

Supplementary Materials: Multi-scale Change-Aware Transformer for Remote Sensing Image Change Detection

Anonymous Authors

1 ANNOTATION OF MACD DATASET

Our remote sensing change detection dataset undergoes a meticulous annotation process, executed by a team of seasoned experts with in-depth knowledge of remote sensing interpretation and the complexities of mining area topography. The multi-tiered annotation strategy ensures dataset accuracy and reliability for change detection analysis:

Coarse Localization of Mining Areas: In the initial phase, our annotation process commences with an analysis of large-scale imagery using geographical coordinates to identify the approximate boundaries of mining areas. To ensure the completeness of the region under study, we establish a buffer zone extending 200 to 300 pixels from the epicenter of the mining area, thereby creating a preliminary map for coarse localization.

Image Registration and Cropping: Following the coarse localization phase, we meticulously register images of the same mining area captured at different times by various sensors. This registration process is essential as it ensures the comparability of image data acquired at different time points. Once registration is complete, we standardize the images by resizing them to a uniform dimension of 128×128 pixels. This standardization facilitates a consistent and in-depth annotation process.

Fine Annotation: In the critical phase of detailed data annotation, our team of experts employed ArcGIS 10.2 software to conduct comprehensive pixel-level annotation of mining areas across two distinct time frames. This meticulous process involved the precise delineation of boundaries and the accurate identification of internal features within each sample, thereby ensuring the integrity and quality of the annotated data for our dataset. As a result of this precise work, we have generated individual mask maps for the dual temporal imagery.

Generation of Change Maps: By subtracting the mask maps derived from dual temporal imagery, we generated change maps that accurately pinpoint locations where alterations have occurred within the mining areas. This process is complemented by a robust annotation protocol and stringent quality assurance measures, including cross-validation among annotators and expert panel reviews. These measures collectively ensure the exceptional accuracy and robustness of our dataset, making it highly suitable for refining and evaluating sophisticated change detection models.

2 EXPERIMENTS

In this section, we begin by presenting an overview of the datasets involved, along with the pertinent experimental parameters, comparative methods, and evaluation metrics. Subsequently, we augment the discussion with additional visual examples from each dataset, providing a more comprehensive understanding of the data characteristics. Furthermore, we conduct a thorough statistical analysis of the performance results for all methods across each dataset.

Experimental Datasets. Our experiments are conducted on three datasets: MACD, LEVIR, and WHU, each providing unique challenges for change detection algorithms.

MACD Dataset: Our MACD dataset consists of 2133 image pairs at a resolution of 128×128 pixels. The dataset is split into 1801 pairs for training and 332 for testing. Spanning from 2018 to August 2023, data collection has been conducted multiple times, capturing a diverse range of solar elevation angles, seasons, and weather conditions. A key feature of MACD is the diverse scale and complexity of change areas, characterized by intricate shapes and curved edges, which distinguishes it from datasets focused on urban structures.

LEVIR Dataset: The LEVIR dataset [3] is an optical dataset widely used for remote sensing change detection. It includes 637 high-resolution images of 1024×1024 pixels, covering various regions in Texas with significant landscape and structural changes. The dataset’s dual-temporal images record substantial alterations, and the variations in acquisition times, seasons, and lighting conditions present challenges for neural network performance. We use a default sub-image size of 256×256 pixels, and the data is randomly partitioned into training, validation, and test sets following a ratio of 7 : 1 : 2, yielding set sizes of 7120, 1024, and 2048, respectively.

WHU Dataset: The WHU dataset [8] contains detailed building information with a resolution of 0.075m/pixel and an original size of 32507×15354 pixels. We cropped the images non-overlappingly to 256×256 pixels. Given the relatively small size of the WHU dataset, we partitioned it into training, validation, and test sets with a ratio of 8 : 1 : 1, resulting in set sizes of 6096, 762, and 762, respectively.

In summary, the MACD dataset introduces a high level of complexity with its diverse and intricately shaped change areas, reflecting the dynamic nature of mining regions and presenting a significant challenge for change detection algorithms. The LEVIR dataset, with its high-resolution imagery capturing various urban landscapes, emphasizes the need for algorithms to handle significant alterations and environmental variations. Lastly, the WHU dataset, with its fine-grained building information, tests the algorithms’ ability to detect changes in dense urban structures. Together, these datasets provide a comprehensive evaluation platform, enabling the assessment of algorithmic performance across a range of scenarios and complexities, and thus driving the advancement of change detection technologies.

