AdaptFormer: An Adaptive Hierarchical Semantic Approach for Change Detection on Remote Sensing Images

Teng Huang, Yile Hong, Yan Pang, *Member, IEEE*, Jiaming Liang, Jie Hong, Lin Huang, Yuan Zhang, Yan Jia, and Patrizia Savi, *Senior Member, IEEE*

***Abstract*— Change detection (CD) in remote sensing (RS) aims to consistently track alterations in specific regions over time. While current methods employ hierarchical architectures to analyze semantic details, they often miss crucial changes across different semantic levels, resulting in partial representations of environmental shifts. Addressing this, we propose AdaptFormer, uniquely designed to adaptively interpret hierarchical semantics. Instead of a one-size-fits-all approach, it strategizes differently across three semantic depths: employing straightforward operations for shallow semantics, assimilating spatial data for medium semantics to emphasize detailed interregional changes, and integrating cascaded depthwise attention for in-depth semantics, focusing on high-level representations. The experimental evaluations reveal that AdaptFormer surpasses many leading benchmarks, showcasing exceptional accuracy on LEVIR-CD and DSIFN-CD datasets. AdaptFormer showcases impressive performance with F1 and intersection over union (IoU) scores of 92.65% and 86.31% on the LEVIR-CD dataset, and 97.59% and 95.29% on the DSIFN-CD dataset, respectively. The datasets are available at https://github.com/aigzhusmart/AdaptFormer.**

***Index Terms*— Change detection (CD), deep learning, hierarchical representation learning, remote sensing (RS), representation fusion.**

I. Introduction

CHANGE detection (CD) has emerged as a crucial field of remote sensing (RS), primarily focusing on the systematic identification of alterations within a region [1], [2]. This identification is realized through the comparative analysis of images captured at distinct temporal intervals [3]. By leveraging the concept of binary labeling for each pixel, CD techniques facilitate the automated extraction of pertinent information [4]. The strength of contemporary CDs largely stems from their ability to extract and compare semantic information [5]. This process empowers the techniques to identify, characterize, and comprehend changes within RS data. The insights gleaned from this process are invaluable, driving informed decision-making across a plethora of applications, including urban development [6], disaster management [7], deforestation [8], environmental surveillance [9], [10], etc.

The CD in RS represents a significant challenge due to the need for meticulous analysis and comparison of coregistered images obtained at different time points. Existing methodologies [11], [12] employ complex hierarchical architectures, where semantic information is dissected and compared across various levels. A common category of CD techniques emphasizes detecting changes predominantly at the deepest levels [13], [14]. Although this approach yields a detailed understanding of advanced-level changes, it may overlook critical alterations at more rudimentary layers, potentially resulting in an incomplete depiction of overall environmental transformations.

An alternative set of CD techniques involves a systematic and repeated extraction of semantic information at each hierarchical level, followed by an exhaustive comparison of this data [15], [16]. However, this method tends to lack nuanced interpretation across the levels and may result in inaccuracies. Specifically, the simplistic and repeated comparison process might fail to detect intricate inter-level relationships, or it might disproportionately emphasize certain changes, thereby affecting the overall quality and accuracy of change detection (CD). The existing challenges highlight the urgent need for an efficient investigative manner for ensuring accurate and comprehensive analysis across all semantic levels in RS applications.

The hierarchical structure of RS image analysis allows for the extraction of semantic information at various depths, each possessing distinct characteristics and challenges [17], [18], [19]. Shallow semantic information, gleaned from the initial layers of the hierarchy, is adept at identifying rudimentary features such as edges and basic shapes but may struggle with

Manuscript received 15 November 2023; revised 21 February 2024; accepted 20 March 2024. Date of publication 11 April 2024; date of current version 24 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB2704300, and in part by the Scientific Research Project for Guangzhou University under Grant YJ2023041. The Associate Editor coordinating the review process was Dr. George Dan Mois. *(Corresponding author: Yan Pang.)*

Teng Huang, Yile Hong, Yan Pang, Jiaming Liang, and Jie Hong are with the Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China (e-mail: huangteng1220@gzhu.edu.cn; yile.hong@e.gzhu.edu.cn; yanpang@gzhu.edu.cn; jiaming.liang@e.gzhu.edu.cn; hongjie@e.gzhu.edu.cn).

Lin Huang is with the Department of Engineering and Engineering Technology College of Aerospace, Computing, Engineering, and Design (ACED), Metropolitan State University of Denver, Denver, CO 80204 USA (e-mail: LHuang1@msudenver.edu).

Yuan Zhang is with the School of Information, North China University of Technology, Beijing 100144, China (e-mail: zhangyuan@ncut.edu.cn).

Yan Jia is with the Department of Surveying and Geoinformatics, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: jiayan@njupt.edu.cn).

Patrizia Savi is with the Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy (e-mail: patrizia.savi@polito.it).

Digital Object Identifier 10.1109/TIM.2024.3387494

intricate details, particularly when considering the tiny objects frequently found in RS images [20], [21]. Medium semantic information, sourced from intermediate layers, recognizes complex shapes and patterns with increased accuracy but can overlook subtler details or minor objects. Conversely, deep semantic information from advanced layers can comprehend broader contextual relationships and substantial structures but can neglect smaller objects or nuanced changes [22], [23]. Given the unique challenges presented by the numerous small objects common in RS images, it is crucial to develop an adaptive method that efficiently extracts semantic information at different levels based on their inherent properties. Such an approach to CD would improve accuracy and efficiency and would be of particular value in RS applications.

In order to solve the above challenges, we present AdaptFormer, a novel framework that probes into hierarchical semantic interpretations. The AdaptFormer deviates from the conventional method by systematically and repetitively investigating semantic information at each hierarchical level. Instead, it adopts an adaptive technique for interpreting hierarchical representations at three distinct semantic stages: shallow, medium, and deep, as illustrated in Fig. 1. This framework progressively captures salient semantic representations, aligning with the idiosyncrasies of different hierarchical architecture states in RS imagery. For shallow semantics associated with small objects, AdaptFormer employs straightforward operations to identify local representations. In contrast, for medium semantics, it assimilates spatial information to accentuate finer interregional details across different temporal intervals. Furthermore, it introduces cascaded depthwise attention for deep semantics, thereby enabling the effective learning of high-level representations. Rigorous testing against 11 established benchmarks on popular CD datasets, including LEVIR-CD and DSIFN-CD, attests to the superior performance of AdaptFormer, marking it as a trailblazer in the realm of CD. In addition, AdaptFormer holds significant potential value in the industrial domain, with applications extending to areas such as agricultural CD [24], land use change analysis [25], deforestation monitoring [8], flood monitoring [26], climate change impact assessment [27], and water body CD [28].

The main contributions in this article are summarized as follows.

1) We present an innovative, end-to-end approach called AdaptFormer enables the adaptive interpretation of hierarchical representations for CD on RS imagery.
2) Designed for precise and differentiated semantic interpretation at multiple hierarchical levels, AdaptFormer implements unique strategies across shallow, medium, and deep semantic layers, showcasing its versatility and specificity.
3) The AdaptFormer outperforms various established CD baselines, setting new records on two benchmark datasets, LEVIR-CD and DSIFN-CD.

II. Related Work

In the field of CD, techniques have emerged in tandem with the rise of aerial imagery technology, increasingly gaining importance in managing large-scale image data [1], [29]. The FC series approaches, encompassing FC-EF, FC-Siam-DI, and FC-Siam-Conc, first incorporate the fully convolutional neural network architecture into CD tasks [30]. These methodologies are remarkable for their ability to be applied to any RS CD dataset. However, their performance is often compromised by disruptive elements like shadows and backgrounds, leading to misinterpretation of image features. Responding to these challenges, newer techniques such as DTCDSCN, STANet, and DASNet [6], [31], [32] integrate attention modules into their frameworks, leveraging interdependencies between channels and spatial positions to enhance feature perception.

