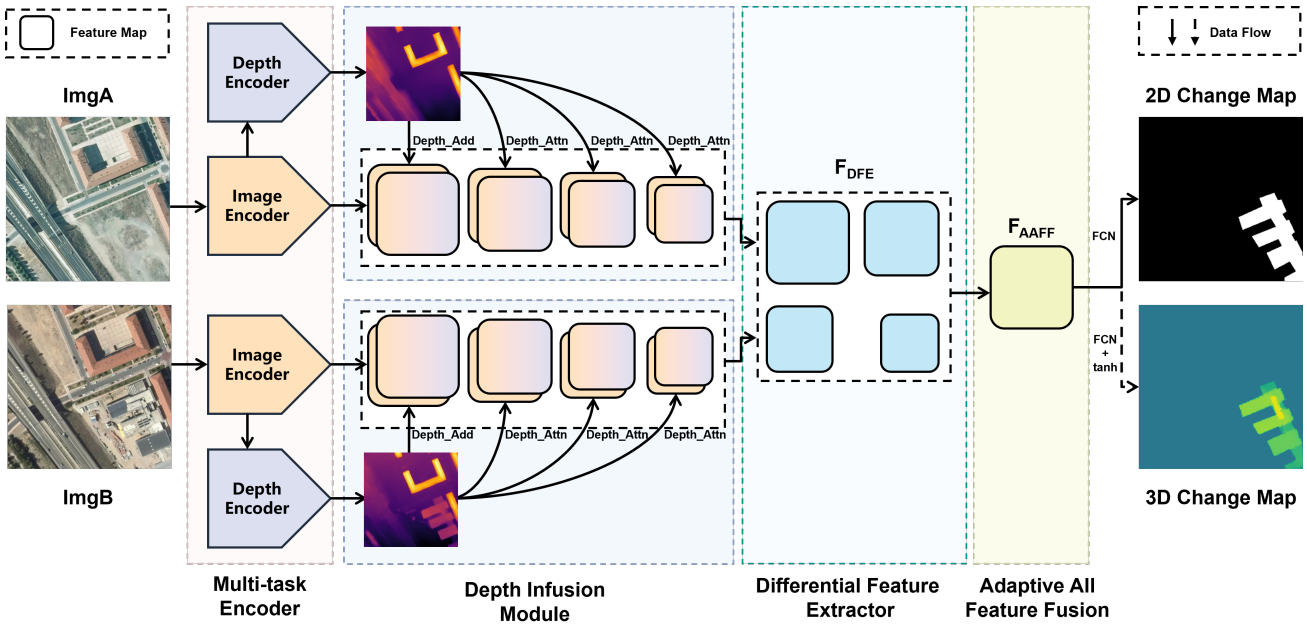

Fig. 1. Architecture of the proposed ChangeDA. Depth_Add and Depth_Attn are two methods for infusing depth information into feature maps. F_{DFE} refers to the differential feature pyramid extracted from bi-temporal features at the same level after passing through the DFE module. F_{AAFF} denotes the feature maps obtained by cross-level feature fusion of the differential feature pyramid. 2D CD and 3D CD results are generated through different branches of the network.

to measure the feature differences between input images, enabling the network to better utilize content information in the image sequence, thereby improving the accuracy of urban land change detection. The ChangerEx [39] network employs FDAF, an innovative differential feature extraction method, to learn differential feature mappings. Specifically, it first estimates the pixel offset between the two temporal images, then aligns the bitemporal feature maps, and finally extracts difference information through subtraction, explicitly highlighting changed regions to improve detection accuracy. After concatenation via 3D convolutions, AFCF3DNet [40] directly processes the bi-temporal images. By leveraging the internal fusion characteristic of 3D convolutions, it acquires the fusion features of the bi-temporal images, and thus the difference information is implicitly obtained. SEIFNet [41] enhances the feature representation ability for changed objects through its ST-DEM module. Specifically, the ST-DEM is composed of the subtraction and connection branches. These two branches jointly capture the global and local change information, strengthening the extraction of change features. FFBDNet [42] extracts the difference information using the FIFM module. The features are initially enhanced by the FIFM through the interleave fusion, followed by addition and subtraction to obtain the difference features from the interacted ones.

Despite efforts to enhance network expressiveness through various modules [43], [44], leading to positive outcomes, these networks still rely on the framework of semantic segmentation. In differential feature extraction, they primarily enhance the network's feature expression through basic techniques such as cosine similarity, feature subtraction, or concatenation. Since they do not sufficiently emphasize and design targeted modules for learning change features, these methods struggle to effectively capture and distinguish different patterns and subtle changes, leading to limitations in performance when handling complex RSCD tasks.

To enhance the saliency of change features in feature maps and accurately delineate changed regions, we developed the Difference Feature Extractor (DFE). This innovative algorithm deeply analyzes subtle changes within various semantic regions across bitemporal remote sensing images. By integrating a combination of optical flow consistency verification, feature subtraction, and cosine similarity into its differential feature extraction strategy, DFE provides an advanced perspective for bitemporal remote sensing imagery change detection. As a core component of ChangeDA, DFE enables us to effectively distill change features from bitemporal feature pyramids, precisely capturing the differences between two remote sensing images.

III. Method

A. Architecture Overview

In this study, we propose ChangeDA, as depicted in Fig. 1, a framework that innovatively extracts and integrates information from original images using a multi-task learning approach. ChangeDA focuses on exploiting the synergy of different yet complementary information within the RGB image. It contains visual characteristics like color, texture, and edges, along with implicit depth information. In the multi-task learning process, the network learns from both these aspects simultaneously. For example, in building change analysis, visual features help initially identify and locate building structures, and the depth information is crucial for precisely understanding height

differences and three-dimensional alterations. By jointly leveraging this information, ChangeDA can achieve better results in RSCD.

To evaluate the superiority of ChangeDA, we will conduct comparisons with multimodal networks on multimodal datasets. We aim to demonstrate that without relying on traditional multimodal data combinations (from different sensor types), our network can achieve better results, highlighting its unique design and contribution to the RSCD field. Through meticulously designed differential feature extraction and cross-level feature fusion modules, it presents a novel and efficient solution for RSCD. Our framework consists of four main components: a multi-task encoder, a depth infusion module, a differential feature extraction module, and a decoder based on cross-level feature fusion.

In the first part, we employ a Siamese image encoder (ResNet18) to process dual-temporal remote sensing images in parallel, generating multi-level feature maps with decreasing spatial resolution. Simultaneously, a Siamese depth encoder is used to obtain depth maps. This process not only captures the basic structure and depth information of the images but also lays a multi-scale foundation for subsequent analyses. The second component, the Depth Infusion Module (DIM), integrates depth maps with multi-level feature maps. By incorporating the third-dimensional depth aspect into the two-dimensional visual features, this step significantly enhances the model's spatial understanding and detail discernment capabilities, enabling it to analyze scene changes from a more enriched dimensional perspective. The third part utilizes the Differential Feature Extraction (DFE) module to transform these depth-enriched multiscale feature maps into differential feature maps that focus on revealing temporal differences between the dual-phase images. Finally, in the fourth part, through the Adaptive All-feature Fusion (AAFF) module, cross-level fusion of the differential feature maps is performed, followed by the prediction of the fused features. This strategy facilitates effective integration of features across different levels, markedly improving the model's accuracy and generalization performance in complex changing scenarios.

B. Multi-task Encoder

The multi-task encoder of ChangeDA comprises an image encoder and a depth encoder, both utilizing Siamese subnetworks with shared weights to extract representative information within a consistent feature space. The image encoder employs ResNet18 to process the original images and generate a feature pyramid (F_1 to F_4) for each temporal phase.

The depth encoder takes the multi-scale feature pyramid (F_1 - F_4) produced by the image encoder as the input. Inspired by the AdaBins algorithm [15], we refine the feature maps using a fusion and upsampling strategy. This approach can not only strengthen the model's understanding of the global scene structure but also significantly enhance its depth inference capability across diverse scenes by integrating the cross-level feature information. Leveraging these fused high-level features, we implement a dynamic binning technique combined with a depth center prediction approach. This mechanism enables ChangeDA to output the depth maps that match the spatial resolution of the coarsest level F_1, ensuring meticulous and accurate depiction of the depth information for each pixel. It provides a nuanced representation for scene understanding in three-dimensional perspective. The underlying formula is as follows:

$$depth\_map = Depth\_Encoder(F_1, F_2, F_3, F_4) \tag{1}$$

where Depth_Encoder represents the depth encoder and depth_map denotes the extracted depth map.

In the depth encoder, the multi-scale features F_1 to F_4 are initially processed via a Feature Pyramid Network (FPN) to obtain F_d. The FPN enhances the expressive capability of multi-scale features, making F_d richer in information and aligning its spatial resolution with F_1. Next, F_d is fed into a Mini Vision Transformer (mViT) to get F'_d, ensuring effective association and transformation of features across different levels and capturing long-range dependencies and cross-scale information. This process can be described as follows.

$$F_d = FPN(F_1, F_2, F_3, F_4) \tag{2}$$

$$F'_d = mViT(F_d) \tag{3}$$

Subsequently, the different dimensional features of F'_d are processed separately to obtain bin widths w and bin probabilities p. Here, we define b to represent a bin, which serves as a division interval in the depth domain for categorizing depth values. Specifically, b_i represents the right boundary of the i-th bin. There are a total of n bins, and the bin centers are denoted as c. Finally, the depth value for each pixel is calculated by multiplying and summing the bin centers c_i and their corresponding probabilities p_i. This process is broadcasted to each pixel to generate the final depth map, depth_map, as detailed below.

$$w = Softmax(MLP(F'_d[0, :])) \tag{4}$$

$$p = Softmax(PW(C_{3\times3}(F_d)) \cdot PW(F'_d[1:, :])) \tag{5}$$

$$c_i = \frac{b_i}{2} + \sum_{j=1}^{i-1} b_j \tag{6}$$

$$depth\_map = \sum_{i=1}^{n} c_i \odot p_i \tag{7}$$

Here,$C_{3\times3}$ denotes a 3×3 convolutional layer, PW() stands for Point-Wise convolution, · represents matrix multiplication, and ⊙ represents broadcasting multiplication.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3532468

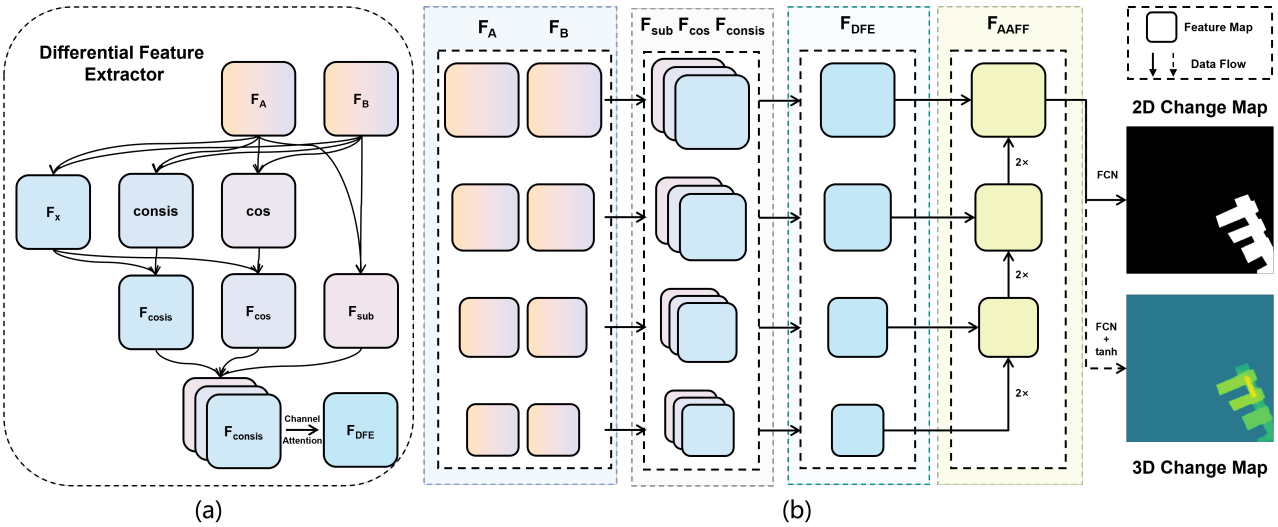

Fig. 2. (a) Structure of the DFE module. (b) Implementation details of the DFE and AAFF modules applied to the bi-temporal feature pyramids F_A and F_B, which are produced by the DIM module, in ChangeDA.