Implemented Details. We stack (3, 6, 6, 3) Dynamic Change-Aware Attention modules in each of the four stages of our One-stream Change Detection Transformer. The local window size L_i of Dynamic Change-Aware Attention for each stage is set to (3, 3, 5, 5). The C_i is set to (64, 128, 250, 320).

Pseudo-Code The inference detail of our Multi-scale Change-Aware Transformer (MACT) for change detection is shown in Algorithm 1.

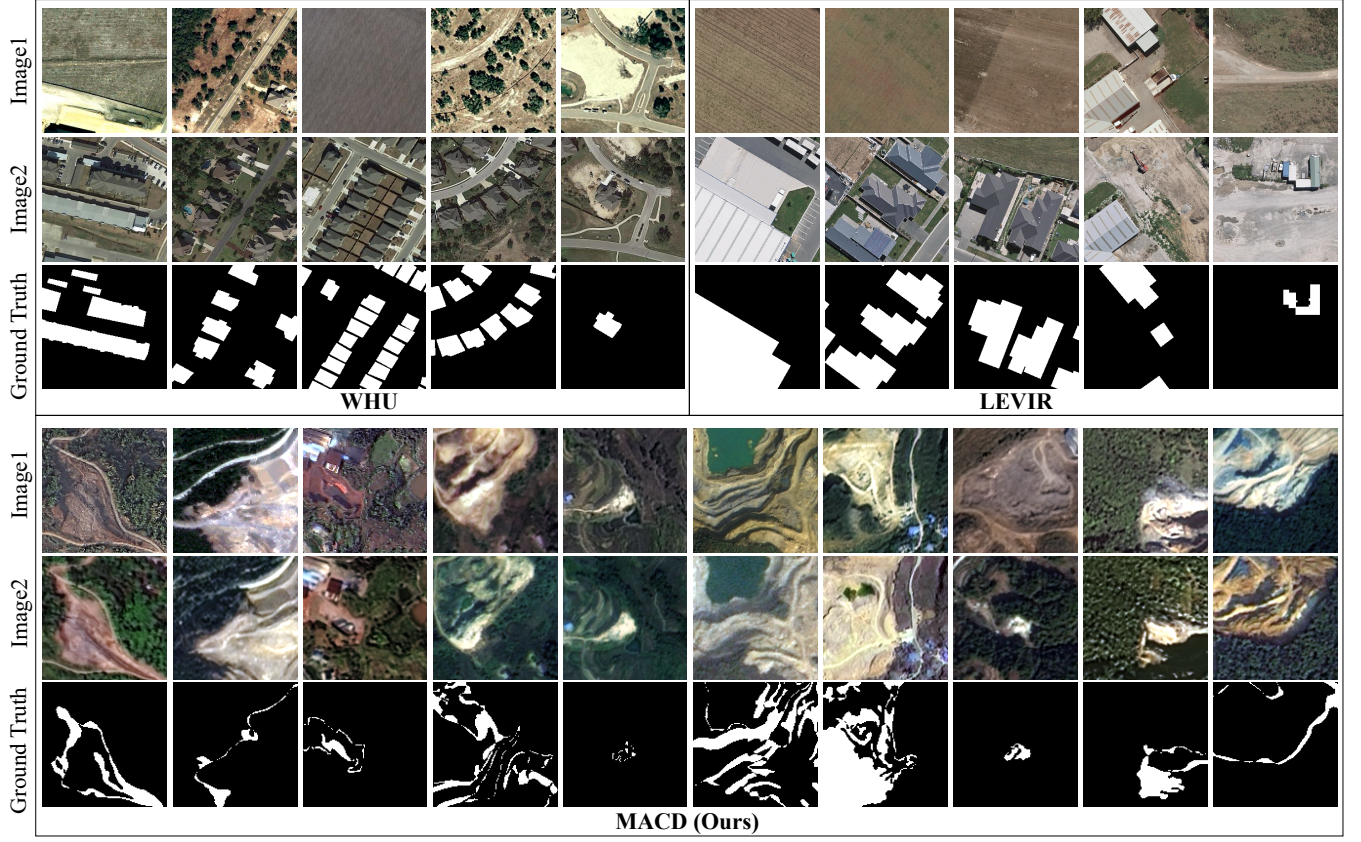


Figure 1: Comparison of our dataset with WHU and LEVIR datasets reveals that the change regions in WHU and LEVIR are predominantly regular in shape, whereas our MACD dataset exhibits a greater diversity in scale and complexity of changes.