As we transition into a newer era of CD, the robust representational capabilities of the Transformer model have received increased attention, showcasing comparable performance to convolutional models in various visual tasks. In fact, BiT [33] integrates the Transformer model with convolution layers. The ChangeFormer [15] supports the idea that the Transformer encoder on its own is capable of extracting fundamental features, analyzing intricate details from dual-temporal images, and integrating feature differences at various scales. Then, Changer [34] introduces feature interaction to allow the sharing of feature information between two branches of a network, thereby improving the perception of contextual semantic information differences. Despite these advancements, both ChangeFormer and Changer fall short in differentiating cross-level feature information due to their uniform module usage for semantic extraction at varying levels. Addressing these limitations, our proposed AdaptFormer emphasizes the differences in semantic information between different levels and adaptively employs selective modules for shallow, medium, and deep semantic layers, thereby demonstrating its versatility and specificity.

III. Method

In this section, we introduce the architecture of a pioneering framework designated as AdaptFormer, devised for the purpose of CD. This framework harnesses the power of an adaptive, transformer-based model arranged in a hierarchical fashion, which is described in detail in Section III-A.

A. Hierarchical Adaptive Mechanism

AdaptFormer is a cutting-edge architecture that prioritizes adaptive feature learning and comparative analysis. Designed to cater to the intrinsic hierarchical semantic features, it delves into various representation levels: shallow, medium, and deep. This methodical approach to feature learning unfolds across three distinct stages, with the pivotal difference module bolstering each stage's unique operations. The intricate details of its structure, inclusive of the operational nuances and the integral role of the difference module, are depicted in Fig. 1.

AdaptFormer's operational flow begins with the intake of two sets of images, which represent the same geographical region captured at different time intervals, referred to as pre-change and post-change images. These images are processed through a sequence of three differentiated stages. Each stage involves the essential tasks of downsampling and feature

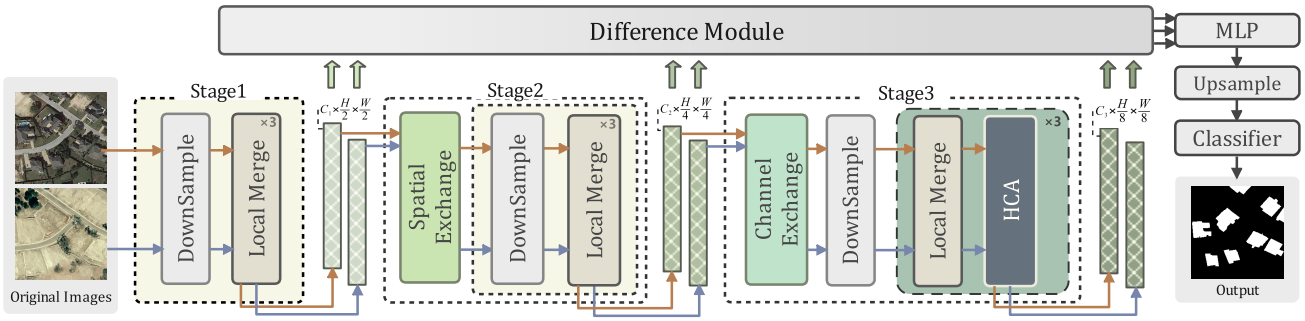

Fig. 1. Schematic representation of the AdaptFormer architecture. The proposed AdaptFormer employs distinct strategies from straightforward operations for shallow levels, spatial data assimilation for medium levels, to cascaded depthwise attention for deeper semantics.

selection, applied in a manner that respects the semantic depth associated with each stage. As a culmination of these stages, the differences in the resulting outputs are fused by the difference module. This module computes the dissimilarities between the stage outputs and then undergoes an upsampling process to match the size of the original input images. This systematic approach ensures a comprehensive analysis and comparison of changes at various semantic levels, reinforcing the accuracy of the CD process.

Our proposed AdaptFormer implements an ingenious design to facilitate adaptive feature learning and comparison, effectively catering to the varied levels of representation, i.e., shallow, medium, and deep, inherent in hierarchical semantic features. In essence, the system integrates a local merge module at each stage, enhancing the model's feature extraction capabilities, and thus optimizing the utility of semantic information across different levels in RS images. These stages also encompass the introduction of stage-specific modules, such as the spatial exchange module in stage 2, designed to augment the model's performance by bolstering precise semantic interpretations.

Moving deeper into the system, stage 3 benefits from the addition of the channel exchange module [34] and the hierarchical collaborative attention (HCA) module. These modules are instrumental in adapting to more abstract information encapsulated within deeper-level semantics, leading to favorable segmentation results. Remarkably, AdaptFormer's design provides for the relative independence of the encoders that process pre-change and post-change images, contributing to the system's robustness. Each stage within an encoder operates on a distinct set of images, employing the difference module to facilitate difference detection of image processing results across various time domains. Such a methodology, harnessing both the independence of image processing and the interconnectedness of module application, contributes to AdaptFormer's superior performance in CD.

1) Stage 1—Shallow Semantic: As the initiating phase of the AdaptFormer, stage 1 is integral for the selection and extraction of rudimentary, or shallow, semantic features. The image being processed, denoted as X_{in} with dimensions $W \times H \times C$ (representing width, height, and channels, respectively), is subjected to downsampling by the Downsample module. The Downsample module, employing a 3×3 convolution operation and group normalization

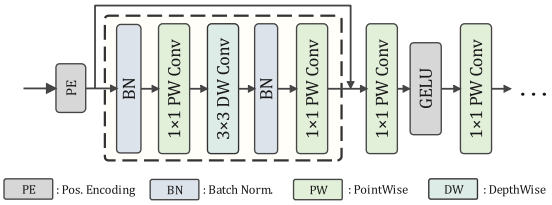

Fig. 2. Structure of local merge.

with a stride of 2, modifies X_{in} to a dimensionality of $(W/2) \times (H/2) \times C$. The output tensor, consequent to the downsampling process, primarily encapsulates basic shallow semantic information such as shapes and textures. To efficiently manage these features, we integrate the local merge module at this juncture of the framework.

Local merge prioritizes dual learning in spatial and channel dimensions of the data, as shown in Fig. 2. Utilizing depthwise separable convolution, it aggregates local features across both domains, enriching data analysis. This approach promotes the integration of channel-specific information into input features, thereby elevating the predictive accuracy of the CD model. Equation (1) provides an in-depth mathematical insight into the local merge module's operations

$$\begin{aligned} X_1 &= \text{PW}(\text{BN}(\text{PE}(X_{\text{in}}))) \\ X_2 &= \text{DW}(X_1) \\ X_3 &= \text{PW}(\text{BN}(\text{DW}(X_2)) \\ Y &= \text{PW}(\varphi(\text{PW}(X_3))) \end{aligned} \tag{1}$$

where BN and φ denote batch normalization and GELU activation functions [35]. Y represents the output of the local merge module that employs a position-wise (PW) and a depth-wise (DW) convolutional layer, designed for effective local feature aggregation. The PW convolves input data across spatial dimensions, while DW focuses on local feature aggregation. This structure is augmented by a depthwise convolution layer, or PE, extracting relative positional information to enhance image understanding. Through this configuration, the local merge module efficiently generates rich semantic features, vital for precise CD.