C. Depth Infusion Module

During the feature extraction process in the image encoder, as the network depth increases, the size of the feature maps decreases while the number of channels increases. Shallower-level feature maps (such as F_1) excel at capturing edge and texture information of images, whereas deeper-level feature maps (F_2 to F_4) encompass more abstract and rich semantic content. Capitalizing on this characteristic, our proposed DIM adopts a hierarchical fusion strategy: at the initial F_1 layer, depth information is explicitly incorporated into the feature representation through a direct depth_add operation; for the more deeper layers F_2 to F_4, a depth_attn mechanism is employed, which subtly and dynamically embeds depth perception within the features, aiming to finely harmonize depth information with semantic understanding, thereby further optimizing the expressiveness and precision of change detection.

Through the DIM's fusion of multi-task information, the model's scope for recognizing details and overall structure in remote sensing scenes is expanded, promoting multidimensional enhancement of image features under the multi-task learning framework. The ingenuity of this approach lies in how the integration of depth maps complements the inadequacies of conventional texture and semantic features, especially in revealing subtle changes in three-dimensional spatial structures, such as minute variations in surface object heights that are otherwise difficult to discern directly in two-dimensional image comparisons. Depth data, along with textural and semantic features, validate and reinforce each other, collectively elevating the accuracy and robustness of change detection, ensuring the model can make accurate and reliable judgments amidst various environmental changes. The specific formula is illustrated as follows.

$$DIM(depth\_map, F_i) = \begin{cases} Depth\_Add(depth\_map, F_i) \\ i = 1 \\ Depth\_Attn(depth\_map, F_i) \\ i = 2, 3, 4 \end{cases} \tag{8}$$

$$Depth\_Add(depth\_map, F_i) = R(C_{3\times3}(F_i)) + CA(R(C_{3\times3}(depth\_map))) \tag{9}$$

$$Depth\_Attn(depth\_map, F_i) = softmax(QK_i^T)V_i + F_i \tag{10}$$

$$Q = PW(depth\_map) \tag{11}$$

$$K_i = PW(F_i) \tag{12}$$

$$V_i = PW(F_i) \tag{13}$$

Where R represents the ReLU activation function, and CA() signifies channel attention mechanism. The depth_add operation takes F1 and the depth map as inputs, with the depth map first being mapped through a convolution to match the dimensionality of the input feature map. Afterwards, it undergoes adaptive adjustment via channel attention before being added to the feature map, thus explicitly imbuing depth information into the feature representation.

For depth_attn, which operates on F_2 through F_4 along with the depth map, a residual cross-attention mechanism is employed to implicitly infuse depth information into the higher-level feature maps. Here, Q represents the depth map after undergoing Point-Wise convolution, while K_i and V_i denote the key and value components respectively, resulting from applying Point-Wise convolutions to the input features F_i.

D. Differential Feature Extractor

The essence of RSCD involves effectively identifying and leveraging subtle variations between bi-temporal images, with a primary goal being the suppression of unchanged information while highlighting changes [45]. Traditional methods often rely on single-dimensional comparisons, such as basic pixel-level difference analysis, but these are frequently limited by environmental variation interference and neglect of deeper feature changes. Moreover, while some studies have attempted to guide change detection using semantic understanding, these strategies are often hindered by computational burdens and error propagation or lack direct interpretability in practical applications.

For bi-temporal features, differences are primarily manifested in three aspects: numerical discrepancies between pixels, dissimilarities in pixel feature vectors, and semantic differences. When addressing pixel-level differences in bi-temporal images, directly subtracting dual-phase images can effectively highlight variations in higher-level feature maps due to their relative stability. However, this approach is less suitable for shallower-level feature maps, as they may exhibit varied representations of the same object across time due to varying lighting, viewpoints, and other external factors during image capture, leading to subtraction results that inadequately reflect actual changes and necessitate more meticulous analysis to distinguish real alterations from environmental interference-induced false positives. Regarding differences in pixel feature vectors, calculating cosine similarities between bi-temporal features offers a good solution, providing an effective measure of dissimilarity. When it comes to extracting semantic differences, the traditional method of first classifying and then differencing the information is computationally intensive and prone to error accumulation.

Building on these insights, we innovatively designed the DFE module to precisely capture and decipher change information within remote sensing imagery. This module initially computes the absolute differences between features of the dual-time phases and employs PW convolution to adaptively map these difference features, optimizing the distribution of information across channels. This effectively discriminates non-substantive changes introduced by differing acquisition conditions, particularly crucial when handling shallower-level visual features. Subsequently, we integrate cosine similarity analysis to gauge the degree of matching between pixel-level feature vectors in bi-temporal feature maps, further refining the precision of change feature identification. To delve deeper into semantic changes, we introduce an optical flow consistency checking module, which evaluates forward and backward optical flows generated from bi-temporal features based on their consistency scores, accurately pinpointing regions where changes in remote sensing objects occur.

In the optical flow consistency check, particular attention is given to mismatched areas, as they typically indicate key dynamic changes in the scene. Specifically, when bidirectional optical flows fail to exhibit expected consistency, this usually signifies the disappearance of existing objects, the emergence of new objects, or significant object displacements. In RSCD tasks, these mismatches are interpreted as strong indicators of changing objects rather than simple environmental disturbances or computational errors. Therefore, we leverage this observation, using the detection of inconsistencies in bidirectional optical flows as potent evidence for identifying semantic change objects. By implementing a reasonable threshold strategy to distinguish normal errors from abnormal mismatches, our system accurately identifies changing objects in complex dynamic environments without being misled by transient illumination changes or similar textures.

In summary, the computational steps of DFE are outlined as follows:

$$F_{DFE} = CA(concat(F_{sub}, F_{cos}, F_{consis})) \tag{14}$$

$$F_{sub} = PW(abs(F_a - F_b)) \tag{15}$$

$$F_{cos} = F_x \times sigmoid(1 - cosine\_similarity(F_a, F_b)) \tag{16}$$

$$F_{consis} = F_x \times sigmoid(Consistant(Flow\_make(F_a, F_b), Flow\_make(F_b, F_a))) \tag{17}$$

$$F_x = FPN(concat(F_a, F_b)) \tag{18}$$

Here, F_{DFE} represents the ultimate differential feature after concatenating outputs from three distinct feature difference extraction modules (F_{sub}, F_{cos}, F_{consis}) and passing them through a Channel Attention (CA) module. F_{sub} signifies the discrepancy in pixel values, where F_a and F_b denote the feature maps of dual-phase images, having integrated textural, semantic, and depth information; F_a corresponds to the initial temporal phase, while F_b represents the subsequent temporal phase. PW() denotes point-wise convolution; the absolute difference between the bi-temporal features F_a and F_b, post point-wise convolution mapping, constitutes F_{sub}.

F_x represents the fused bi-temporal features. Following concatenation of features F_a and F_b, they are integrated with multi-level information through a Feature Pyramid Network (FPN), yielding F_x. Fcos embodies the dissimilarity in pixel feature vectors; cosine_similarity is computed between F_a and F_b, the complement of 1 is taken, and then normalized through a sigmoid function before being multiplied with F_x to obtain F_{cos}.

F_{consis} signifies the disparity in pixel semantic information. By employing the optical flow extraction algorithm Flow_make, forward and backward optical flows are obtained from F_a to F_b and vice versa. These flows are then input into the forward-backward consistency check algorithm Consistent. The magnitude of the output, reflecting the inconsistency between the forward and backward optical flows, is normalized via a sigmoid function to represent the probability of changed areas. This probability map is subsequently multiplied with F_x, resulting in F_{consis}.

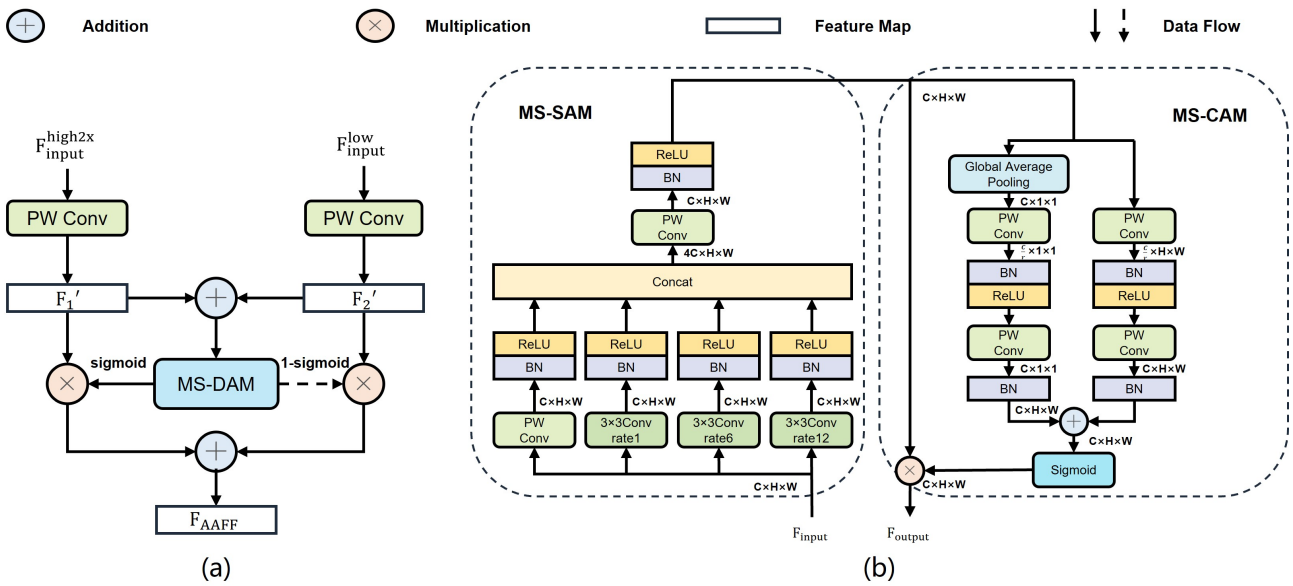

Fig. 3. (a) Structure of the AAFF module. (b) Structure of the MS-DAM module within AAFF.

E. Adaptive All Feature Fusion

AAFF is dedicated to the cross-level fusion of the difference feature pyramid from F_{DEF}, driving accurate change detection. This process permeates from deeper-level features to lower ones, leveraging semantic understanding to guide the selective fusion of textural and contour details in the shallower-level features. Specifically, when deeper-level features indicate no significant changes in an area, the system reduces the incorporation of shallower-level features from that region; conversely, if signs of change are present, it reinforces the integration of edge details. The visual manifestation of this fusion strategy is depicted in Figure. 3. After applying AAFF, the network performs the final prediction utilizing a large-sized feature map that integrates depth information, semantic content, and edge detail, resulting in highly accurate change detection outcomes.