Algorithm 1: MCAT for change detection

Input: Two temporal remote sensing images (I_1, I_2)

Output: Chaneg map P

- 1: **Step1:** Chang-aware Encoder
 - 2: **for** stage $i \in \{1, 2, 3, 4\}$ **do**
 - 3: $Y_i = \text{Dynamic Change-aware Attention}(Y_{i-1})$;
 - 4: **end**
 - 5: **Step2:** Change-enhanced Multiscale Aggregator
 - 6: **for** layer $i \in \{1, 2, 3, 4\}$ **do**
 - 7: $M'_i = \text{Change-enhanced Module}(Y_i)$;
 - 8: $M_i = \text{Multi-scale Change Aggregator}(M_{i-1} + M'_i)$;
 - 9: **end**
 - 10: **Step3:** Change Map Generation
 - 11: $P = \text{Decoder}(M_i)_{i \in \{1, 2, 3, 4\}}$;
 - 12: **Return:** P
-

Comparison Methods. We have conducted a comprehensive comparison of our model against several state-of-the-art approaches in remote sensing change detection, encompassing FC-EF [5], FC-Siam-Diff [10], FC-Siam-Conc [5], IFNet [12], DASNet [4], DTCD-SCN [11], SNUNet [6], BIT [2], SARAS [1], ChangeFormer [7], and USSFC-Net [9]. Next, we briefly describe the main ideas of these ten methods.

FC-EF [5]: the network architecture of this method is based on UNet, using an early additive fusion strategy. Raw diachronic images are concatenated as network inputs and processed through a one-stream convolutional network to detect changes.

FC-Siam-Diff [10]: this method employs a post-fusion strategy based on FC-EF networks. Multi-scale features are extracted from a dual convolutional network of diachronic images and algebraic operations are used to obtain parallax features to detect changes.

FC-Siam-Conc [5]: the method uses the U-Net architecture. Bitemporal features transmitted by a shallow network are fused through connections and integrated with deep semantic information.

IFNet [12]: the method employs channel attention and spatial attention to enhance the extraction of disparity features, respectively. A deep supervision mechanism is used to supervise the disparity feature extraction process.

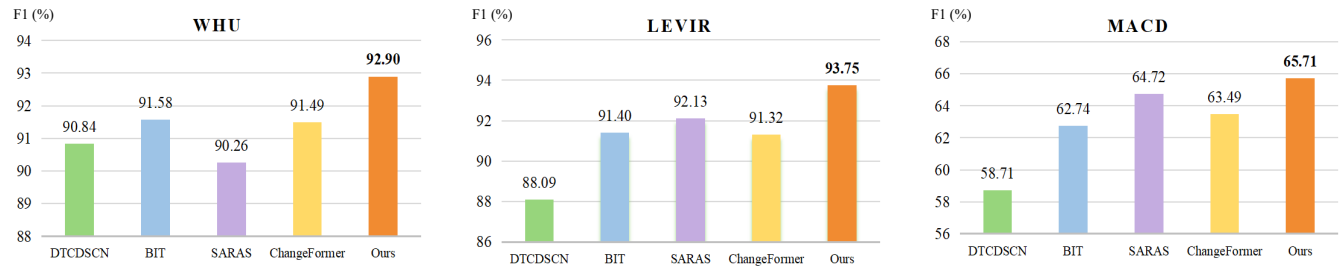


Figure 2: F1 Score Comparison Across Three Datasets. The bar graph illustrates the F1 values obtained by our method on the LEVIR, MACD, and an additional third dataset. Notably, our approach outperforms others on all datasets, with particularly consistent results observed on both the LEVIR and MACD datasets, indicating a robust performance in change detection tasks.