2) Stage 2—Medium Semantic: In stage 2 of our model, the emphasis is placed on the adept extraction and processing of intermediate-level semantics, characterized by their abstract and semantically rich attributes. This contrasts with

the more rudimentary characteristics inherent to shallow-level semantics. In order to address the challenges associated with extracting these complex features, we have integrated the spatial exchange module into stage 2. This module is an enhancement over stage 1, capitalizing on the associational strength inherent to intermediate-level semantics by evaluating diverse spatial perspectives present in data channels. Consequently, this strategic augmentation facilitates a more robust capability for the extraction and interpretation of abstract features synonymous with intermediate-level semantics. The details of spatial exchange are as follows.

Spatial exchange plays a pivotal role in CD models by adeptly integrating change region features. These features are learned through a dual-encoder system, highlighting the intricate interplay of correlations across varied temporal domains. A defining characteristic of this integration is the exchange of grayscale images stemming from the double temporal domain processing outcomes, all while operating at half the spatial dimension. This strategic inclusion bolsters the CD model's proficiency and amplifies its capability to forge spatial object associations [34]. Specifically, the execution flow of spatial exchange is shown in the following equation:

$$
\begin{aligned}
M_i &= \begin{cases} 1, & \text{if } i \bmod \alpha = 0 \\ 0, & \text{otherwise} \end{cases} \\
Y_e &= X_e \odot M + \hat{X}_e \odot (1 - M) \\
\hat{Y}_e &= X_e \odot (1 - M) + \hat{X}_e \odot M
\end{aligned} \tag{2}
$$

where e represents the dimension that the input feature needs to be exchanged, α represents the channel exchange mask displacement, M_i represents the ith element of the 1-D mask M, and X_e, $\hat{X}_e$, Y_e, $\hat{Y}_e$ represent the representation of X, $\hat{X}$, Y, $\hat{Y}$ in the channel dimension, respectively.

In stage 2, we designate e as the width (W) dimension of the input features and $\alpha = 2$. This deliberate selection enables the effective comparison and fusion of middle-level semantic features across distinct temporal instances, effectively capturing the relational information between diverse spatial regions.

Subsequently, the exchanged feature vectors continue to undergo further processing through the Downsample module and the local merge module. The resulting processed feature vectors are then fed into the difference module and subsequently passed on to the next stage for subsequent analysis or utilization.

3) Stage 3—Deep Semantic: After stage 2, stage 3 processes semantic features related to objects, scenes, or advanced concepts. These features' global information is vital for quality CD results. Understanding the interplay between encoders representing the same region at different times enhances the model's grasp of temporal relations between spatial elements in a scene. Consequently, we integrated channel exchange and HCA modules in stage 3. Details of these modules are presented below.

Channel exchange contrasts with spatial exchange by operating in the channel dimension, where it swaps half of the input images from both sides based on (2) with e set as the channel (C) dimension. This approach avoids the potential spatial ambiguity that might arise from exchanging features in the plane dimension. Exchanging along the channel dimension enhances the capture of deep semantic interactions across temporal instances within a specific region. Following this exchange, the feature vectors proceed to the local merge and HCA modules.

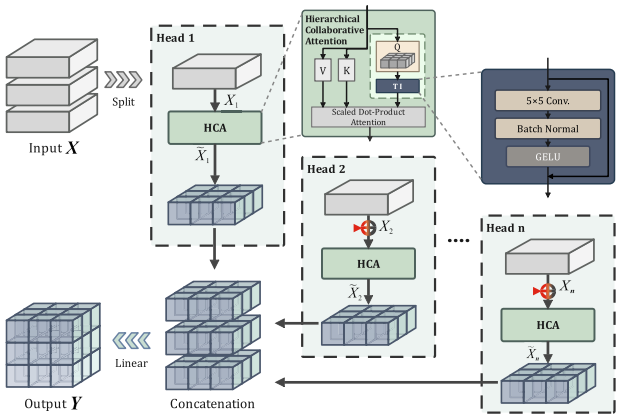

Fig. 3. Overview of HCA.

HCA is designed to discern spatial relationships in the input image through feature clipping and attention computations. It extracts refined global features from a feature vector rich in temporal and abstract semantic information. The HCA's workflow is depicted in Fig. 3, with its computational details provided in the following equation:

$$
\begin{aligned}
&[X_1, X_2, \ldots, X_{i-1}, X_i, \ldots, X_n]_d = X_{\text{in}} \\
&X_i = \tilde{X}_{i-1} + X_i \\
&\tilde{X}_i = \text{Attn}\left(X_i W_i^Q, X_i W_i^K, X_i W_i^V\right) \\
&Y = \tilde{X}_1 \parallel \tilde{X}_2, \ldots, \parallel \tilde{X}_{i-1} \parallel \tilde{X}_i, \ldots, \parallel \tilde{X}_n
\end{aligned} \tag{3}
$$

where n denotes the number of segments and Y represents the output, with X_i as the ith segment of input X_{in}. After the *Attn* operation, X_i yields $\tilde{X}_i$. Here, W_i^Q, W_i^K, and W_i^V are projection layers mapping input features into distinct subspaces, and the $\parallel$ indicates the concatenation.

The HCA is designed to enhance the handling of feature vectors. By partitioning data along the channel dimension, C, it allows for individualized attention computations on each segment, streamlining the computational process and boosting model parallelism. The model's understanding of local structures in input images is further enriched by incorporating a sequence of convolution, batch normalization, and the GELU activation function after the query phase. To preserve information throughout the process, a residual connection is integrated.

A significant trait of HCA is its feedback mechanism. The output from one attention computation serves as the input for the subsequent one, reinforcing feature representation. Given the depth of semantic feature analysis, the model determines that a partition count (n) of four is optimal for extracting global features. Within stage 3, the combination of three HCAs with local merge modules forms the backbone, drawing out deep semantic features and enhancing the model's proficiency in CD.

4) Difference Module: The difference module calculates the variance between pre-change and post-change image encodings produced at each stage. By merging the two outputs in the channel CC dimension, their distinctions are discerned using convolutional operations. This computation procedure is detailed in the following equation:

$$\begin{aligned} X &= \mathrm{DW}(X_1 \parallel X_2) \\ D &= \mathrm{DW}(\mathrm{BN}(\sigma(X)) \end{aligned} \tag{4}$$

where X_1 and X_2, respectively, represent the output of two encoders in the same stage, the σ is the RELU function [36], and D represents the output of the difference module.

B. Loss Function

To facilitate the CD task, we consider employing the cross-entropy loss function [37] for training the model, which is expressed by the following equation:

$$\mathcal{L}_{\mathrm{ce}}(G, Y) = -\frac{1}{N}\sum_{i=1}^{N}\big[Y(i)\log(G(i)) + (1 - Y(i))\log(1 - G(i))\big] \tag{5}$$

where N represents the number of pixels in the input binary masks, G represents the real binary masks of the changed region, and Y represents the predicted CD mask.

Since the outputs of different levels contain feature representations with different levels of abstraction, by using the multilayer output to calculate the loss, these features can be considered comprehensively, thereby improving the modeling ability of the target task. This loss calculation can be expressed by the following equation:

$$\begin{aligned} \mathcal{L}_3 &= \mathcal{L}_{\mathrm{ce}}(G, \mathrm{Up}(\mathrm{fuse}(D_3)) \\ \mathcal{L}_2 &= \mathcal{L}_{\mathrm{ce}}(G, \mathrm{Up}(\mathrm{fuse}(D_2 + D_3)) \\ \mathcal{L}_1 &= \mathcal{L}_{\mathrm{ce}}(G, \mathrm{Up}(\mathrm{fuse}(D_1 + D_2 + D_3))) \\ \mathcal{L}_{\mathrm{total}} &= \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 \end{aligned} \tag{6}$$

where D_1, D_2, and D_3 represent the results of each stage after passing through the difference modules. The Up operation is to upsample the input tensor size to G size. The details of the fuse operation are as follows:

$$\begin{aligned} D &= \mathrm{BN}(\sigma(\mathrm{DW}(D_{\mathrm{in}}))) \\ \mathrm{fuse}(D_{\mathrm{in}}) &= \mathrm{DW}(D) \end{aligned} \tag{7}$$

where $\mathcal{L}_j$ indicates that the output of the jth stage is cross-entropy calculated with G, and the coefficient λ_j before each layer loss $(\lambda_j > 0) j \in \{1, 2, 3\}$. We use the total loss $\mathcal{L}_{\mathrm{total}}$ to measure model capability.