The detailed structure of AAFF begins with the application of a double upsampling technique to map the deeper-level features to match the resolution of the shallower-level ones. Subsequently, PW convolution aligns the number of channels, ensuring consistency across different levels in both spatial and channel dimensions, thus laying the groundwork for profound information interaction. The specific formula is as follows, where PW() denotes Point-Wise Convolution, B stands for Batch Normalization, and δ represents the ReLU activation function, with $F_1^{'}$ and $F_2^{'}$ being the deeper-level and shallower-level features after alignment in spatial dimensions and channel numbers, respectively.

$$F_1^{'} = \delta\left(B\left(PW\left(F_{input}^{high2\times}\right)\right)\right) \tag{19}$$

$$F_2^{'} = \delta\left(B\left(PW\left(F_{input}^{low}\right)\right)\right) \tag{20}$$

During the fusion phase, we initiate the integration of deeper- and shallower-level features through an element-wise addition operation, laying the groundwork for subsequent fine-grained processing. Subsequently, the core component—the Multi-Scale Dynamic Attention Module (MS-DAM)—comes into play. This module comprises dual subsystems: the Multi-Scale Spatial Attention Mechanism (MS-SAM) and the Multi-Scale Channel Attention Mechanism (MS-CAM). MS-SAM leverages multi-scale dilated convolutions to capture rich spatial context, particularly enhancing edge sharpness and texture details. Meanwhile, MS-CAM employs adaptive pooling strategies at the channel level to cluster information, assigning weights to each channel that reflect its discriminative power, thereby balancing global and local information. The detailed formulations are as follows.

$$F_{coarse} = F_1^{'} + F_2^{'} \tag{21}$$

$$MDA(F) = MCA(MSA(F)) \tag{22}$$

$$MSA(F) = C_{1,1}([C_{1,1}(F), C_{3,1}(F), C_{3,6}(F), C_{3,12}(F)]) \tag{23}$$

$$MCA\left(F\right) = F \times \sigma\left(CL\left(F\right) \oplus CG\left(F\right)\right) \tag{24}$$

$$C_{a,b}\left(F\right) = \delta\left(B\left(Conv_{k=a\times a, d=b}\left(F\right)\right)\right) \tag{25}$$

$$CL\left(F\right) = B\left(PW\left(\delta\left(B\left(PW\left(F\right)\right)\right)\right)\right) \tag{26}$$

$$CG(F) = CL(G(F)) \tag{27}$$

Where F_{coarse} denotes the features after preliminary integration, MDA() represents the operation of MS-DAM, MSA() corresponds to MS-SAM, MCA() to MS-CAM, $Conv_{(k=a\times a, d=b)}$ denotes a convolution operation with a

kernel size of $a \times a$ and a dilation rate of b, $\oplus$ signifies element-wise broadcast addition, and $\times$ symbolizes element-wise multiplication.

MS-DAM, by synthesizing the attention weights generated by MS-SAM and MS-CAM, dynamically adjusts the contributions of individual features, realizing intelligent feature selection. In the ultimate adaptive feature enhancement phase, based on the computed weights, an element-wise weighted fusion operation is executed on F_1 and F_2, ensuring the output features maintain profound semantic comprehension while precisely preserving crucial detail information. This significantly enhances the model's accuracy in change detection within complex scenarios. The specific algorithmic description is as follows.

$$F_{AAFF} = MDA\left(F_{coarse}\right) \times F_1^{'} + \left(1 - MDA\left(F_{coarse}\right)\right) \times F_2^{'} \tag{28}$$

Where + denotes element-wise addition, and F_{AAFF} represents the final output feature.

F. Prediction Head and Loss Function

In RSCD, distinct prediction heads are employed for determining 2D and 3D CD maps using the final features F_{AAFF} obtained from previous steps.

For 2D CD, the primary objective is to detect areas that have experienced changes. This is achieved by first upsampled the output features to align with the dimensions of the original input, followed by using a Fully Convolutional Network (FCN) to classify each pixel as either changed or not. The FCN module ultimately generates the final 2D prediction map $\mathbf{R}_{2d} \in \mathbb{R}^{2\times H\times W}$. The 2D change detection task is considered as a binary classification problem to differentiate between changed and unchanged pixels, using the Binary Cross-Entropy (BCE) loss function below:

$$Loss_{2D} = -\frac{1}{N}\sum_{i=1}^{N}\left[\omega_c g_i \log(p_i) + \omega_u (1-g_i)\log(1-p_i)\right] \tag{29}$$

Here, i represents the i-th pixel. $p_i \in [0,1]$ is the predicted probability of the i-th pixel belonging to a certain category. $g_i \in \{0,1\}$ is the ground truth value, with $g_i = 0$ indicating unchanged and $g_i = 1$ signifies changed. ω_c and ω_u are the weights for the changed and unchanged categories, respectively. N is the total number of pixels in the image.

For 3D CD, the depth values associated with changed pixels are crucia, with a different method used to process the feature map. Initially, the final feature map F_{AAFF} is upsampled to generate the initial outputs using an FCN. A tanh activation function is then used to produce the final 3D prediction map $\mathbf{R}_{3d} \in \mathbb{R}^{H\times W}$. The Mean Squared Error (MSE) loss function is utilized to quantify the difference between the predicted and actual depth values, given by:

$$Loss_{3D} = \frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2 \tag{30}$$

Here, z_i represents the predicted depth value of the i-th pixel, and $\hat{z}_i$ is the actual depth value. When performing joint 2D and 3D CD, the overall loss function Loss is defined as a weighted combination of the 2D and 3D loss functions as follows:

$$Loss = \omega_{2D} Loss_{2D} + \omega_{3D} Loss_{3D} \tag{31}$$

In this equation, ω_{2D} and ω_{3D} are the weights assigned to the 2D and 3D loss functions, respectively. If only 2D CD is being performed, ω_{3D} is set to 0.

IV. Experimental results

A. Experiment Details

A series of meticulously planned comparative experiments are set to comprehensively assess the performance of the proposed ChangeDA model across diverse scenarios. For the single-modal dataset experiments, the intention is to validate how the incorporation of depth information and each module impacts the model's performance. To this end, ChangeDA will be benchmarked against several leading deep learning approaches that have exhibited remarkable performance on single-modal datasets. In the multimodal dataset experiments, the objective is to illustrate that the ChangeDA model can outperform multimodal networks relying on DSM-RGB inputs even when only bi-temporal images are used. A selection of DSM-RGB multimodal change detection algorithms will be chosen for comparison. Through these experiments, we endeavor to establish that the ChangeDA model can maintain superior detection capabilities without relying on supplementary modal information, thus highlighting its practical application advantages.

The proposed ChangeDA is constructed using the PyTorch framework and trained on a setup with an Intel(R) Core(TM) i9 - 13900KF CPU of the 13th generation and an NVIDIA GeForce RTX 4090 GPU. The AdamW optimizer is adopted to direct the optimization process, accompanied by a constant weight decay of 0.05 to alleviate overfitting. For the loss function, different hyperparameters are set according to the task. When ChangeDA only performs 2D CD tasks, the loss function is a standard BCE loss with ω_c = 0.5, ω_u = 0.5, ω_{2D} = 1, and ω_{3D} = 0. When ChangeDA conducts the multi - task of 2D CD and 3D CD, following the hyperparameter weight convention on the dataset, ω_c = 0.05, ω_u = 0.95, ω_{2D} = 1, and ω_{3D} = 1. All models are trained for 160,000 iterations across diverse datasets. Data augmentation strategies such as random cropping, flipping, and photometric distortion are uniformly applied to bolster generalization.

B. Evaluation Metrics

To assess the network's performance, we employ a set of key metrics that provide a comprehensive evaluation against the ground truth. For 2D CD, we use the following metrics: Precision (Prec), Recall (Rec), F1-score (F1), Intersection over Union (IoU), and Overall Accuracy (OA). For 3D CD, we introduce two additional metrics: Root Mean Squared

TABLE I
CONFUSION MATRIX.

	Predicted Results	
Ground Truth	**Changed**	**Unchanged**
Changed	TP	FN
Unchanged	FP	TN

Error (RMSE) and Change-specific Root Mean Squared Error (cRMSE). The formulas for these metrics are as follows:

$$Prec = \frac{TP}{TP + FP} \tag{32}$$

$$Rec = \frac{TP}{TP + FN} \tag{33}$$

$$F1 = \frac{2 \cdot Prec \cdot Rec}{Prec + Rec} \tag{34}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{35}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{36}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\Delta H}_i - \Delta H_i)^2} \tag{37}$$

$$cRMSE = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(\widehat{\Delta H}_i^C - \Delta H_i^C)^2} \tag{38}$$

Here, TP, FP, TN, and FN represent True Positives, False Positives, True Negatives, and False Negatives, respectively. Their confusion matrix is illustrated in Table I. n is the total number of pixels, n_c is the number of changed pixels, $\widehat{\Delta H}_i$ is the predicted depth change, and ΔH_i is the ground truth depth change. $\widehat{\Delta H}_i^C$ and ΔH_i^C represent the predicted and ground truth depth changes for changed pixels, respectively.

Precision measures the accuracy of predicting positive classes, aiming to minimize false positives. Recall emphasizes the ability to capture actual changes, striving to reduce false negatives. The F1-score, as a harmonic mean of precision and recall, balances the two, providing a more robust measure of the model's performance. IoU quantifies the spatial overlap between predicted and ground truth labels, offering an intuitive measure of localization accuracy. OA provides a holistic view of the correct classifications, reflecting the proportion of correctly classified pixels out of all pixels.

RMSE measures the average magnitude of error between predicted and ground truth depth changes across all pixels, providing a general measure of depth prediction accuracy. cRMSE focuses on the depth errors only for pixels where actual changes occur, making it particularly useful for assessing errors in critical change regions.

C. Benchmark Methods

To validate the superiority of our proposed ChangeDA model, a series of meticulously planned comparative experiments have been conducted across diverse scenarios. The benchmark methods are divided into two categories based on the type of datasets they are mainly applied to: single-modal and multi-modal datasets.

For single-modal dataset comparisons, we selected a series of well-known methods: BIT [46], ICIFNet [47], DDLNet [48], CDNet [49], ISNet [50], STANet [51], MFATNet [52], FHD [53], ChangerEx [39], ChangeFormer [54], IFNet [55], LRNet [56], CGNet [57], C2FNet [58], HCGMNet [59], DTCDSCN [60], SNUNet [61], HANet [62], P2V [63], L-UNet [64], AFCF3DNet [40], SEIFNet [41], FFBDNet [42], HMCNet [65], and DARNet [66]. These methods are widely recognized for their performance on single-modal datasets.

In the context of multi-modal dataset comparisons, several representative methods have been selected: SUNet [61], ChangeFormer [54], MTBIT [33], CSCLNet [34], and MMCD [8]. MTBIT and CSCLNet are multi-task learning networks that use bi-temporal optical images to perform both 2D CD and 3D CD tasks. Originally single-task networks, SUNet and ChangeFormer have been adapted by incorporating a 3D output head, inspired by MTBIT, to handle both 2D CD and 3D CD using bi-temporal optical images. MMCD, on the other hand, is a multi-modal network specifically designed to extract and analyze information from pre-DSM and post-RGB data.

D. Datasets

Single-modal Datasets:

LEVIR-CD dataset serves as a pivotal benchmark for building change detection, providing 637 pairs of very high-resolution (0.5 meters per pixel) remote sensing images at a size of 1024×1024 pixels. Spanning a period of 5 to 14 years, it focuses on documenting extensive land-use shifts, particularly in building advancements across diverse structures such as villas, skyscrapers, garages, and warehouses. Accurate annotations differentiate altered and unmodified building areas, totaling over 31,333 instances of changes. Following the dataset's guidelines, the evaluation setup includes 445 image pairs for training, 64 for validation, and 128 for testing.

S2Looking dataset, compiled between 2017 and 2020, is tailored for rural building change detection utilizing high-resolution satellite imagery. It comprises 5,000 image pairs at a resolution of 0.5 to 0.8 meters per pixel within a 1024×1024 frame, presenting 65,920 annotated change instances. This dataset serves as a comprehensive resource for in-depth remote sensing studies. In alignment with the dataset's official structure, our study adopts the designated partitioning: 3,500 pairs for training, 500 for validation, and 1,000 for testing.