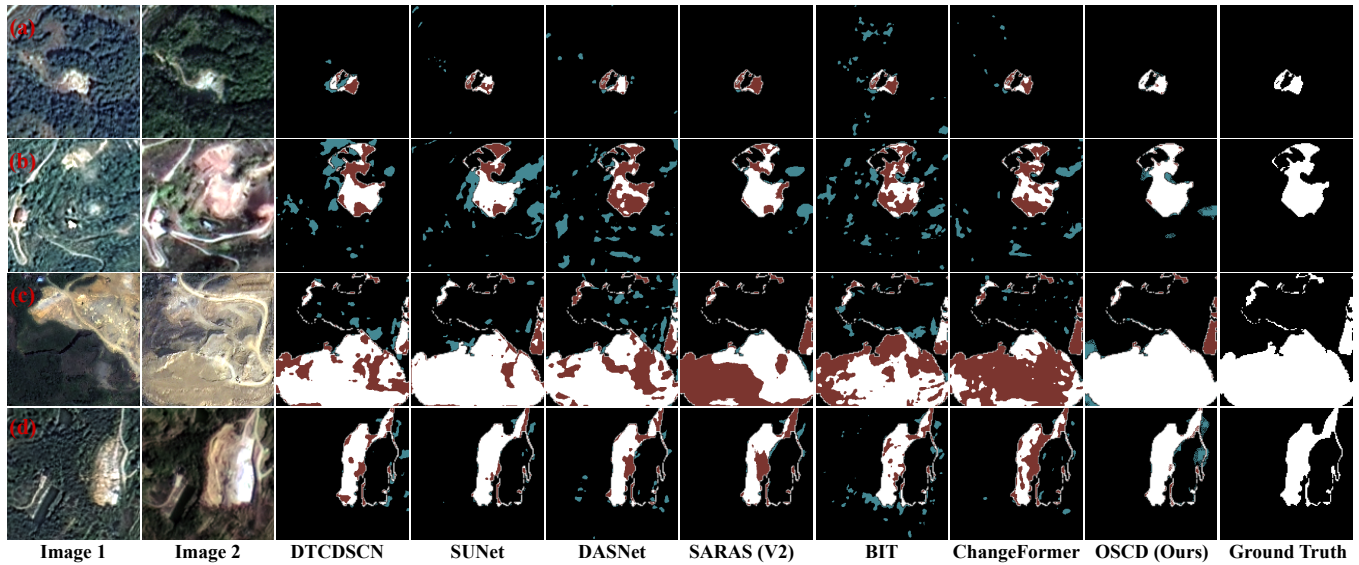


Figure 3: Visual results on the MACD dataset are depicted. (a)–(d) four representative samples. White represents true positives, black denotes true negatives, red indicates false negatives, and green represents false positives.

DASNet [4]: DASNet is a fully convolutional Siamese neural network based on dual attention, extracting spatial and channel information through two separate attention mechanisms. It addresses the issue of sample imbalance by incorporating a weighted bilateral contrastive loss.

DTCDSCN [11]: this method creates a dual attention module with channel attention and spatial attention to enhance feature information based on SE-ResNet.

SNUNet [6]: SNUNet uses the UNet++ architecture to compensate for details and semantic information through dense hopping connections, with the core being an integrated channel attention module that handles multi-scale semantic information.

BIT [2]: this method uses ResNet as the backbone. The use of two decoder transformers is used to establish long-term dependencies on deep semantic information. Detailed semantic information is emphasized by subtracting and taking the absolute value of the difference features.

SARAS [1]: the network is a multiscale architecture that combines ResNet and Transformer, addressing boundary noise from objects of different scales through a scale-aware module and a relation-aware module. Additionally, the method utilizes a cross-transform module to fuse features from different scales, enhancing the representation for improved change detection

ChangeFormer [7]: this method uses the transformer architecture as a backbone network. It creates a difference module to continuously aggregate feature information for the difference features in each layer.

USSFC-Net [9]: USSFC-Net is an ultra-lightweight remote sensing change detection network. It uses a pseudo-siamese U-Net as the backbone to flexibly capture multi-scale features of change objects through multi-scale decoupled convolution. In addition, spatial spectral features are cooperated with the strategy to better capture the change-related features.