IV. Experiments and Discussion

A. Datasets

We evaluate the performance of the CD task using two large-scale remote building CD sensing datasets.

LEVIR-CD [6], a benchmark dataset for building CD, comprises 637 bitemporal image patch pairs sourced from Google Earth, each having a very high resolution of 0.5 m/pixel and dimensions of 1024 × 1024 pixels. Spanning a time frame of 5–14 years, these images vividly capture significant land-use transformations, especially construction growth. The dataset encompasses a variety of building morphologies, from villa residences and tall apartments to small garages and large warehouses. Primarily emphasizing building-related dynamics, it specifically categorizes changes as building growth or decline. Expert RS interpreters annotated these images with binary labels, denoting change (1) or no change (0), with every annotation undergoing a rigorous double-check process to ensure accuracy. For experimental divisions, patches of size 256 × 256 yielded 7120, 1024, and 2048 samples for training, validation, and testing sets, respectively.

DSIFN-CD [38] dataset comprises six large, bitemporal, high-resolution images that span six Chinese cities, namely Beijing, Chengdu, Shenzhen, Chongqing, Wuhan, and Xian. Initially obtained manually from Google Earth, the images are pre-processed into default pairs with dimensions of 512 × 512 pixels. For experimental consistency, these are further segmented into non-overlapping 256 × 256 blocks, yielding 14 400 training, 1360 validation, and 192 test samples.

B. Evaluation Metrics

F1-score (F1) [39] is a statistical measure used in the context of binary and multiclass classification to evaluate a model's accuracy. The $F1$-score combines recall, which gauges correct change identification, with the minimization of false detection, serving as an overall indicator of a model's accuracy in detecting RS image changes [40]. Metric formulations are as follows:

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{8}$$

where TP represents true positives, FP denotes false positives, TN signifies true negatives, and FN refers to false negatives.

Intersection over union (IoU) [41] is a widely adopted metric in the domain of CD using RS imagery to gauge the agreement between predicted change areas and ground-truth (GT) annotations [40]. It quantifies the ratio of the intersecting area to the union area of the predicted and actual change regions, providing a value ranging from 0 (no overlap) to 1 (complete overlap). Metric formulations are as follows:

$$\mathrm{IoU} = \frac{Y \cap G}{Y \cup G}. \tag{9}$$

Overall accuracy (OA) [42] serves as a performance metric to evaluate the proportion of correctly classified pixels relative to the total number of pixels in RS imagery. It provides a comprehensive measure of the model's effectiveness in accurately detecting both changed and unchanged areas across the entire spatial extent of the image under the CD task [43]. Metric formulations are as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{10}$$

Recall [44] evaluates the fraction of true positive changes that were correctly identified by a model relative to the

total actual changes [45]. This metric is crucial to gauge the model's proficiency in capturing all pertinent alterations within the satellite images, ensuring that no significant changes are overlooked [46]. Metric formulations are as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{11}$$

C. Implementation Details

AdaptFormer is trained on eight NVIDIA A100-PCIE-40G. Each GPU has a batch size of 24 with a patch size of 256 × 256. The AdamW optimizer is utilized with a cosine annealing strategy, setting an initial learning rate of 0.0006 and a weight decay of 0.05. The training procedure is configured for a total of 600 epochs. Additionally, we have configured the weights for model multilayer output and label calculation loss in a ratio of 5:5:5:8 during training, and our data loader utilizes four subprocesses to load data in parallel, improving data loading speed and efficiency.

D. CD Performance

Our experimental evaluation benchmarked AdaptFormer's performance on the LEVIR-CD and DSIFN-CD datasets, as shown in Table I. Performance was assessed using four critical metrics: F1, IoU, OA, and Recall, and juxtaposed with 11 established CD methods, including notable performers such as ChangeFormer, P2V-CD, and Changer. Each of these employed unique strategies for CD: ChangeFormer utilized the difference module to gauge the variance in decoder output feature maps, P2V-CD resolved the problem via temporal–spatial transformations, and Changer integrated feature interaction strategies, achieving metrics of 92.24%, 85.59%, 99.20%, and 91.20%, respectively.

AdaptFormer, however, through its innovative methodologies, presents an evident advancement in the performance metrics across both datasets. Specifically, on the LEVIR-CD dataset, AdaptFormer manifests scores of 92.65%, 86.31%, 99.19%, and 92.59% for the F1, IoU, OA, and Recall metrics, respectively. Despite a marginal decrement of 0.01% in the OA metric compared to Changer, the F1, IoU, and Recall metrics exhibit enhancements of 0.41%, 0.72%, and 1.39%, respectively. The superiority of AdaptFormer is further emphasized in the DSIFN-CD dataset. Here, it significantly surpasses P2V-CD, the runner-up, with an impressive F1-score of 97.59%—a striking 5.77% advancement.

E. Ablation Study

1) Stage Depth Setting: This section is dedicated to assessing the impact of depth at each model stage, denoted as N1, N2, and N3, for the first, second, and third stages, respectively. As shown in Fig. 4 with an initial configuration of [3, 3, 3], the F1, IoU, OA, and Recall values register at 92.65%, 86.31%, 99.19%, and 92.59%. It is notable that any decrease in depth at each stage reflects in a consequent decrease in all performance metrics, exemplified when N1, N2, and N3 are set to [1, 1, 3], causing decreases of 1.31%, 2.25%, 0.12%, and 2.31% in F1, IoU, OA, and Recall, respectively. This

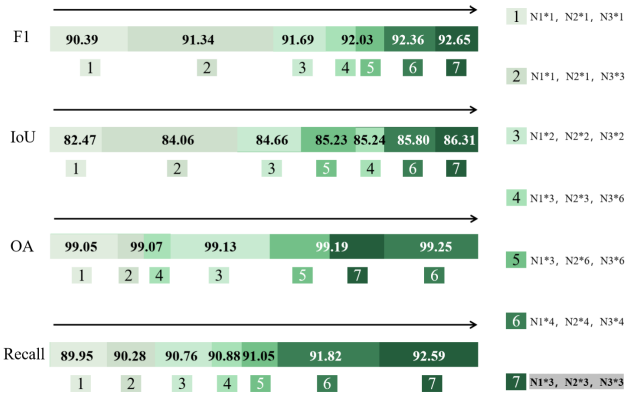

Fig. 4. Quantitative comparison with different stage depths of AdaptFormer on the LEVIR-CD dataset.

scenario implies a shortfall in feature extraction by shallow models, thereby negatively affecting accuracy. Conversely, an attempt to increase depth also instigates similar metric decreases, such as when N1, N2, and N3 are set to [3, 3, 6], resulting in decreases of 0.62%, 1.07%, 0.12%, and 1.71% in F1, IoU, OA, and Recall, respectively. Interestingly, with the configuration [4, 4, 4], the F1-value slightly elevates to 99.25%, outperforming the base by 0.06%, yet other metrics underperform, suggesting an over-extraction of deep semantic features due to excessive stages. After a thorough examination of all these dynamics, the configuration of [3, 3, 3] is retained as the optimal choice.

