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

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

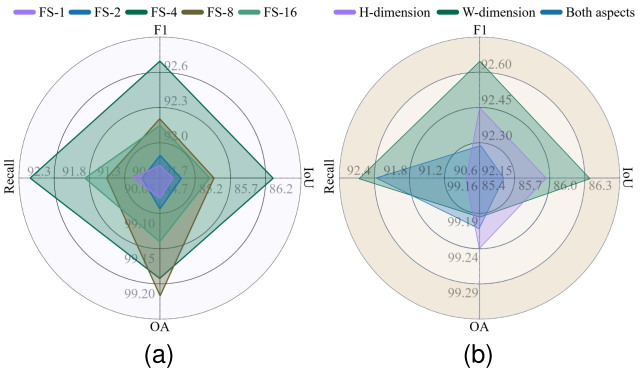

Fig. 5. Conducting a comparative quantitative analysis involves examining (a) various feature splits within the HCA module and (b) diverse dimensional configurations in the spatial exchange module. Both aspects are evaluated using the LEVIR-CD dataset. FS: feature split.

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

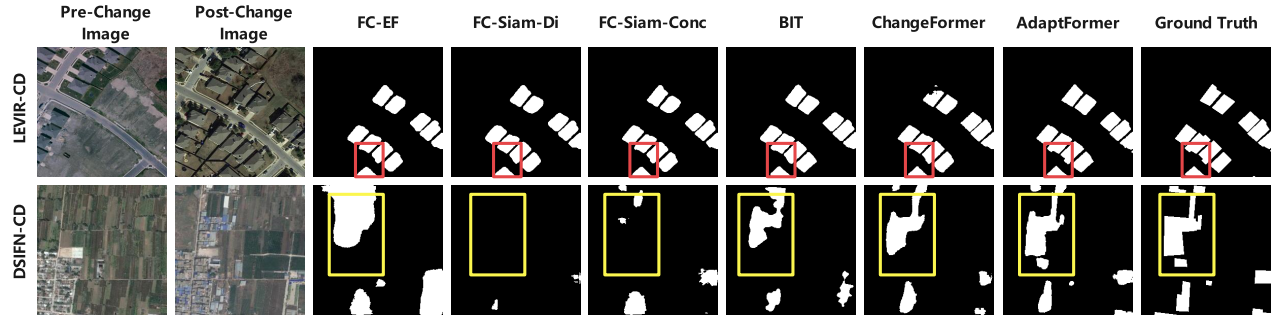

Fig. 6. Comparative display of CD performance from divergent CD frameworks applied to LEVIR-CD and DSIFN-CD datasets.