To ensure a fair and rigorous comparison, we have reimplemented all methods, optimizing each by selecting the parameter

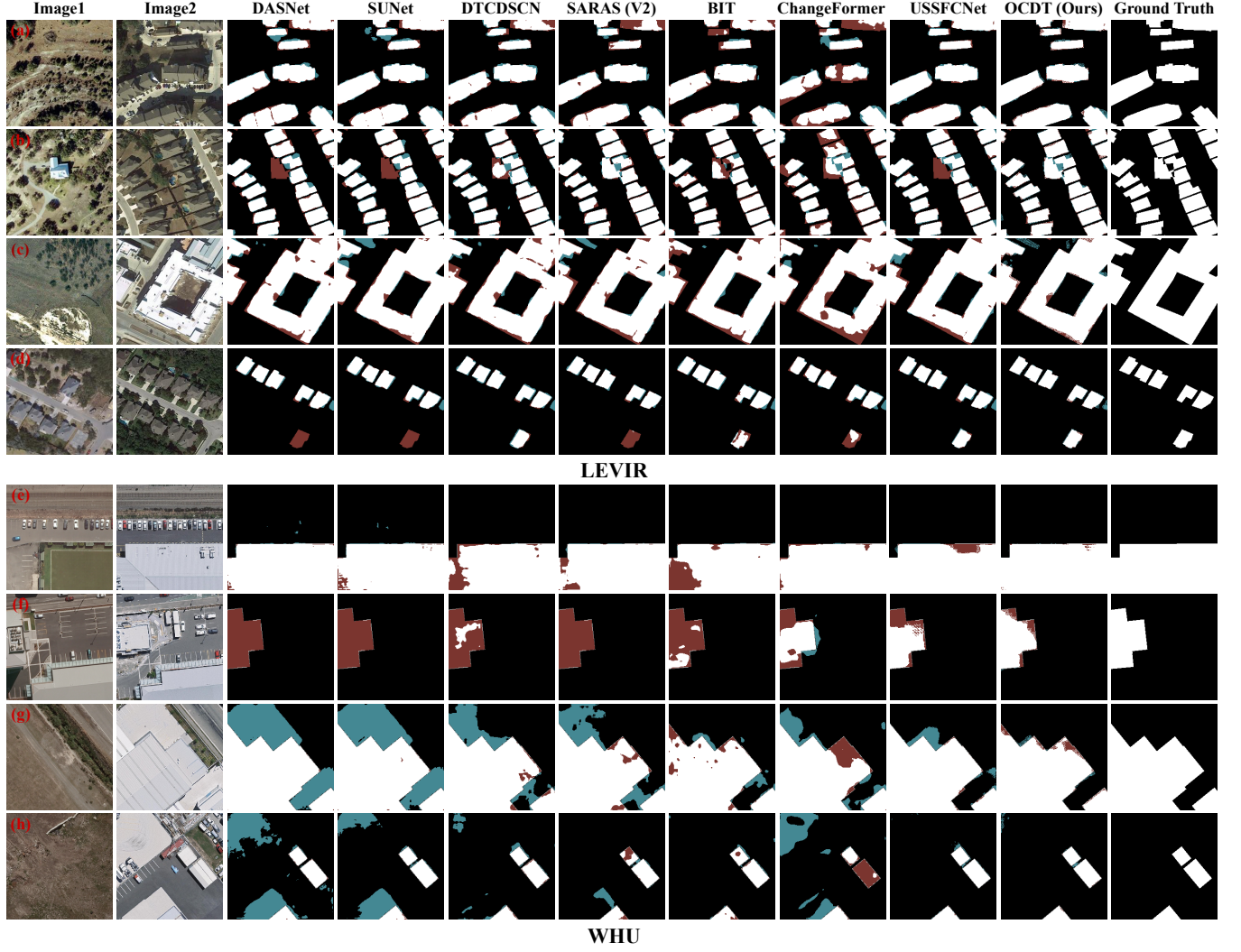


Figure 4: Qualitative results of different change detection methods on LEVIR and WHU dataset. (a)–(h) eight representative samples.

set that yielded the highest F1 scores on the validation set during training.

Evaluation Metrics. The effectiveness of our model is quantified using the F1 score and Intersection over Union (IoU) as the primary metrics. We also report additional metrics including precision (Prec.), recall (Rec.), and overall accuracy (OA). These metrics are computed using the formulas below:

$$\text{precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{OA} = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (4)$$

$$\text{F1} = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \quad (5)$$

Here, TP (True Positives) is the count of correctly identified changed pixels. FP (False Positives) is the number of pixels incorrectly labeled as changed. TN (True Negatives) represents the unchanged pixels correctly identified as such. FN (False Negatives) is the count of changed pixels incorrectly classified as unchanged.