2) Feature Splits: Splitting input features into a specified number affects the model performance. The goal of this section is to evaluate the impact of feature splits on the model performance. As shown in Fig. 5(a), we notice that the model achieves the best performance when the feature splits are set to 4, with F1, IoU, OA, and Recall of 92.65%, 86.31%, 99.19%, and 92.59%, respectively. When the feature splits are less than 4, the model's performance decreases. For example, when the feature splits are 1, the model's F1, IoU, OA, and Recall decrease by 0.82%, 1.42%, 0.11%, and 1.42%, respectively. This is because fewer feature hierarchies are not conducive to the model learning feature representations from multiple perspectives, which leads to performance degradation. On the other hand, when the feature splits are greater than 4, the model's performance also decreases. For example, when the feature splits are set to 16, the four indicators of the model decreased by 0.50%, 0.87%, 0.05%, and 0.73%, respectively. This is due to an excessive number of feature splits causing the model to easily overfit the training data, leading to a decrease in generalization performance. Considering the above factors, we believe that setting the feature hierarchy to 4 is a reasonable choice.

3) Spatial Exchange Setting: The objective of this section is to evaluate the impact of spatial swapping positions on the model's performance for the spatial exchange module. The experimental results are shown in Fig. 5(b). When performing spatial swaps only in the h-dimension, the model's F1 and IoU are 92.45% and 85.97%, respectively. When swapping in the w-dimension, the model's performance improves, with F1 increasing by 0.20% and IoU increasing by 0.34%.

TABLE I

PERFORMANCE OF EACH MODEL IN THE LEVIR-CD DATASET AND THE DSIFN-CD DATASET. ALL VALUES ARE REPORTED IN PERCENTAGE (%). * SYMBOL INDICATES THAT THE CHANGER MODULE UTILIZES THE EXCHANGE MODULE, AND THE BACKBONE OF THE MODEL EMPLOYS RESNEST-101. THE PERFORMANCE OF OUR PROPOSED MODEL IS MARKED IN GRAY

Method	LEVIR-CD				DSIFN-CD			
	F1	IoU	OA	Recall	F1	IoU	OA	Recall
FC-EF [30]	83.40	71.53	98.39	80.17	61.09	43.98	88.59	52.73
FC-Siam-Di [30]	86.31	75.92	98.67	83.31	62.54	45.50	86.63	65.71
FC-Siam-Conc [30]	83.69	71.96	98.49	76.77	59.71	42.56	87.57	54.21
DTCDSCN [31]	87.67	78.05	98.77	86.83	63.72	46.76	84.91	77.99
STANet [6]	87.26	77.40	98.66	91.00	64.56	47.66	88.49	61.68
IFNet [38]	88.13	78.77	98.87	82.93	60.10	42.96	87.83	53.94
SNUNet [47]	88.16	78.83	98.82	87.17	66.18	49.45	87.34	72.89
BIT [33]	89.31	80.68	98.92	89.37	69.26	52.97	89.41	70.18
ChangeFormer [15]	90.40	82.48	99.04	88.80	86.67	76.48	95.56	84.94
P2V-CD [48]	91.32	83.88	99.12	89.76	91.82	84.88	96.07	90.18
Changer* [34]	92.24	85.59	**99.20**	91.20	-	-	-	-
AdaptFormer	**92.65**	**86.31**	99.19	**92.59**	**97.59**	**95.29**	**99.10**	**97.20**

TABLE II

IMPACT OF EXCHANGE OPERATIONS ACROSS DIFFERENT STAGES ON THE LEVIR-CD DATASET. FOR EACH PERFORMANCE METRIC, PERFORMANCE DECLINES ARE DENOTED IN **GREEN**, WHILE ENHANCEMENTS ARE HIGHLIGHTED IN **RED**. THE PERFORMANCE OF THE RECOMMENDED CHOICE IS MARKED IN GRAY

Setting	Stage 2	Stage 3	F1	IoU	OA	Recall
Baseline	-	-	91.93	85.02	99.06	92.01
Group I	SE	-	91.96 +0.03	85.12 +0.10	99.02 -0.04	90.82 -1.19
	-	CE	92.00 +0.07	85.12 +0.10	99.02 -0.04	91.06 -0.95
Group II	SE	SE	92.50 +0.57	85.98 +0.96	99.08 +0.02	91.36 -0.65
	CE	CE	91.84 -0.09	84.91 -0.11	99.06 -0.00	90.83 -1.18
Group III	CE	SE	90.89 -1.04	83.31 -1.71	98.94 -0.12	90.53 -1.48
	SE	**CE**	**92.65** +0.72	**86.31** +1.29	**99.19** +0.13	**92.59** +0.58

However, when both the h-dimension and w-dimension are swapped simultaneously, compared to swapping only in the w-dimension, the model's F1 and IoU decrease by 0.41% and 0.71%, respectively. This is because the spatial exchange module's effectiveness lies in providing the encoder with semantic information from another temporal aspect, while the encoder itself plays a crucial role in extracting semantic features from the current temporal aspect. The excessive information exchange during swapping in both the h-dimension and w-dimension causes the encoder to lose too much image feature information, leading to a suboptimal extraction of semantic features for the current temporal aspect and resulting in performance degradation. Therefore, we choose the w-dimension as the spatial swapping position for the spatial exchange module.

4) Exchange Positions: Building on the established spatial exchange settings from earlier experiments, this section specifically investigates how spatial and channel exchanges are positioned across stages 2 and 3, with findings outlined in Table II. The baseline performance metrics, derived from a model without either exchange module and serving as a control, are as follows: 91.93% for F1, 85.02% for IoU, 99.06% for OA, and 92.01% for Recall, as indicated in the first row of Table II.

In the first comparative experiment (Group I), the model was tested with only a spatial exchange in stage 2 or a channel exchange in stage 3. Results reflected a slight increase of less than 0.1% in F1 and IoU, while observing substantial drops

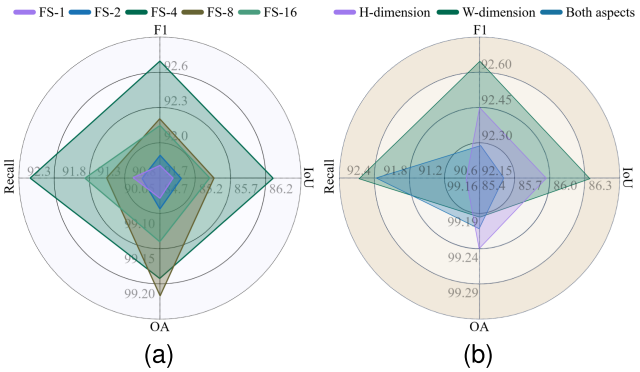

Fig. 5. Conducting a comparative quantitative analysis involves examining (a) various feature splits within the HCA module and (b) diverse dimensional configurations in the spatial exchange module. Both aspects are evaluated using the LEVIR-CD dataset. FS: feature split.

in OA and Recall by 1.19% and 0.95%, respectively, hinting that isolating feature dimension transformations might hamper the overall model efficiency.