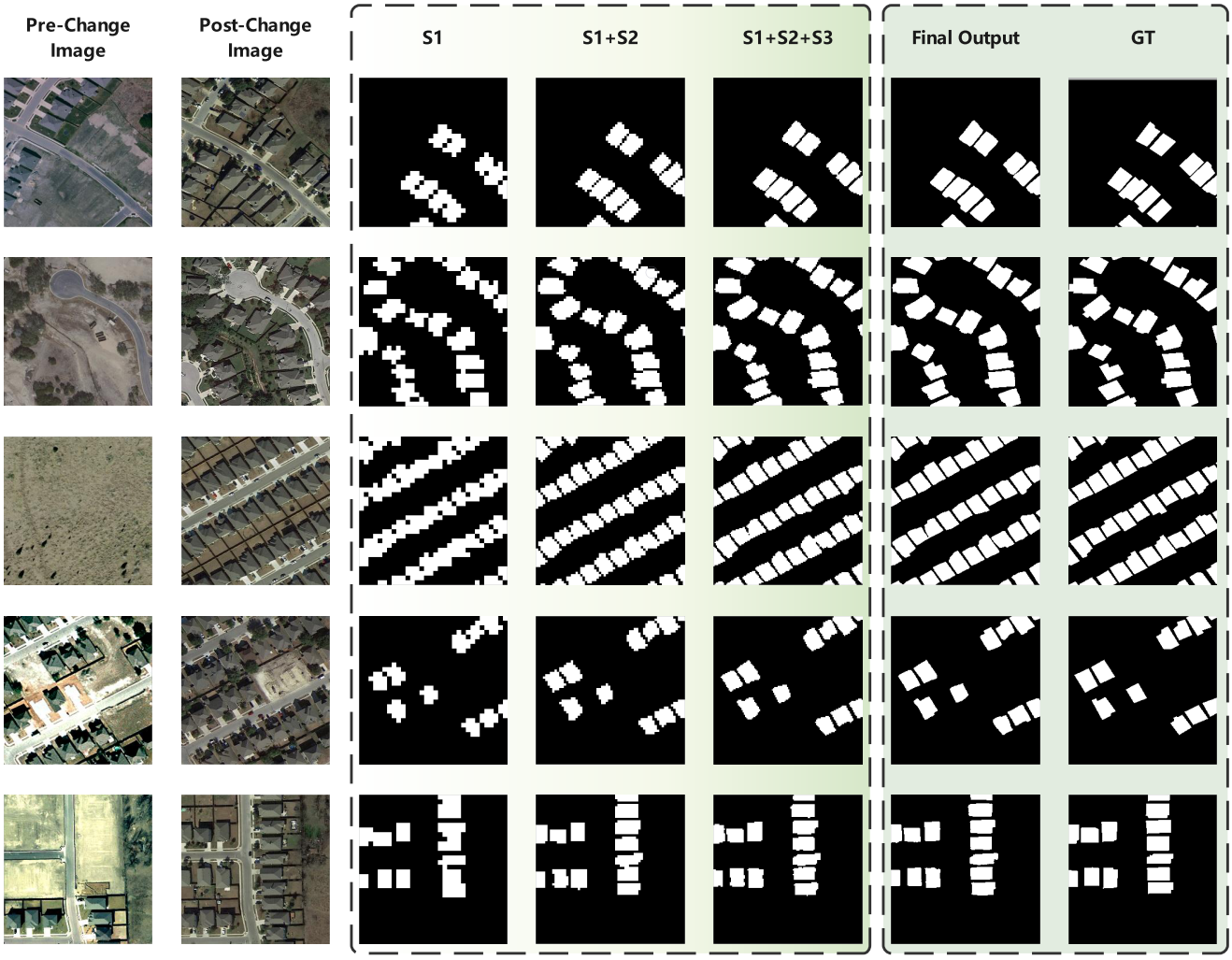

Fig. 7. Visual journey through our model's three stages in AdaptFormer.

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

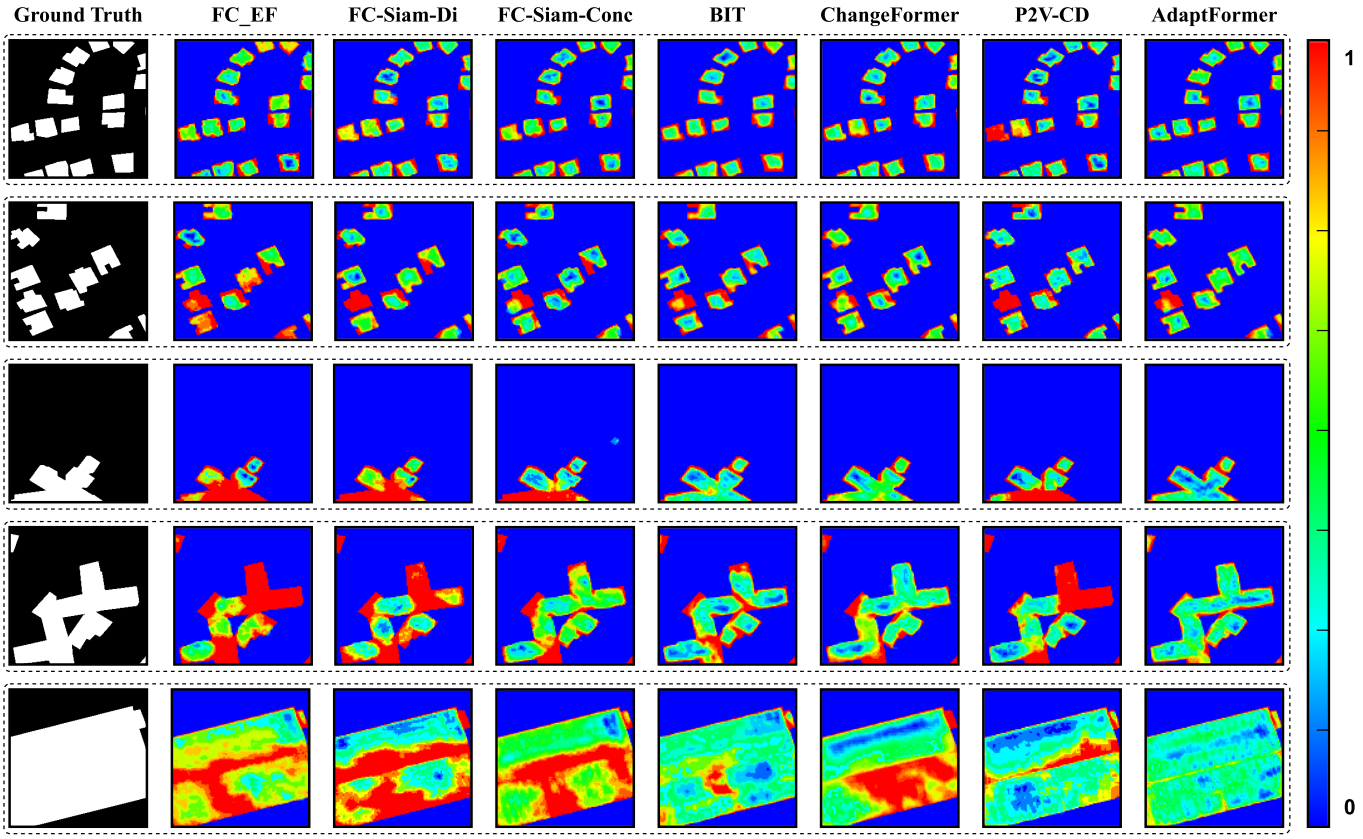

Fig. 8. Comparison of error maps resulting from different CD frameworks on the LEVIR dataset. The error maps are computed by subtracting the GT from the CD predictions. The lower the value at a position point, the more confident the model is about that point.

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