WHU-CD dataset is a significant public asset for building change detection, presenting a high-resolution (0.2 meters per pixel) aerial image pair covering an expansive 32,507×15,354 pixel area. Due to the absence of official segmentation instructions, a preprocessing method involving cropping into non-overlapping 512×512 pixel patches was adopted. These

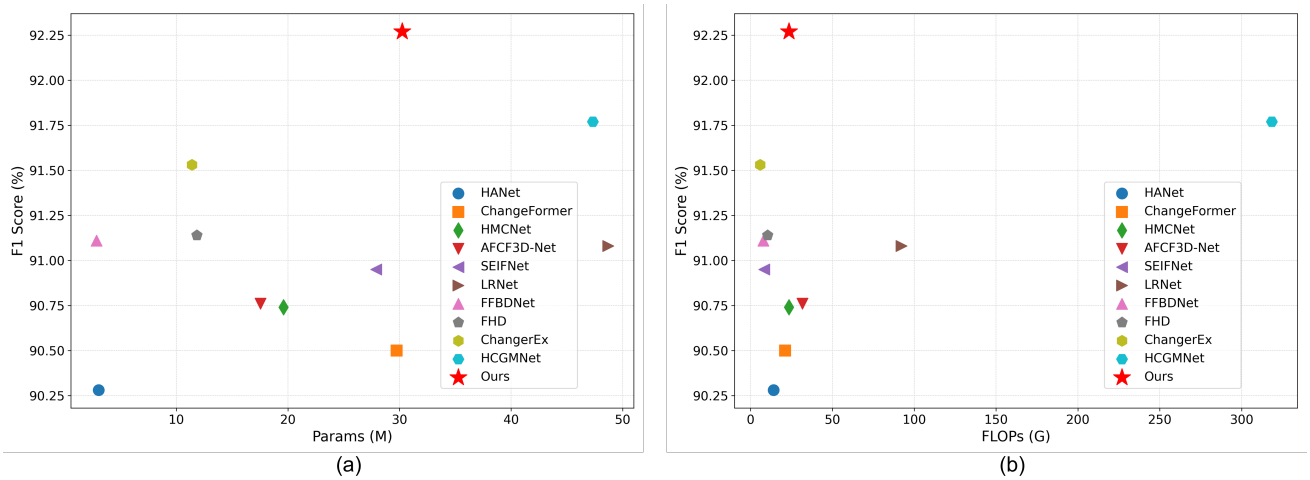

Fig. 4. Performance Comparison of Various Methods on LEVIR-CD Dataset: (a) Number of Parameters (M), (b) Floating Point Operations (G).

patches are then randomly allocated to form a dataset supporting rigorous assessment, divided into 2,153 for training, 270 for validation, and another 270 for testing.

SYSU-CD dataset is a cornerstone in urban change analysis, compiling 20,000 pairs of high-resolution aerial images at 256×256 pixels. Chronicling Hong Kong's urban evolution from 2007 to 2014, it encompasses a wide range of transformations, including new building construction, suburban growth, pre-construction groundwork, vegetation shifts, road enhancements, and maritime infrastructure development. It systematically segments into a training set of 12,000 image pairs, a validation set of 4,000 pairs, and a test set of 4,000 pairs, adhering to standard practices for algorithmic evaluations.

Multi-modal Dataset:

3DCD dataset is an essential resource for detecting changes in urban environments, specifically focusing on the historical center and surrounding commercial areas of Valladolid, Spain. This dataset comprises 472 pairs of high-resolution optical orthoimages and corresponding DSMs, sourced from two distinct aerial surveys conducted in 2010 and 2017. Additionally, it features 472 2D CD maps that document alterations to structures and 472 3D CD maps illustrating height variations between the survey periods. The dataset is meticulously partitioned to support rigorous assessment, divided into 320 pairs for training, 42 pairs for validation, and 110 pairs for testing. Notably, the training subset encompasses elevation changes ranging from -25 to 35 meters, capturing a wide spectrum of urban transformations.

E. Performance Comparison on Single-modal Datasets

As shown in Tables II-V, we have thoroughly evaluated ChangeDA across four single-modal datasets (LEVIR-CD, S2-Looking, WHU-CD, and SYSU-CD), with - symbols denoting missing data from the original papers. Analyzing the metrics, ChangeDA demonstrates superior performance across all datasets, with notable improvements in IoU/F1 scores compared to previous state-of-the-art (SOTA) models: increases of 0.86%/0.50%, 1.19%/1.07%, 2.69%/1.53%, and 2.44%/2.00%

TABLE II
QUANTITATIVE RESULTS ON THE LEVIR-CD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. ALL RESULTS ARE DESCRIBED AS PERCENTAGES. (%)

Model	Backbone	Prec	Rec	F1	IoU	OA
CDNet	ResNet18	83.61	84.14	83.87	72.21	98.35
DTCDSCN	SE-Res34	88.53	86.83	87.67	78.05	-
STANet	ResNet18	92.49	85.80	89.02	80.20	98.92
BIT	ResNet18	90.65	87.53	89.48	80.86	98.95
SNUNet	UNet++	91.31	88.67	89.97	81.77	98.99
HANet	-	91.21	89.36	90.28	82.27	99.02
MFATNet	ResNet18	91.85	88.93	90.36	82.42	99.03
ChangeFormer	MiT-b1	90.83	90.18	90.50	82.66	99.04
HMCNet	-	91.68	89.82	90.74	83.05	99.07
AFCF3DNet	-	91.35	90.17	90.76	83.08	-
SEIFNet	ResNet18	92.49	89.46	90.95	83.40	99.09
LRNet	VGG16	92.19	90.00	91.08	83.63	99.10
FFBDNet	EfficientNet	92.28	89.98	91.11	83.67	-
FHD	ResNet18	91.97	90.32	91.14	83.72	99.10
ChangerEx	ResNet18	92.97	90.13	91.53	84.38	99.15
HCGMNet	VGG16	92.96	90.61	91.77	84.79	99.18
ChangeDA	**ResNet18**	**93.67**	**90.92**	**92.27**	**85.65**	**99.22**

TABLE III
QUANTITATIVE RESULTS ON THE S2-LOOKING DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. ALL RESULTS ARE DESCRIBED AS PERCENTAGES. (%)

Model	Backbone	Prec	Rec	F1	IoU	OA
STANet	ResNet18	38.75	56.49	45.97	29.84	-
DTCDSCN	SE-Res34	68.58	49.16	57.27	40.12	-
HANet	-	61.38	55.94	58.54	41.38	99.04
CDNet	ResNet18	67.48	54.93	60.56	43.43	-
BIT	ResNet18	72.64	53.85	61.85	44.77	-
C2FNet	VGG16	**74.84**	54.14	62.83	45.80	99.22
ChangeFormer	MiT-b1	72.82	56.13	63.39	-	-
HCGMNet	VGG16	72.51	57.06	63.87	46.91	99.22
FHD	ResNet18	74.09	56.71	64.25	47.33	-
CGNet	ResNet18	70.18	59.38	64.33	47.41	99.20
ChangerEx	ResNet18	71.64	60.07	65.35	48.53	99.23
ChangeDA	**ResNet18**	72.26	**61.45**	**66.42**	**49.72**	**99.25**

TABLE IV
QUANTITATIVE RESULTS ON THE WHU-CD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. ALL RESULTS ARE DESCRIBED AS PERCENTAGES. (%)

Model	Backbone	Prec	Rec	F1	IoU	OA
SNUNet	-	75.03	92.31	82.77	70.61	98.22
BIT	ResNet18	82.04	89.74	85.71	75.00	98.62
SEIFNet	ResNet18	87.01	85.77	86.39	76.04	98.90
DARNet	-	84.20	89.85	86.93	76.89	98.75
HANet	-	88.30	88.01	88.16	78.82	99.16
ICIFNet	ResNet18	90.79	87.58	89.16	80.43	99.01
DTCDSCN	SE-Res34	90.15	89.35	89.75	81.40	-
IFNet	VGG16	93.78	87.13	90.33	82.37	99.14
DDLNet	ResNet18	91.56	90.03	90.56	82.75	99.13
HCGMNet	VGG16	92.08	**93.93**	90.31	85.33	99.45
P2V	-	94.18	90.91	92.52	86.07	99.32
LRNet	VGG16	95.11	90.04	92.51	86.06	99.47
CGNet	VGG16	94.47	90.79	92.59	86.21	99.48
FFBDNet	EfficientNet	93.60	92.95	93.27	87.39	-
AFCF3DNet	-	93.47	92.69	93.58	87.93	-
ChangeDA	**ResNet18**	**96.45**	91.91	**94.12**	**88.90**	**99.54**

TABLE V
QUANTITATIVE RESULTS ON THE SYSU-CD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. ALL RESULTS ARE DESCRIBED AS PERCENTAGES. (%)

Model	Backbone	Prec	Rec	F1	IoU	OA
HANet	-	78.71	76.14	77.41	63.14	89.52
C2FNet	VGG16	75.44	**80.67**	77.97	63.89	89.25
CDNet	ResNet18	79.34	77.29	78.30	64.34	89.90
ISNet	ResNet18	76.41	80.27	78.29	64.44	90.01
SNUNet	-	83.58	75.87	79.54	66.02	90.79
L-UNet	-	81.24	78.08	79.63	66.15	90.58
HCGMNet	VGG16	86.28	74.15	79.76	66.33	91.12
CGNet	VGG16	**86.37**	74.37	79.92	66.55	91.19
FFBDNet	EfficientNet	84.48	76.15	80.10	66.81	-
IFNet	VGG16	80.98	79.37	80.17	66.90	90.74
ICIFNet	ResNet18	83.37	78.51	80.74	68.12	91.24
DARNet	-	83.04	79.11	81.03	68.10	91.26
SEIFNet	ResNet18	84.81	79.98	82.32	69.96	91.90
ChangeDA	**ResNet18**	86.01	79.72	**82.74**	**70.56**	**92.16**

on LEVIR-CD, S2-Looking, WHU-CD, and SYSU-CD, respectively, highlighting significant and meaningful advancements.

In the detailed breakdown, our experiments on the LEVIR-CD dataset, a benchmark for high-resolution urban change detection, reveal that ChangeDA excels. It surpasses leading models such as HCGMNet, ChangerEx, and FHD across all metrics. The algorithm, leveraging deep feature enhancement and adaptive aggregation strategies, achieves remarkable improvement over prior baselines. Specifically, it attains 93.67% precision, 90.92% recall, 92.27% F1 score, 85.65% IoU, and 99.22% overall accuracy, setting new performance standards and demonstrating high precision and efficiency in complex urban landscape change detection.

Turning to the challenging S2-Looking dataset, our algorithm continues to demonstrate robust detection capabilities, particularly in detecting rural building changes. Relative to top models like ChangerEx, CGNet, and HCGMNet, our model makes a substantial leap in F1 score to 66.42% and raises IoU to 49.72%, indicating heightened sensitivity in identifying small-scale changes and in high-density areas. The algorithm also sees a marked increase in recall to 61.45% (+1.38%), significantly reducing missed detections in practical applications, further validating its comprehensiveness and practical utility.

On the WHU-CD dataset, our algorithm shines in dense urban construction detection, with precision reaching 96.45%, F1 score climbing to 94.12%, and IoU hitting 88.90%. This reinforces its strength in handling high-density building changes and reiterates its effectiveness and adaptability in complex urban scenarios.

Confronted with the diverse targets and intricate changes in the SYSU-CD dataset, the algorithm maintains strong performance. Its F1 score rises to 82.74%, IoU to 70.56%, outperforming algorithms such as ICIFNet, DARNet, and IFNet, illustrating broad applicability and high reliability across different target types, thereby affirming its robustness and efficiency in real-world applications.

In Figs. 5-8, we visually compare the test results of ChangeDA against recent classic algorithms on the LEVIR-CD, S2-Looking, WHU-CD, and SYSU-CD datasets. Using color-coding – green for TP, black for TN, yellow for FP, and red for FN regions – these visualizations compellingly demonstrate that ChangeDA achieves exceptional detection outcomes across varying datasets and scenarios, closely aligning with ground truth annotations and underscoring its superior performance.