The second comparison (Group II) aimed to discern the effect of utilizing identical exchange modules, either channel or spatial, in both stages. Introducing the channel exchange too prematurely, especially when semantics were not adequately deep, led to a retention of redundant information from the medium stage, which negatively influenced the deep-stage feature comparison. Specifically, this resulted in a decline in

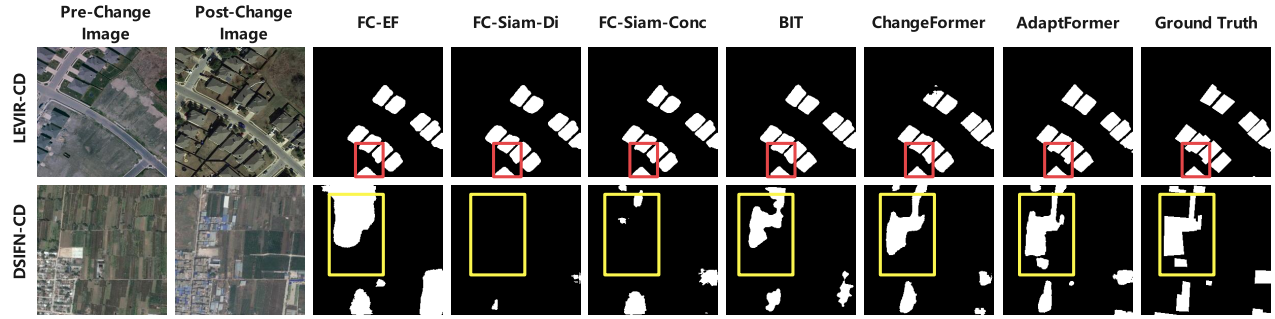

Fig. 6. Comparative display of CD performance from divergent CD frameworks applied to LEVIR-CD and DSIFN-CD datasets.

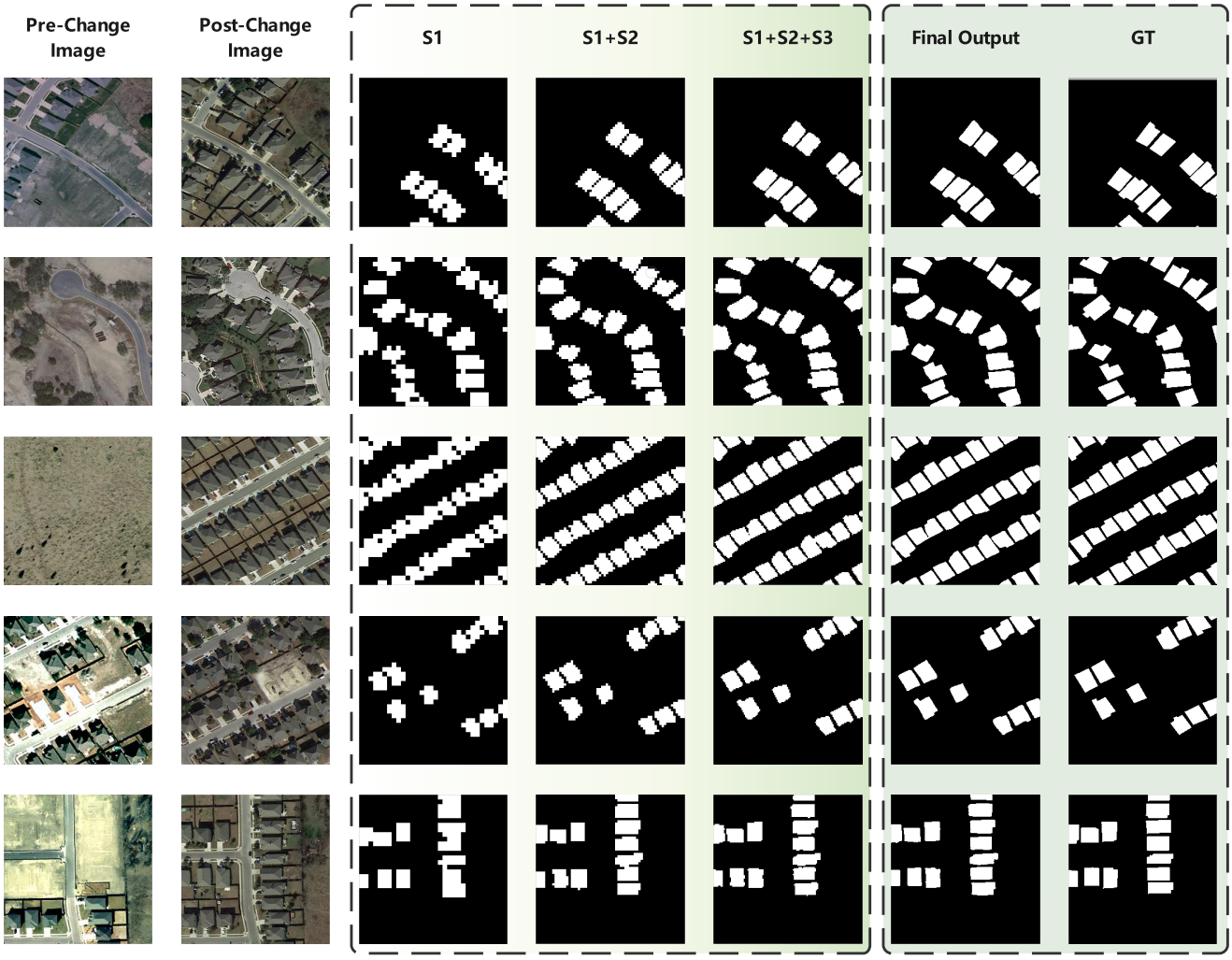

Fig. 7. Visual journey through our model's three stages in AdaptFormer.

all four metrics, with Recall dropping by 1.18%. Conversely, replacing channel exchange with spatial exchange in stage 3 revealed that simplistic exchanges at this depth adversely affected high-level semantic representation, witnessing the steepest metric drops, particularly with IoU and Recall plummeting to 83.31% and 90.53%.

Based on these outcomes, the third comparison (Group III) was conceptualized. The spatial exchange was positioned in stage 2, showing an increase in F1, IoU, and OA by 0.57%, 0.96%, and 0.02%, respectively, although Recall decreased by 0.65%. This highlighted the efficacy of the spatial exchange in enhancing CD accuracy at a mid-level semantic layer. Furthermore, deploying the channel exchange in stage 3 proved most effective, registering the best performance among the comparative groups with metrics soaring to 92.65% for F1, 86.31% for IoU, 99.19% for OA, and 92.59% for Recall. This underscored that the spatial exchange is more potent for abstract mid-level semantics in stage 2, while the channel

