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

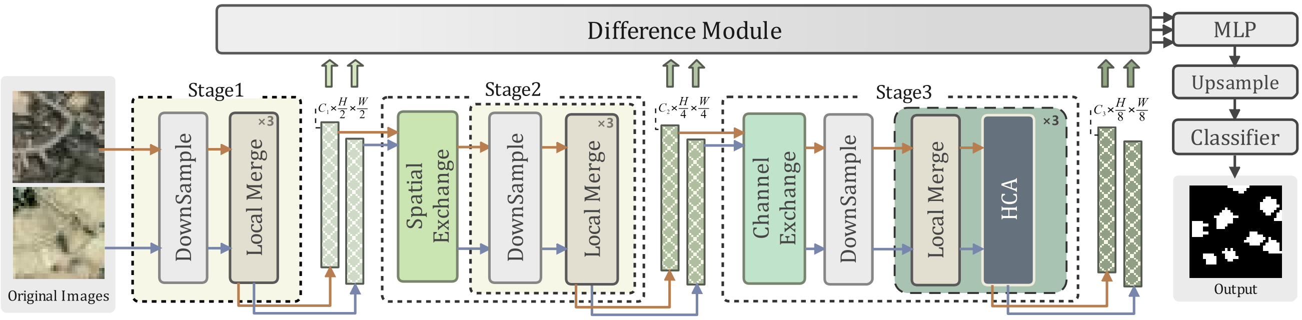

Fig. 1. Schematic representation of the AdaptFormer architecture. The proposed AdaptFormer employs distinct strategies from straightforward operations for shallow levels, spatial data assimilation for medium levels, to cascaded depthwise attention for deeper semantics.

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 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.

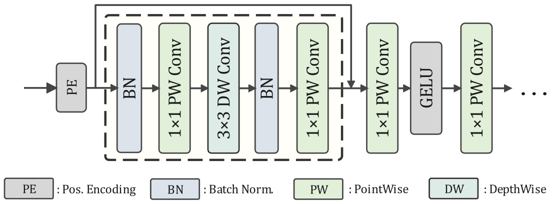

Fig. 2. Structure of local merge.

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.

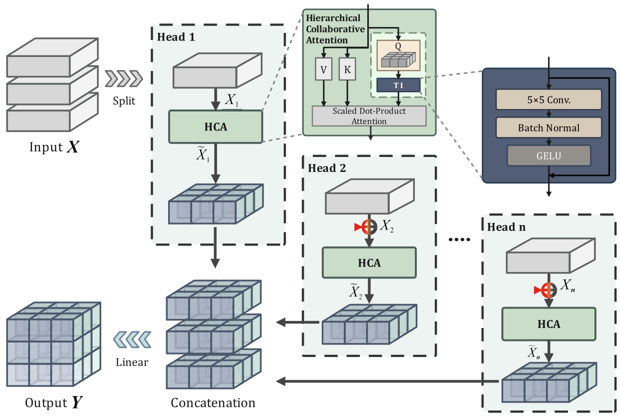

Fig. 3. Overview of HCA.

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:

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

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 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.

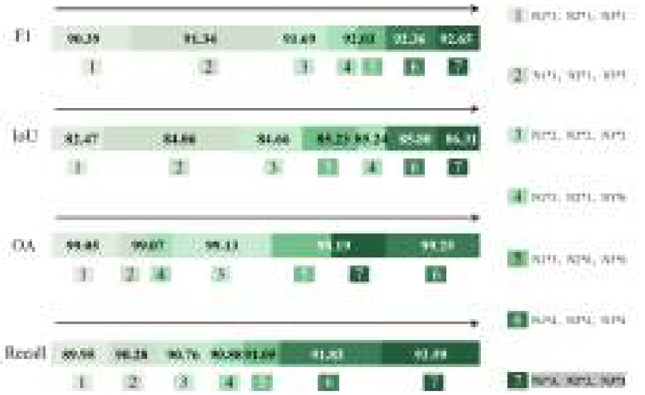

Fig. 4. Quantitative comparison with different stage depths of AdaptFormer on the LEVIR-CD dataset.

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%.