F. Performance Comparison on Multi-modal Dataset

We compare the performance of the ChangeDA with other representative methods on the multi-modal 3DCD dataset, with the quantitative results presented in Table VI, including those from the SUNet, ChangeFormer, MTBIT, MMCD, CSCLNet, and our ChangeDA, using both 2DGT and 2DGT + 3DGT for supervised learning.

For the 2D CD task, ChangeDA with 2DGT + 3DGT for traning achieves an F1 score of 63.52 and an IoU of 46.54. Compared to the multi-modal network MMCD, which has an F1 score of 41.33 and an IoU of 26.05, ChangeDA shows significantly improved performance. This indicates that ChangeDA can effectively leverage the depth information estimated from bi-temporal optical images, even without the direct DSM data, demonstrating the effectiveness of its depth estimation network. Among other multi-task learning networks like CSCLNet, ChangeDA outperforms them slightly in terms of both F1 and IoU, achieving SOTA performance in multi-task learning scenarios.

When considering only 2D GT for supervbised training of the ChangeDA in the 2D CD single-task, it attains excellent results with an F1 score of 65.73 and an IoU of 48. 95, the best among all compared algorithms. This highlights the strong adaptability and excellent performance of the ChangeDA model even in a simpler supervision setting.

In the 3D CD task, ChangeDA with 2DGT + 3DGT for training achieves an RMSE of 1.20 and a cRMSE of 4.78. Compared to MMCD, which has an RMSE of 1.76 and a cRMSE of 4.86, ChangeDA shows better performance with a lower RMSE and a slightly lower cRMSE. Additionally,

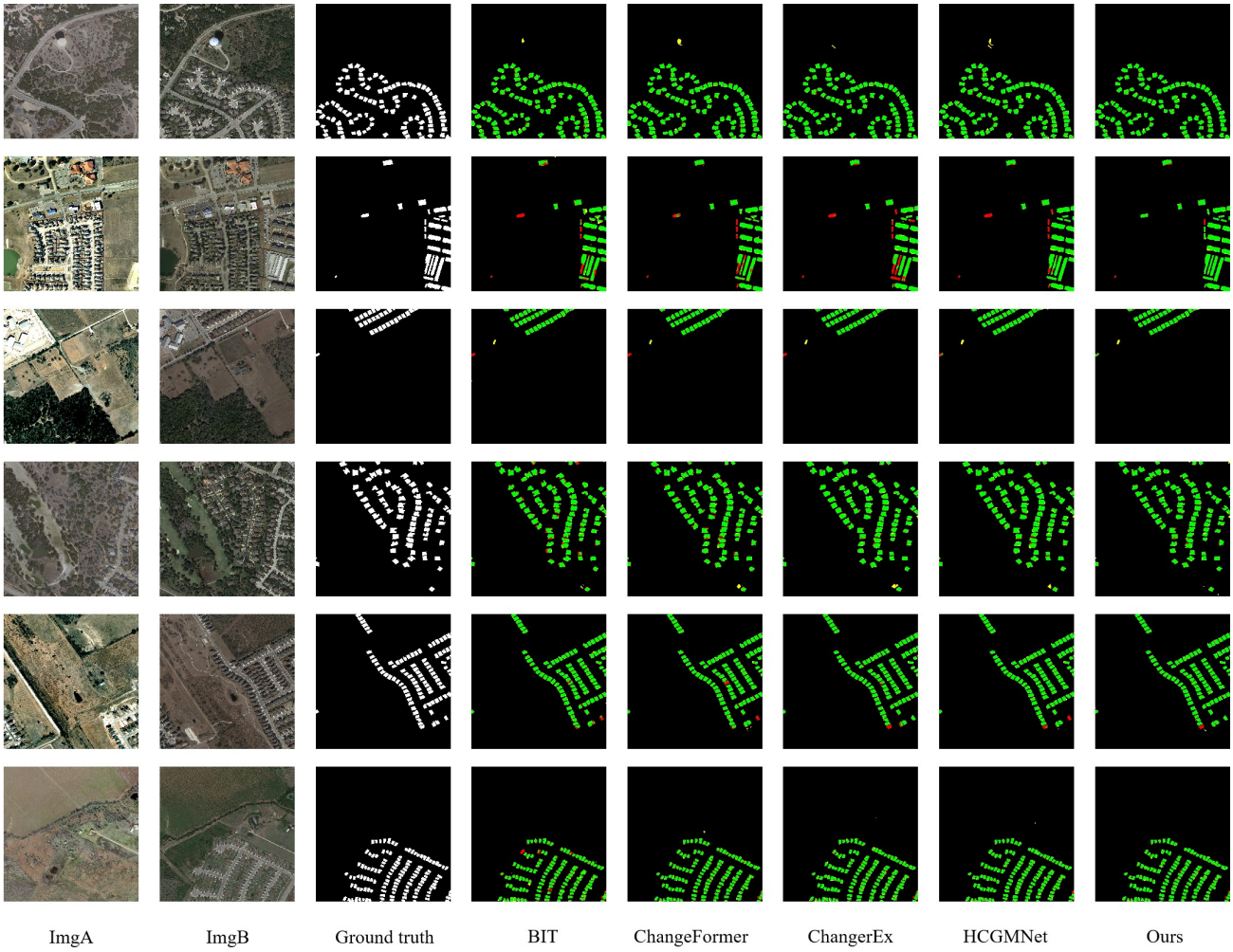

Fig. 5. Visualization results of different methods on the LEVIR-CD Dataset. The green, yellow, black and red colors represent TP, FP, TN and FN.

TABLE VI
QUANTITATIVE RESULTS ON THE 3DCD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Model	Year	Input	Supervision	Params (M)	2D CD		3D CD	
					F1(%)	IoU(%)	RMSE(m)	cRMSE(m)
SUNet	2021	RGB	2DGT+3DGT	27.9	52.39	35.50	1.47	5.04
ChangeFormer	2022	RGB	2DGT+3DGT	29.7	41.00	25.78	1.58	5.13
MTBIT	2023	RGB	2DGT+3DGT	13.1	63.28	46.29	1.31	5.52
MMCD	2024	DSM+RGB	2DGT+3DGT	11.7	41.33	26.05	1.76	4.86
CSCLNet	2024	RGB	2DGT+3DGT	26.9	63.10	46.09	1.24	4.97
ChangeDA	2024	**RGB**	**2DGT**	30.2	**65.73**	**48.95**	-	-
ChangeDA	2024	**RGB**	**2DGT+3DGT**	31.7	**63.52**	**46.54**	**1.20**	**4.78**

ChangeDA outperforms other multi-task learning networks such as SUNet, ChangeFormer, and CSCLNet in 3D CD tasks.

Fig. 9 provides a visual comparison of different models on the 3D CD dataset. The first three rows show the 2D CD results, where SUNet, ChangeFormer, MTBIT, MMCD, and CSCLNet exhibit misclassifications or missed detections. In contrast, ChangeDA (Ours) aligns more closely with the ground truth, with fewer misclassifications and missed detections. The last three rows display the 3D CD results, where ChangeDA also demonstrates higher similarity to the ground truth, showcasing its advantage in 3D change detection. Overall, these results confirm that ChangeDA performs well in both 2D and 3D change detection tasks, outperforming other methods.

G. Model Complexity

We evaluate the model complexity of various algorithms on the LEVIR dataset for predicting a 256×256 image from two aspects: i.e. the number of parameters (Params) and the number of floating-point operations (FLOPs). The values of Params and FLOPs are directly related to network complexity.

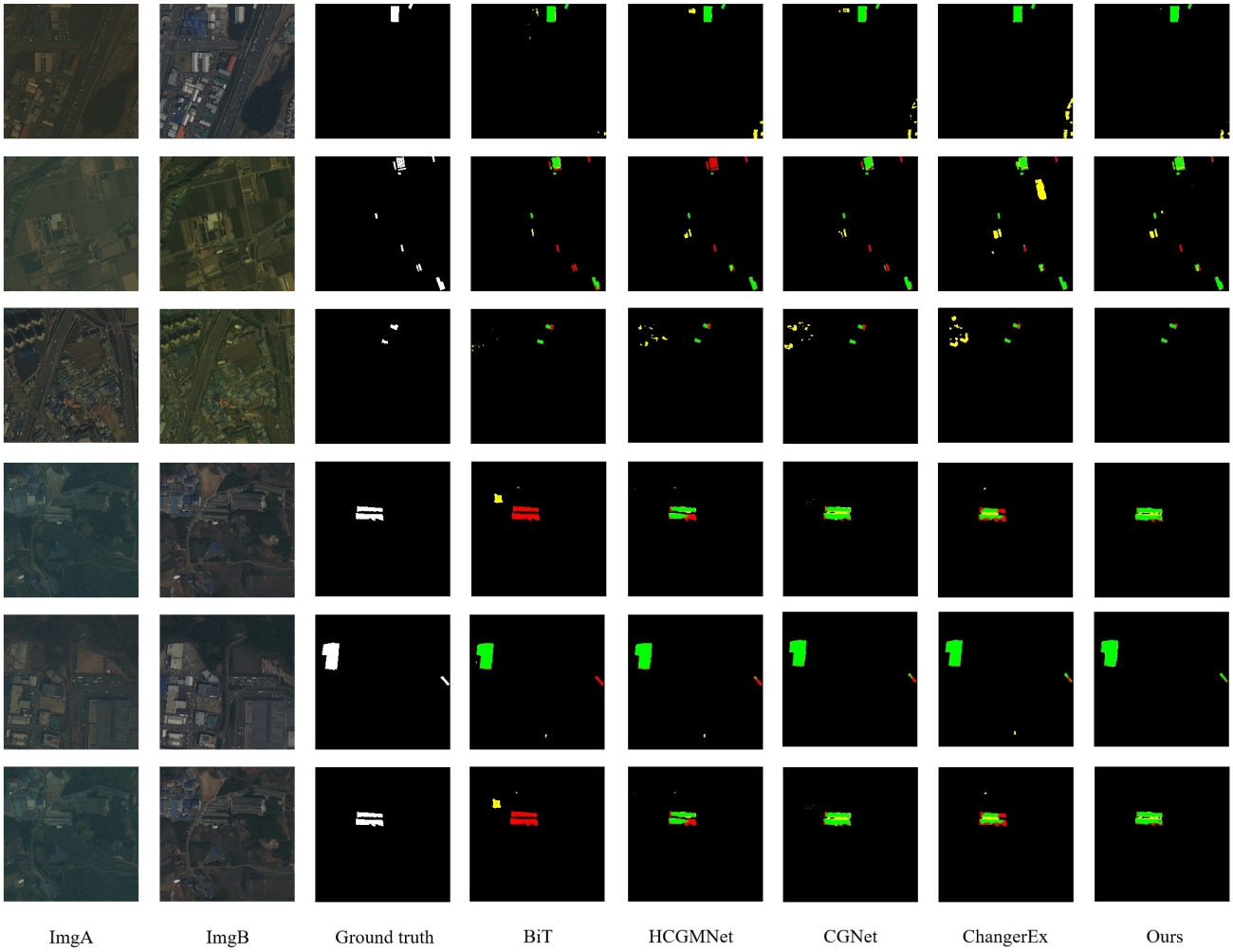

Fig. 6. Visualization results of different methods on the S2Looking Dataset. The green, yellow, black and red colors represent TP, FP, TN and FN.

TABLE VII
MODEL COMPLEXITY COMPARISONS ON THE LEVIR-CD DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Model	Params(M)	FLOPs(G)	F1
CDNet	14.33	21.52	83.87
IFNet	50.44	82.26	87.58
DTCDSCN	41.07	7.21	87.67
DSIFN	50.46	50.77	88.42
STANet	16.89	6.43	89.02
BIT	17.99	15.17	89.48
DMINet	6.24	14.55	89.90
SNUNet	12.03	54.83	89.97
HANet	3.03	14.07	90.28
ChangeFormer	29.75	21.18	90.50
HMCNet	19.61	23.54	90.74
AFCF3DNet	17.54	31.72	90.76
SEIFNet	27.91	8.37	90.95
LRNet	48.71	92.23	91.08
FFBDNet	2.85	7.81	91.11
FHD	11.83	10.43	91.14
ChangerEx	11.39	5.95	91.53
HCGMNet	47.32	318.41	91.77
ChangeDA	30.25	23.49	**92.27**

The relevant values of all compared methods are listed in Table VII. For a visual comparison, scatterplots are presented in Fig. 4.