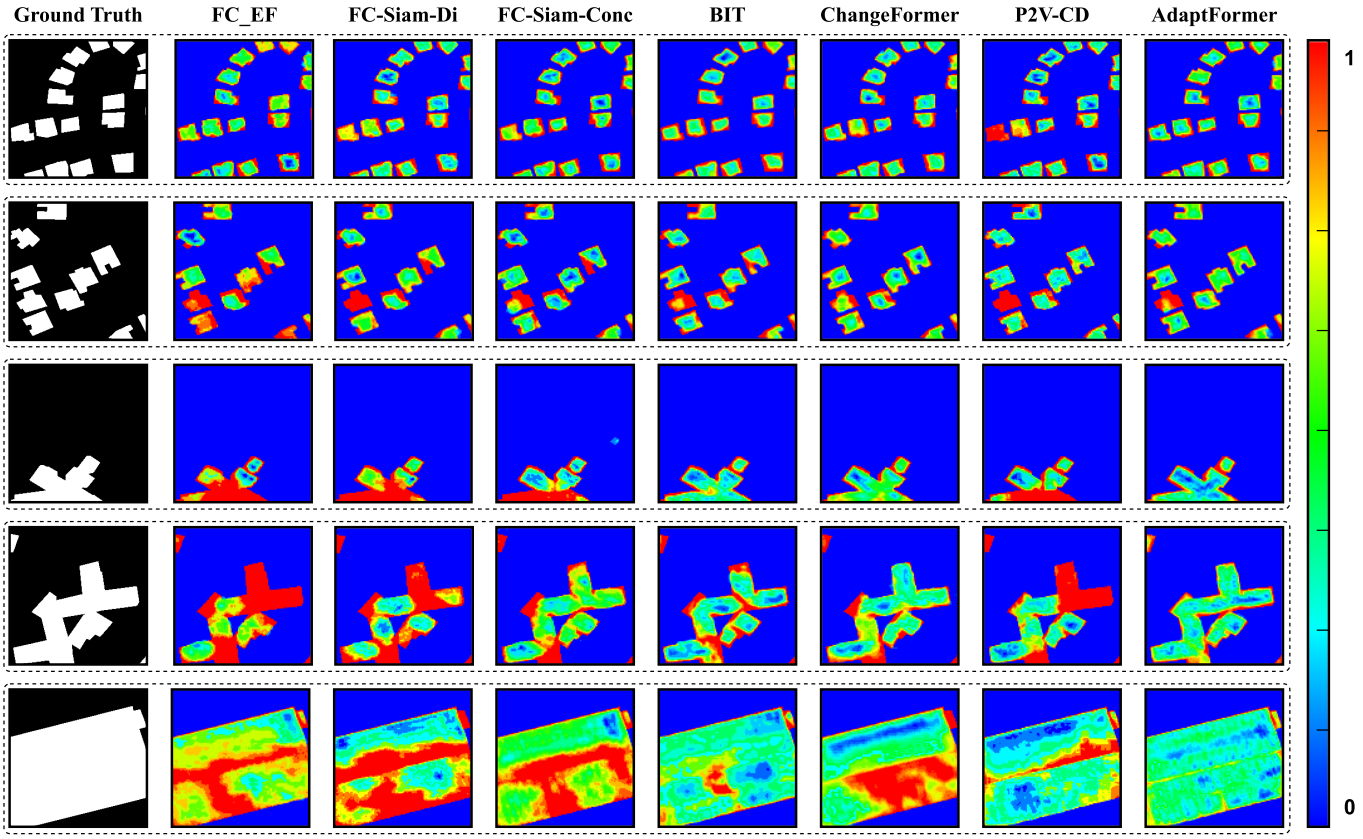

Fig. 8. Comparison of error maps resulting from different CD frameworks on the LEVIR dataset. The error maps are computed by subtracting the GT from the CD predictions. The lower the value at a position point, the more confident the model is about that point.

exchange is optimal for deep-level semantics related to objects, scenes, or advanced concepts in stage 3.

5) Evaluation of the HCA Module: The HCA module stands out for its innovative design tailored to interpret complex, deep representations. Utilizing advanced feature clipping and attention-based computations, it excels at distilling a more precise set of features that are temporally coherent and semantically rich. This feature refinement is particularly vital at stage 3, where the model is expected to make high-level semantic interpretations.

The efficacy of the HCA module is validated through a set of performance metrics. In the absence of the HCA module, the model demonstrated an F1-score of 91.28%, IoU at 83.96%, OA at 98.98%, and Recall at 91.72%. After incorporating the HCA module, each of these metrics showed significant improvement: F1 increased by 1.37%, IoU by 2.35%, OA by 0.21%, and Recall by 0.87%. These measurable gains, detailed in Table III, affirm the HCA module's pivotal role in enhancing the model's capability to make accurate and context-rich semantic judgments.

TABLE III

EFFECT OF INCORPORATING OR EXCLUDING THE HCA MODULE IN ADAPTFORMER ON THE LEVIR-CD DATASET. IN THIS TABLE, "W/O" STANDS FOR "WITHOUT" WHILE "W" INDICATES "WITH." FOR EACH PERFORMANCE METRIC, PERFORMANCE DECLINES ARE DENOTED IN **GREEN**, WHILE ENHANCEMENTS ARE HIGHLIGHTED IN **RED**. THE PERFORMANCE OF THE RECOMMENDED CHOICE IS MARKED IN GRAY

HCA	F1	IoU	OA	Recall
w/o	91.28	83.96	98.98	91.72
w	**92.65** +1.37	**86.31** +2.35	**99.19** +0.21	**92.59** +0.87

F. Visualization

1) Qualitative Performance: As illustrated in Fig. 6, a range of CD models undergo application to the LEVIR-CD and DSIFN-CD datasets, creating a broad canvas for comparison. The initial columns of the figure showcase pre-change and post-change images, offering the bedrock for evaluation. Notably, AdaptFormer, our proposed model, receives representation amidst an array of top-performing models presented in columns 3–7. The red and yellow boxes serve to highlight the areas of maximum variance in the output across the various models on the two datasets. When these results are compared with the GT, provided in the last column, AdaptFormer visibly outperforms others, demonstrating superior overall quality and accuracy, particularly within the designated regions. This juxtaposition thus emphasizes the powerful performance and substantial potential of AdaptFormer in executing CD tasks.

2) Progressive Visualization Through AdaptFormer's CD Stages: Fig. 7 offers a visual journey through our model's three stages in CD. In our analytical framework, the model traverses through a hierarchical structure of semantic analysis across three stages, each delineated by its depth of semantic processing and its implications for CD in RS imagery. Initially, stage 1 lays the groundwork by leveraging shallow semantic insights to pinpoint basic yet pivotal features like

edges and shapes, proving instrumental for the identification of minor changes. However, this stage is limited in its ability to unravel more intricate details. Advancing to stage 2, the model deepens its semantic exploration to intermediate levels, thereby refining its detection capabilities to encompass moderate changes through the discernment of more complex shapes and patterns, albeit with remaining challenges in capturing the finest nuances. The culmination occurs in stage 3, where an intensive dive into deep semantic realms enables the model to grasp comprehensive contextual relationships and substantial structural shifts, thus extending its detection acumen to substantial changes. This graduated approach aligns closely with GT data, indicating minimal discrepancies and highlighting the model's adaptability and scalability. The framework effectively addresses the diverse requirements of CD in RS imagery, accommodating changes across a wide range of magnitudes.

3) Error Maps: We employ error maps as a visual technique to rigorously assess the effectiveness of CD on RS images, highlighting discrepancies between predicted and true values. Fig. 8 elucidates the confidence visualization results for various CD models when applied to the LEVIR-CD dataset. Primarily, the majority of the figures—columns 1 to 6—display error analysis from several mainstream models on their respective test images, whereas the concluding column distinctively represents the outcomes of our AdaptFormer approach. A unique measurement system was employed wherein the differences between the model outputs and the GT were visualized on a scale from 0 to 1. A shade closer to blue (indicating a value nearer to 0) epitomizes high confidence in detection, while a hue leaning toward red (signifying a value approaching 1) designates lesser assurance.

In this visualization, AdaptFormer's adeptness is consistently evident across various test images. Particularly notable is its proficiency in small object detection, where the near absence of the red hue in the first row suggests its enhanced capability to identify scattered minor entities. For medium-sized objects, many contemporary models manifest continuous red zones, indicating lapses in their detection confidence. In stark contrast, AdaptFormer's results, especially in the fourth row, underscore its superiority by almost flawlessly identifying these areas. This prowess extends to large object detection as well, as observed in the fifth row, where the dearth of red regions in our method's visualization stands testament to its exceptional confidence and accuracy in recognizing substantial object changes.

V. Conclusion

This study presents AdaptFormer, a groundbreaking solution to CD in RS imagery. Distinctly adaptive, AdaptFormer systematically interprets hierarchical semantics, tailoring its operations across three depth levels: simple techniques for shallow semantics, spatial data assimilation for medium details, and cascaded depthwise attention for in-depth insights. Our experimental evaluations, particularly on the LEVIR-CD and DSIFN-CD datasets, showcase AdaptFormer's superior accuracy and performance over other models, underscore its potential in applications from urban development to environmental surveillance. In essence, AdaptFormer emerges as a benchmark in CD, ushering in new avenues for future research and development in the domain. In future work, we aim to enhance the computational efficiency of the AdaptFormer model to better support real-time analysis, while maintaining its accuracy and effectiveness in CD tasks.