Among these algorithms, CDNet has relatively low Params and FLOPs due to its simple design, but its F1 score is quite low, indicating its poor performance. On the other hand, LRNet has very high Params and FLOPs, but its performance is not the best. Our algorithm, ChangeDA, has the Params of 30.25M and FLOPs of 23.49G, achieving an F1 score of 92.27%, which shows that ChangeDA strikes a good balance between the model complexity and performance.

H. Ablation Studies

1) Differential Information Constituents in DFE Composition: During our ablation study conducted on the LEVIR-CD and S2Looking datasets, focusing exclusively on the Difference Method without incorporating depth information, we initially performed a comparative examination of the performance of three standalone modules designed for extracting change information—Consis, Cos, and Sub. As shown in Table VIII, the empirical results highlighted that our innovatively devised optical flow consistency estimation algorithm (Consis)

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3532468

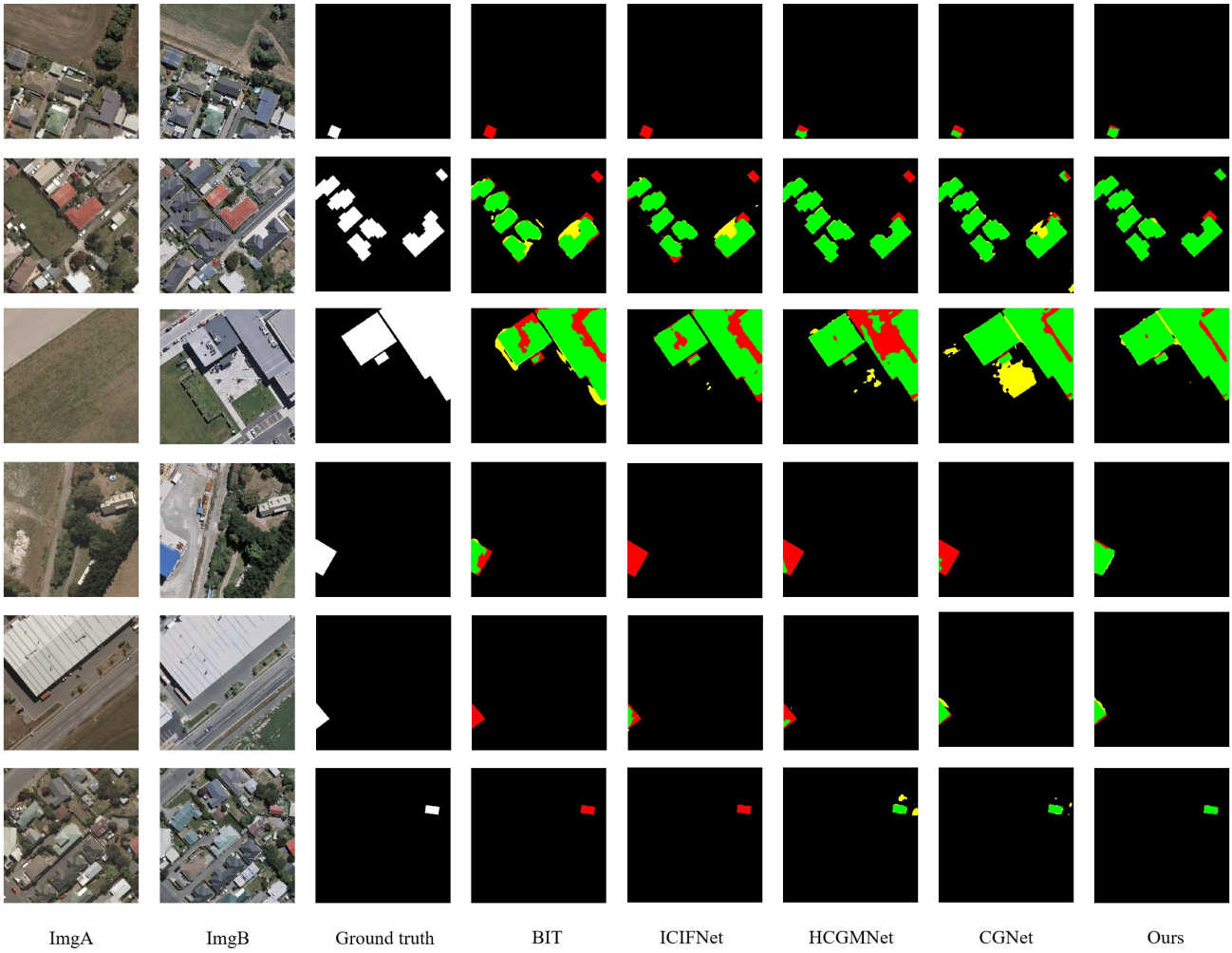

Fig. 7. Visualization results of different methods on the WHU-CD Dataset. The green, yellow, black and red colors represent TP, FP, TN and FN.

excelled when deployed individually, setting a strong baseline for subsequent evaluations.

Building on this robust foundation, we incrementally integrated Consis with the other two modules for capturing differential information, Cos and Sub. Each integration contributed to a steady enhancement in performance. Significantly, the system's performance peaked when Consis, Cos, and Sub were collectively utilized for differential information extraction, corroborating our initial theoretical assessments. This underscores that our DFE module, by synthesizing discrepancies in pixel values, spatial similarities across pixel channels, and semantic differences at the pixel level, is capable of comprehensively and deeply elucidating the variation characteristics between bi-temporal feature maps, thereby enhancing the accuracy and comprehensiveness of change detection.

2) Mechanisms for Depth Infusion into Feature Maps: In the designed experiment, we systematically investigated the effects of integrating depth information using varying strategies across different network levels. The experimental outcomes are summarized in Table IX, where F1-F4 denote the feature maps produced at four successive stages of the backbone. Here, "A" signifies depth_add, "B" represents depth_attn, and "N" denotes no depth infusion. Initially, all four stages of feature maps underwent uniform treatment with the addition of depth information via the depth_add method. We observed a marked improvement in network performance due to the inclusion of depth information.

Proceeding further, we delved into determining the most effective approach for depth integration. Beginning with the highest level feature map, F4, we progressively shifted towards employing the attention mechanism (depth_attn) for implicit depth fusion, culminating in its exclusive application across all stages. The empirical evidence revealed that the network achieved its optimal state when shallow-level feature map F1 was augmented with depth through the straightforward depth_add technique, while deeper feature maps F2 to F4 benefited from the subtler depth integration facilitated by the depth_attn mechanism.

These findings not only validate our hypothesis that, in the shallower network layers where feature maps predominantly embody edge and texture details, the direct supplementation of depth values via A enhances such features effectively; conversely, in the deeper layers rich in semantic information, the B strategy of implicitly blending depth through attention mechanisms is more conducive to exploiting depth cues. The synergy of these two approaches thus realizes the pinnacle of

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3532468

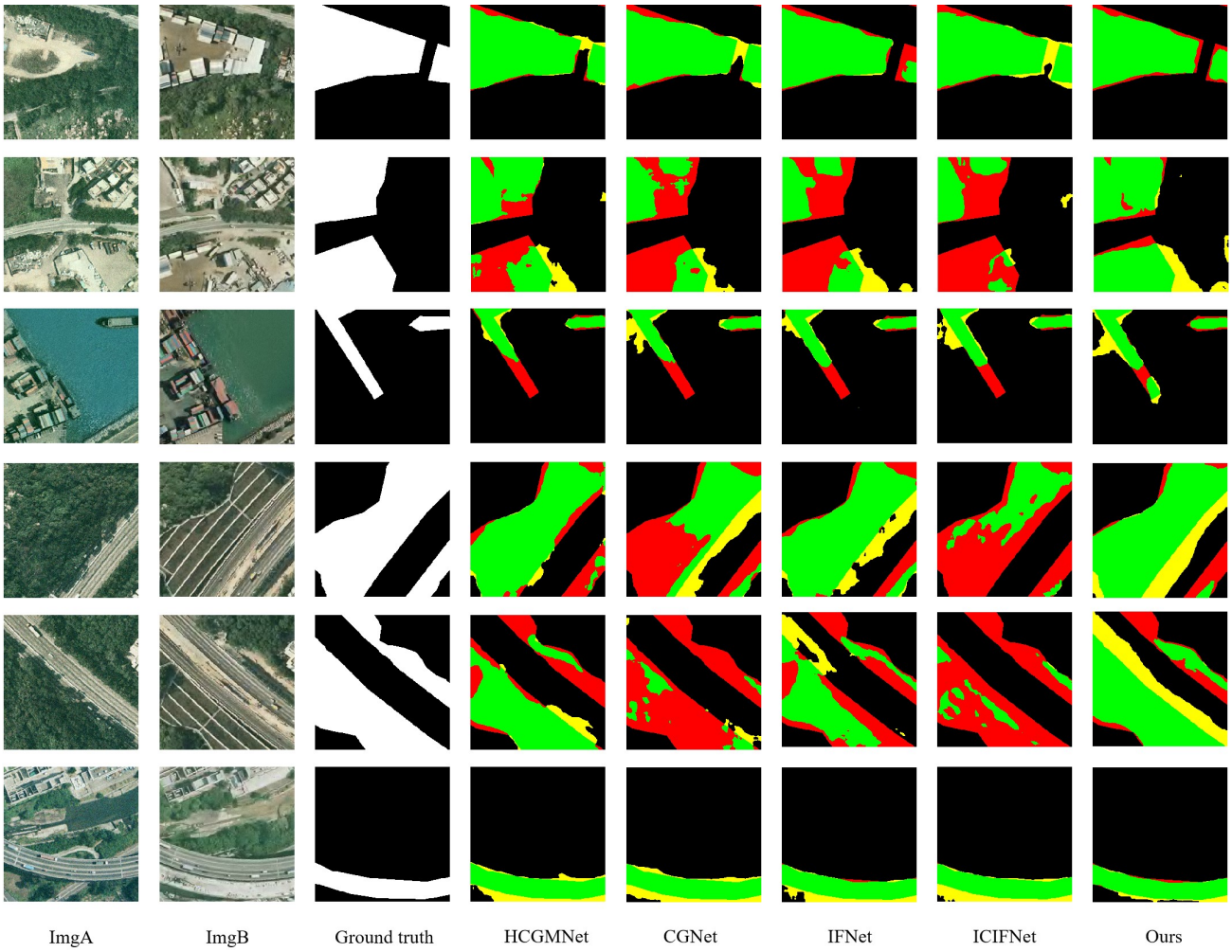

Fig. 8. Visualization results of different methods on the SYSU-CD Dataset. The green, yellow, black and red colors represent TP, FP, TN and FN.

TABLE VIII
ABLATION STUDY ON DIFFERENT METHODS OF ACQUIRING DIFFERENTIAL MAPS ON THE LEVIR-CD AND S2LOOKING DATASETS. Y REPRESENTS THE USE OF THE FEATURE EXTRACTION METHOD, AND N REPRESENTS NO USE.