References

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.

[2] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.

[3] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.

[4] H. Chen, W. Li, S. Chen, and Z. Shi, "Semantic-aware dense representation learning for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630018.

[5] Y. Liang, C. Zhang, and M. Han, "RaSRNet: An end-to-end relation-aware semantic reasoning network for change detection in optical remote sensing images," *IEEE Trans. Instrum. Meas.*

[6] 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.

[7] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," 2019, *arXiv:1910.06444*.

[8] P. de Bem, O. de Carvalho Junior, R. F. Guimar aes, and R. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.

[9] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[10] Y. Pang et al., "Improved crop row detection with deep neural network for early-season maize stand count in UAV imagery," *Comput. Electron. Agricult.*, vol. 178, Nov. 2020, Art. no. 105766.

[11] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[12] 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.

[13] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

[14] T. Lei et al., "Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[15] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 207–210.

[16] Y. Pang et al., "Slim UNETR: Scale hybrid transformers to efficient 3D medical image segmentation under limited computational resources," *IEEE Trans. Med. Imag.*, vol. 43, no. 3, pp. 994–1005, Mar. 2024.

[17] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[18] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[19] Z. Chen et al., "A new approach for detecting urban centers and their spatial structure with nighttime light remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6305–6319, Nov. 2017.

[20] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.

[21] H. Yao, R. Qin, and X. Chen, "Unmanned aerial vehicle for remote sensing applications—A review," *Remote Sens.*, vol. 11, no. 12, p. 1443, Jun. 2019.
[22] H. Yin et al., "Attention-guided Siamese networks for change detection in high resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, Mar. 2023, Art. no. 103206.
[23] Z. Wang et al., "Toward learning joint inference tasks for IASS-MTS using dual attention memory with stochastic generative imputation," *IEEE Trans. Neural Netw. Learn. Syst*.
[24] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," in *Geospatial Information Handbook for Water Resources and Watershed Management*, vol. 2. Boca Raton, FL, USA: CRC Press, 2022, pp. 65–88.
[25] S. I. Toure, D. A. Stow, H.-C. Shih, J. Weeks, and D. Lopez-Carr, "Land cover and land use change analysis using multi-spatial resolution data and object-based image analysis," *Remote Sens. Environ.*, vol. 210, pp. 259–268, Jun. 2018.
[26] S. Schlaffer, P. Matgen, M. Hollaus, and W. Wagner, "Flood detection from multi-temporal SAR data using harmonic analysis and change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 38, pp. 15–24, Jun. 2015.
[27] T. Peterson, C. Folland, G. Gruza, W. Hogg, A. Mokssit, and N. Plummer, *Report on the Activities of the Working Group on Climate Change Detection and Related Rapporteurs*. World Meteorological Organization Geneva, 2001.
[28] K. Rokni, A. Ahmad, A. Selamat, and S. Hazini, "Water feature extraction and change detection using multitemporal Landsat imagery," *Remote Sens.*, vol. 6, no. 5, pp. 4173–4189, 2014.
[29] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.
[30] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.
[31] 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, May 2021.
[32] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020.
[33] 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.
[34] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.
[35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[36] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.
[37] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[38] C. Zhang et al., "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, Aug. 2020.
[39] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.
[40] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proc. 5th Conf. Message Understand.*, 1993, pp. 1–10.
[41] X. Song, Z. Hua, and J. Li, "Remote sensing image change detection transformer network based on dual-feature mixed attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.
[42] J. Huang, Q. Shen, M. Wang, and M. Yang, "Multiple attention Siamese network for high-resolution image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021.
[43] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. L. Rojo-Alvarez, and M. Martinez-Ramon, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.
[44] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, "RDP-net: Region detail preserving network for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635010.
[45] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[46] S. H. Khan, X. He, F. Porikli, and M. Bennamoun, "Forest change detection in incomplete satellite images with deep neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5407–5423, Sep. 2017.
[47] 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.
[48] 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.

Teng Huang received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2019.

Currently, he is working at the Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou, China. His research interests include computer vision, RS, and blockchain.

Yile Hong received the B.S. degree from Huizhou University, Huizhou, China, in 2022. He is currently pursuing the master's degree with Guangzhou University, Guangzhou, China.

His research areas of interest span computer vision, remote sensing and smart contract.

Yan Pang (Member, IEEE) received the Ph.D. degree from the University of Colorado, Denver, CO, USA, in 2021.

From April 2021 to May 2022, he was a Machine Learning Scientist at Moffett AI, Los Altos, CA, USA. From August 2018 to May 2021, he was an Instructor with the Department of Electrical Engineering, University of Colorado Denver, and the Department of Electrical Engineering Technology, Metropolitan State University of Denver, Denver. His research interests span machine learning, computer vision, and efficient deep learning.

Jiaming Liang received the B.S. degree from Foshan University, Foshan, China, in 2021. He is currently pursuing the master's degree with Guangzhou University, Guangzhou, China.

His research areas of interest span computer vision, remote sensing, and medical image analysis. As a master's student, he focuses on advancing the field through innovative research and applications, leveraging his background in electronic technology and a keen interest in the computer vision domain.

Jie Hong received the B.S. degree from Nanjing Institute of Technology, Nanjing, China, in 2021. He is currently pursuing the master's degree with Guangzhou University, Guangzhou, China.

His research areas of interest span computer vision, remote sensing, and medical image analysis.

Lin Huang is a Professor with the Metropolitan State University of Denver, Denver, CO, USA. Her research interests include the areas of biometrics, pattern recognition, signal processing, computer vision, machine learning, embedded system design, and VLSI.

Dr. Huang has been an Active Member of International Conference on Machine Learning and Computing (ICMLC) as a Conference Chair/Keynote Speaker/Session Chair, since 2011. She is a Program Chair of the International Conference on Signal Processing (ICOSP 2023 & 2024). She has been an Editor Board Member of several international journals, for example, she has been the Editor-in-Chief of the *International Journal of Machine Learning and Computing* (IJMLC), since 2012. And she has been reviewing papers on regular basis for some conferences and journals since 2008.

Yuan Zhang received the B.S., M.S., and Ph.D. degrees from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, 2008, and 2017, respectively.

He was a joint Ph.D. Student with German Aerospace Center, Cologne, Germany, and the Technical University of Munich, Munich, Germany, from 2015 to 2016. He is an Associate Professor with the North China University of Technology, Beijing, China. His main research direction is synthetic aperture radar (SAR) imaging.

Yan Jia received the double M.S. degree in telecommunications engineering and computer application technology from Politecnico di Torino, Turin, Italy, and Henan Polytechnic University, Jiaozuo, China, in 2013, and the Ph.D. degree in electronics engineering from Politecnico di Torino in 2017.

Now she is working with Nanjing University of Posts and Telecommunications. In 2013, she was with the Department of Electronics and Telecommunications, Politecnico di Torino, where she performed research on GNSS system construction and GNSS antenna analysis. In 2014, she worked with the SMAT project, mainly focusing on the retrieval of soil moisture and vegetation biomass content by GNSS-R. Her research interests include microwave remote sensing, soil moisture retrieval, and Global Navigation Satellite System Reflectometry (GNSS-R) applications to land remote sensing and antenna design.

Patrizia Savi (Senior Member, IEEE) received the Laurea degree in electronic engineering from the Politecnico di Torino, Turin, Italy, in 1985.

In 1986, she was a Consultant with Alenia, Caselle Torinese, Italy, where she conducted research on the analysis and design of dielectric radomes. From 1987 to 1998, she was a Researcher with Italian National Research Council, Rome, Italy.