Method			**LEVIR-CD**	**S2Looking**
Consis	Cos	Sub	Prec / Rec / F1 / IoU / OA	Prec / Rec / F1 / IoU / OA
Y	N	N	92.17 / 90.19 / 91.17 / 83.77 / 99.11	72.62 / 59.99 / 65.70 / 48.92 / 99.24
N	Y	N	91.91 / 90.29 / 91.10 / 83.65 / 99.10	66.75 / **64.50** / 65.61 / 48.82 / 99.18
N	N	Y	91.48 / 90.57 / 91.02 / 83.53 / 99.09	72.54 / 60.11 / 65.74 / 48.97 / 99.24
Y	Y	N	91.95 / **90.93** / 91.44 / 84.23 / 99.13	**73.38** / 59.47 / 65.70 / 48.92 / **99.25**
Y	N	Y	92.09 / **90.93** / 91.50 / 84.34 / 99.14	72.45 / 60.28 / 65.81 / 49.04 / 99.24
Y	**Y**	**Y**	**92.61** / 90.76 / **91.68** / **84.63** / **99.16**	72.19 / 61.24 / **66.27** / **49.56** / **99.25**

TABLE IX
ABLATION STUDY OF THE DEPTH INFUSION METHODS AT DIFFERENT STAGES ON THE LEVIR-CD AND S2LOOKING DATASETS. F1-F4 REPRESENT FEATURE MAPS AT STAGES 1-4, A DENOTES DEPTH_ADD, B DENOTES DEPTH_ATTN, AND N DENOTES NO DEPTH INFUSION.

Method				**LEVIR-CD**	**S2Looking**
F1	F2	F3	F4	Prec / Rec / F1 / IoU / OA	Prec / Rec / F1 / IoU / OA
N	N	N	N	92.61 / 90.76 / 91.68 / 84.63 / 99.16	72.19 / 61.24 / 66.27 / 49.56 / **99.25**
A	A	A	A	93.17 / 90.54 / 91.83 / 84.90 / 99.18	**73.03** / 60.83 / 66.37 / 49.67 / **99.25**
A	A	A	B	92.56 / 90.83 / 91.69 / 84.65 / 99.16	70.51 / 62.47 / 66.25 / 49.53 / 99.23
A	A	B	B	93.18 / 90.04 / 91.59 / 84.48 / 99.16	71.52 / 61.32 / 66.03 / 49.29 / 99.24
A	**B**	**B**	**B**	**93.67** / **90.92** / **92.27** / **85.65** / **99.22**	72.26 / 61.45 / **66.42** / **49.72** / **99.25**
N	B	B	B	92.61 / 90.16 / 91.37 / 84.11 / 99.13	66.59 / **64.83** / 65.70 / 48.92 / 99.18

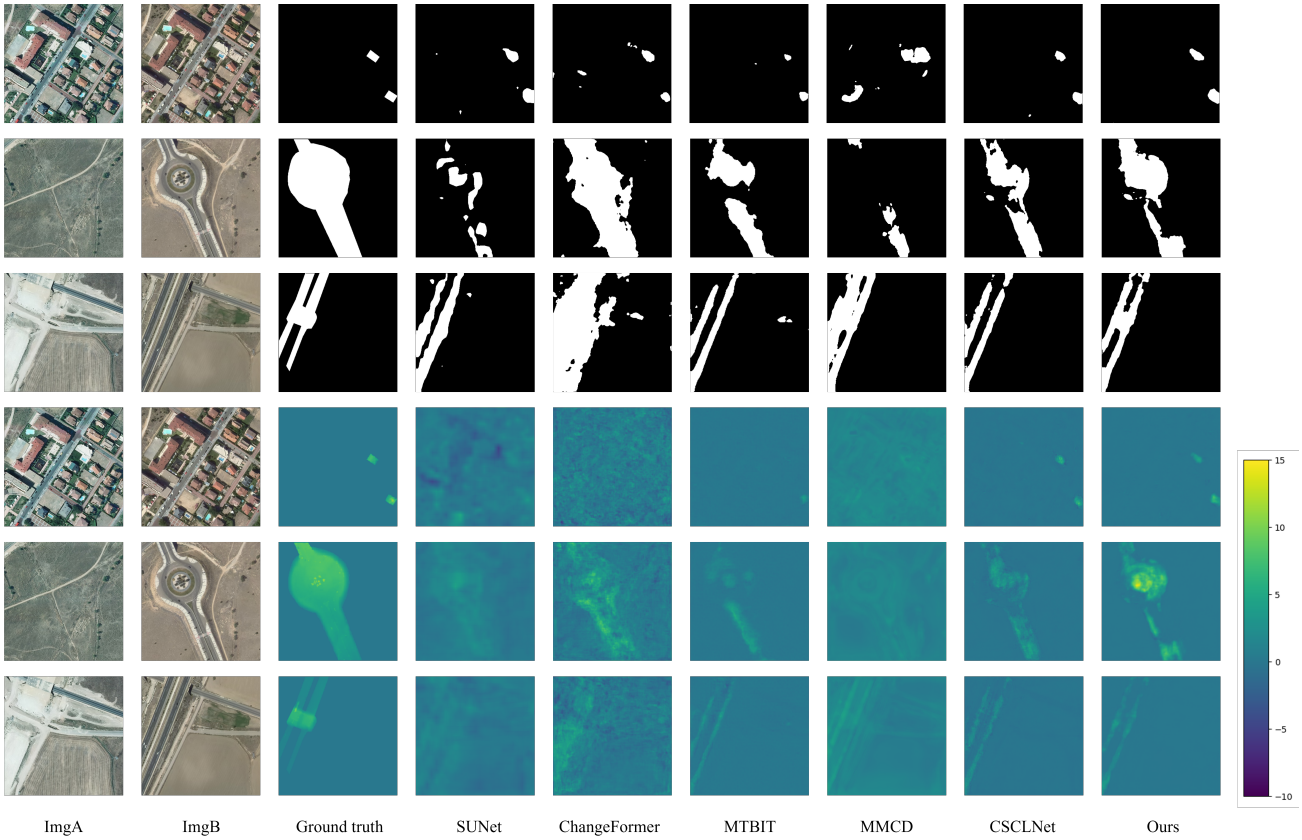

Fig. 9. Visualization results of different methods on the 3DCD Dataset. The first three rows show the 2D CD results, and the last three rows display the 3D CD results.

TABLE X
ABLATION STUDY OF THE CHANGEDA MODULES ON THE LEVIR-CD AND S2LOOKING DATASETS, WITH METRICS INCLUDING IOU AND F1-SCORE. BASELINE: CHANGEDA IMAGE ENCODER WITHOUT DEPTH ENCODER, DFE: DIFFERENTIAL FEATURE EXTRACTOR, DIM: DEPTH INFUSION MODULE, AAFF: ADAPTIVE ALL FEATURE FUSION, SUB: SUBTRACTION, CONCAT&CONV: CONCATENATION AND CONVOLUTION.

Model	Encode	Difference	Decode	LEVIR-CD	S2Looking
Baseline	Image Encoder	Sub	Concat&Conv	84.31/91.48	48.80/65.59
Baseline+DFE+AAFF	Image Encoder	DFE	AAFF	84.63/91.68	49.56/66.27
Baseline+DIM+AAFF	Image / Depth Encoder	Sub	AAFF	84.95/91.86	49.68/66.38
Baseline+DIM+DFE	Image / Depth Encoder	DFE	Concat&Conv	84.41/91.55	49.03/65.80
Baseline+DIM+DFE+AAFF	**Image / Depth Encoder**	**DFE**	**AAFF**	**85.65/92.27**	**49.72/66.42**

performance optimization.

3) Evaluating Component Effectiveness: We selected the LEVIR-CD and S2Looking datasets to conduct systematic ablation experiments on the key innovative components of our proposed ChangeDA model: Depth Infusion Module (DIM), Differential Feature Extractor (DFE), and Adaptive All Feature Fusion (AAFF). The experimental outcomes are summarized in Table X, where comparisons of IoU and F1-scores under various module combinations visually illustrate the individual contributions of each module to the model's performance. Incorporating DIM, DFE, and AAFF modules in different combinations onto the Baseline significantly improved network performance, thereby validating the value of each innovation.

Specifically, removing the DIM led to a decline in IoU by 1.02% and in F1-score by 0.59% on the LEVIR-CD dataset, and by 0.16% and 0.15% respectively on the S2Looking dataset. These decreases highlight the critical role of depth information in accurately detecting vertical structural changes in remote sensing imagery across both datasets. In comparison to the DFE module, relying solely on Sub for change information decreased the IoU by 0.70% and the F1-score by 0.41% on the LEVIR-CD dataset, demonstrating that change information extracted solely based on pixel differences fails to meet the complexity of remote sensing change detection tasks, necessitating the incorporation of additional channel similarity and semantic difference information. Moreover, employing Concat&Conv alone for fusing different feature maps in the decoding phase resulted in a reduction of IoU by 1.24% and F1-score by 0.72% on the LEVIR-CD dataset, and by 0.69% and 0.62% respectively on the S2Looking dataset. These findings indicate that adaptive fusion is crucial for processing multi-level feature maps, ensuring the network can selectively leverage semantic information from higher-level

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3532468

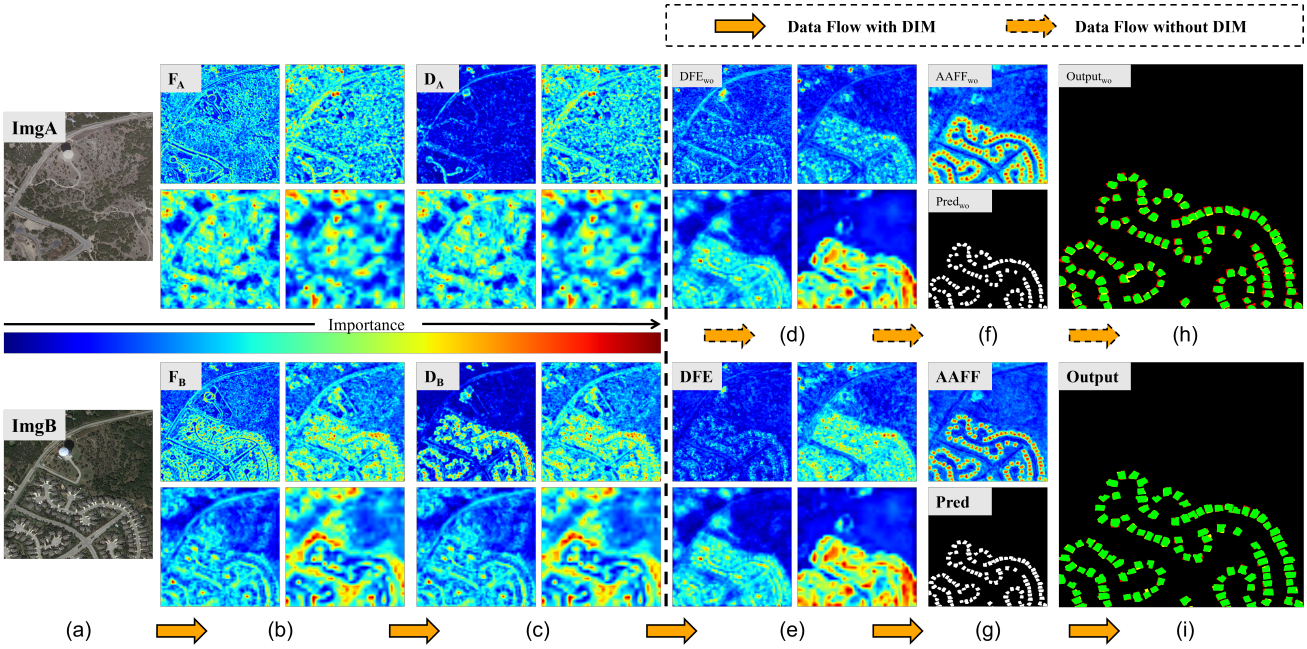

Fig. 10. Network visualization using images from the LEVIR-CD dataset as example: (a) The input image, (b) Multi-level feature maps F_A and F_B generated by the Image Encoder, (c) Multi-level feature maps D_A and D_B after passing through DIM, (d) Multi-level differential feature maps DFE_{wo} obtained from the original F_A and F_B through DFE, (e) Multi-level differential feature maps DFE obtained from D_A and D_B after passing through DFE; (f) Feature maps $AAFF_{wo}$ and prediction map Predwo after processing DFE_{wo} through AAFF, (g) Feature maps AAFF and prediction map Pred after processing DFE through AAFF; (h) and (i) Result images $Output_{wo}$ and Output, showing the comparisons between the prediction maps $Pred_{wo}$ and Pred with the Ground Truth, where green, yellow, black, and red colors represent TP, FP, TN, and FN, respectively.

feature maps and rich textural details from lower ones for accurate predictions.

The complete ChangeDA model, with its modules synergistically interacting and complementing each other's strengths, achieves optimal performance, thereby confirming the model's innovativeness and superiority.

I. Network Visualization

To qualitatively observe the role of each module in ChangeDA, we selected a pair of bi-temporal images from the LEVIR-CD dataset and visualized feature maps at different stages to visually demonstrate the shifting areas of attention during the network's operation. As shown in Fig. 10, starting with the dual-phase images ImgA and ImgB [Fig. 10(a)], we initially derive bi-temporal feature maps F_A and F_B through a backbone network [Fig. 10(b)]. Thereafter, the DIM module infuses these feature maps with depth information, generating new maps D_A and D_B [Fig. 10(c)]. With depth information incorporated, it becomes clear that the network directs enhanced focus toward taller objects. [Fig. 10(e)] presents the differential feature maps produced by the DFE module, which, through careful comparison of the depth-augmented features D_A and D_B, effectively emphasizes zones of notable change in building clusters while reducing emphasis on areas with minimal variation, such as roads and land. Following this, the AAFF module adaptively merges these cross-level differential feature maps and executes an integrated prediction [Fig. 10(g)], highlighting the model's superior discriminatory power. Specifically, the central tower, with its consistently unchanged structure and position across time, is precisely identified by semantic features, thus avoiding excessive attention to this stable area. Upon integrating lower-level features that emphasize local details, the model prudently disregards the tower, avoiding misidentification. Conversely, for the dynamically changing sections of the lower building complexes, the model astutely amplifies the integration of fine-grained features like edge contours, ensuring precise demarcation of change boundaries. This balance between maintaining stability around the tower and sensitivity to variations in building clusters exemplifies the model's precision and nuanced handling.

In highlighting the DIM module's contribution, we contrast the outcomes from processing feature maps through the DFE module without depth information [Fig. 10(d)] against those following DIM pre-processing [Fig. 10(e)]. We note that feature maps devoid of depth information tend to prioritize unchanged terrain and roads, whereas those post-DIM processing are more attuned to actual regions of transformation. Advancing further, by applying the AAFF module to fuse and predict based on these contrasting scenarios, we attain prediction outcomes both without [Fig. 10(f)] and incorporating [Fig. 10(g)] depth information integration. Fig. 10(h) (absence of DIM) and Fig. 10(i) (with DIM) present the comparison of prediction outcomes against Ground Truth, with green signifying TP, red denoting FN, and yellow indicating FP. Through meticulous analysis, we conclude that integrating DIM markedly improves the comprehensiveness and precision of object recognition, effectively decreasing false positives by reducing mispredictions of unmodified areas and minimizing false negatives by capturing more true changes. This compar-

ative finding directly affirms the significant effect of the depth information fusion strategy on boosting model performance and curtailing both false alarms and missed detections, solidly endorsing the critical role of the DIM module in enhancing the accuracy of remote sensing image change detection.

V. Conclusion

In this paper, we introduced ChangeDA, a depth-augmented multi-task network designed to enhance the effectiveness of RSCD. ChangeDA leverages depth information extracted from optical images to improve the detection of subtle changes and structural details in three-dimensional space. The multi-task learning framework enables depth estimation to act as an auxiliary task, sharing feature maps with the primary change detection task. This synergy boosts the accuracy of depth estimation and refines the granularity and detail of change detection, significantly improving overall performance. The network uses a DIM to integrate depth information into the dual-temporal feature maps, enhancing the network's ability to perceive changes over time. The DFE focuses on extracting differences between these feature maps, aiding in the precise localization of changes. The AAFF optimizes the fusion of multi-source features, ensuring that the most relevant information is emphasized during change detection. Our experimental results on prominent single-modal and multi-modal datasets demonstrate the robust adaptability and effectiveness of ChangeDA. Compared to state-of-the-art single-modal networks, ChangeDA captures detailed and subtle changes more effectively due to the integration of depth information. When evaluated against multi-modal networks, ChangeDA shows a competitive edge in handling both 2D and 3D change detection tasks, thanks to its innovative design and effective use of multi-task learning.

References

[1] E. Zhang, H. Zong, X. Li, M. Feng, and J. Ren, "Icsf: Integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024, in press.

[2] U. Stilla and Y. Xu, "Change detection of urban objects using 3d point clouds: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 228–255, 2023.

[3] Z. Fang, J. Ren, J. Zheng, R. Chen, and H. Zhao, "Dual teacher: Improving the reliability of pseudo labels for semi-supervised oriented object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024, in press.

[4] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, 2022.

[5] Y. Li, J. Ren *et al.*, "Cbanet: an end-to-end cross band 2-d attention network for hyperspectral change detection in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, May 2023.

[6] R. Qin, J. Tian, and P. Reinartz, "3d change detection–approaches and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 41–56, 2016.

[7] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103165, 2023.

[8] B. Liu, H. Chen, K. Li, and M. Y. Yang, "Transformer-based multimodal change detection with multitask consistency constraints," *Information Fusion*, vol. 108, p. 102358, 2024.

[9] Y. Xie, X. Yuan, X. X. Zhu, and J. Tian, "Multimodal co-learning for building change detection: A domain adaptation framework using vhr images and digital surface models," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[11] M. F. Reyes, Y. Xie, X. Yuan, P. d'Angelo, F. Kurz, D. Cerra, and J. Tian, "A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 74–97, 2023.

[12] Q. Li, Y. Shi, S. Auer, R. Roschlaub, K. Möst, M. Schmitt, C. Glock, and X. Zhu, "Detection of undocumented building constructions from official geodata using a convolutional neural network," *Remote Sensing*, vol. 12, no. 21, p. 3537, 2020.

[13] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *arXiv preprint arXiv:1802.10249*, 2018.

[14] V.-C. Miclea and S. Nedevschi, "Dynamic semantically guided monocular depth estimation for uav environment perception," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[15] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4009–4018.

[16] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, pp. 151–172, 2007.

[17] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *ECCV*. Springer, 2008, pp. 28–42.

[18] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, 2008.

[19] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Machine Intelligence Res.*, vol. 20, no. 6, pp. 837–854, 2023.

[20] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[21] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*. Springer, 2012, pp. 746–760.

[22] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Inf. Proc. Syst.*, vol. 27, 2014.

[23] L. Wang, J. Zhang, Y. Wang, H. Lu, and X. Ruan, "Cliffnet for monocular depth estimation with hierarchical embedding loss," in *ECCV*. Springer, 2020, pp. 316–331.

[24] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, and J. Li, "Transformer-based dual relation graph for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 163–172.

[25] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[26] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4738–4747.

[27] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4756–4765.

[28] M. Fontana, M. Spratling, and M. Shi, "When multitask learning meets partial supervision: A computer vision review," *Proceedings of the IEEE*, 2024.

[29] Q. Shen, J. Huang, M. Wang, S. Tao, R. Yang, and X. Zhang, "Semantic feature-constrained multitask siamese network for building change detection in high-spatial-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 189, pp. 78–94, 2022.

[30] F. Cui and J. Jiang, "Mtscd-net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103294, 2023.

[31] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[32] D. Hong, C. Qiu, A. Yu, Y. Quan, B. Liu, and X. Chen, "Multi-task learning for building extraction and change detection from remote sensing images," *Applied Sciences*, vol. 13, no. 2, p. 1037, 2023.

[33] V. Marsocci, V. Coletta, R. Ravanelli, S. Scardapane, and M. Crespi, "Inferring 3d change detection from bitemporal optical images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 325–339, 2023.

[34] W. Xiao, H. Cao, Y. Lei, Q. Zhu, and N. Chen, "Cross-temporal and spatial information fusion for multi-task building change detection using multi-temporal optical imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 132, p. 104075, 2024.

[35] P. Ma, J. Ren *et al.*, "Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, March 2023.

[36] H. Tolie, J. Ren, R. Chen, H. Zhao, and E. Elyan, "Blind sonar image quality assessment via machine learning: Leveraging micro-and macro-scale texture and contour features in the wavelet domain," *Engineering Applications of Artificial Intelligence*, vol. 141, p. 109730, 2025.

[37] T. Liu, M. Gong, D. Lu, Q. Zhang, H. Zheng, F. Jiang, and M. Zhang, "Building change detection for vhr remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.

[38] H. Lee, K. Lee, J. H. Kim, Y. Na, J. Park, J. P. Choi, and J. Y. Hwang, "Local similarity siamese network for urban land change detection on remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4139–4149, 2021.

[39] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[40] Y. Ye, M. Wang, L. Zhou, G. Lei, J. Fan, and Y. Qin, "Adjacent-level feature cross-fusion with 3d cnn for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[41] Y. Huang, X. Li, Z. Du, and H. Shen, "Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[42] M. Wang, B. Zhu, J. Zhang, J. Fan, and Y. Ye, "A lightweight change detection network based on feature interleaved fusion and bi-stage decoding," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[43] H. Zheng, M. Gong, T. Liu, F. Jiang, T. Zhan, D. Lu, and M. Zhang, "Hfa-net: High frequency attention siamese network for building change detection in vhr remote sensing images," *Pattern Recognit.*, vol. 129, p. 108717, 2022.

[44] H. Tolie, J. Ren, and E. Elyan, "Dicam: Deep inception and channel-wise attention modules for underwater image enhancement," *Neurocomputing*, vol. 584, p. 127585, June 2024.

[45] T. Bai, L. Wang, D. Yin, K. Sun, Y. Chen, W. Li, and D. Li, "Deep learning for change detection in remote sensing: a review," *Geo-spatial Inf. Sci.*, vol. 26, no. 3, pp. 262–288, 2023.

[46] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[47] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[48] X. Ma, J. Yang, R. Che, H. Zhang, and W. Zhang, "Ddlnet: Boosting remote sensing change detection with dual-domain learning," *arXiv preprint arXiv:2406.13606*, 2024.

[49] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auton. Robots*, vol. 42, pp. 1301–1322, 2018.

[50] G. Cheng, G. Wang, and J. Han, "Isnet: Towards improving separability for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[51] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.

[52] Z. Mao, X. Tong, Z. Luo, and H. Zhang, "Mfatnet: Multi-scale feature aggregation via transformer for remote sensing image change detection," *Remote Sens.*, vol. 14, no. 21, p. 5379, 2022.

[53] G. Pei and L. Zhang, "Feature hierarchical differentiation for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[54] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS*. IEEE, 2022, pp. 207–210.

[55] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.

[56] H. Zhong, C. Wu, and Z. Xiao, "Lrnet: Change detection of high-resolution remote sensing imagery via strategy of localization-then-refinement," *arXiv preprint arXiv:2404.04884*, 2024.

[57] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2023.

[58] C. Han, C. Wu, M. Hu, J. Li, and H. Chen, "C2f-semicd: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[59] C. Han, C. Wu, and B. Du, "Hcgmnet: A hierarchical change guiding map network for change detection," in *IGARSS*. IEEE, 2023, pp. 5511–5514.

[60] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, 2020.

[61] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[62] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "Hanet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[63] M. Lin, G. Yang, and H. Zhang, "Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images," *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2022.

[64] Z. Dong, S. An, J. Zhang, J. Yu, J. Li, and D. Xu, "L-unet: A landslide extraction model using multi-scale feature fusion and attention mechanism," *Remote Sens.*, vol. 14, no. 11, p. 2552, 2022.

[65] L. Wang and H. Li, "Hmcnet: Hybrid efficient remote sensing images change detection network based on cross-axis attention mlp and cnn," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[66] Z. Li, C. Yan, Y. Sun, and Q. Xin, "A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.