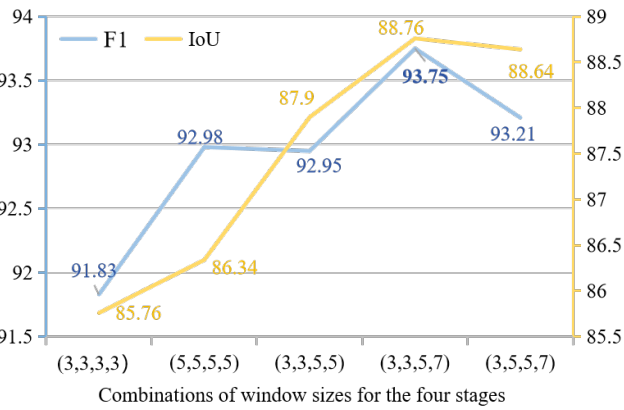


Figure 5: The selection of the window size on different stages.

The F1 score and IoU are scaled between 0 and 1, where values nearing 1 signify higher model accuracy.

Numerical Results Statistics. Fig. 2 presents a comparative analysis of four selected methods, showcasing their performance across the WHU, LEVIR, and our MACD datasets. Our proposed method demonstrates outstanding performance on three remote sensing image change detection datasets: WHU, LEVIR, and MACD. Notably, our method leads across the board in terms of the F1 score, a critical metric for assessing the reliability and effectiveness of change detection tasks. On the WHU and LEVIR datasets, where change areas are regular and the scale variation is minimal, most change detection algorithms can achieve satisfactory results. However, our method not only achieves high accuracy of 92.90% and 93.75%, but also exhibits greater robustness. This advantage stems from the innovative feature extraction and fusion strategies employed by our model, which can precisely capture change areas even when the changes are subtle. When it comes to the MACD dataset, which features large-scale and complex-shaped change areas, the requirements for an algorithm’s multi-scale and shape adaptability are heightened. Current methods often struggle with this dataset, as evidenced by the lower F1 scores, reflecting their limitations in handling complex changes. The SARAS method performs well on MACD because it is specifically designed to address multi-scale change areas. Nonetheless, our method still takes the lead with an F1 score of 65.71%, a result that underscores the significant strength of our model in dealing with complex change areas.

More Qualitative Results. Due to space limitations in the main text, we present visual results for only a subset of methods. To comprehensively showcase the visual superiority of our method, we supplement additional visual predictions on all three datasets.

The samples depicted in Fig. 3 illustrate the robust performance of our method across a spectrum of change detection scenarios, including complex, large-area, and small-region changes. Our approach accurately captures the nuances of genuine semantic changes, providing a detailed representation of the transformations within the imagery. For instance, Fig. 3(a) and (c) highlight our method’s capability to detect scale variations effectively. These figures showcase two distinct instances where other methods falter, particularly in identifying fine details within large change areas. The limitations

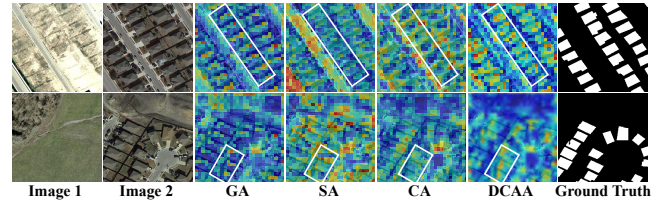


Figure 6: Comparison of attention maps between Global Attention (GA), Self-Attention (SA), Cross-Attention (CA) and Dynamic Change-Aware Attention (DCAA).

of these alternative methods result in a higher number of false negatives, as indicated by the red color in the comparative analysis. Our method, however, excels in these challenging scenarios, delivering a more comprehensive and precise detection of changes.

In Fig. 4(a), (b), and (c), our Multi-scale Change-Aware Transformer is showcased alongside SARAS, where both accurately detect changes within dense building areas while effectively mitigating noise interference. A comparative analysis reveals that our method provides a more comprehensive detection of building edges, as exemplified in Fig. 4(c), (d), (e), and (f). This enhancement is credited to MCAT’s sophisticated feature processing, which contrasts with the BIT model’s direct upsampling of low-level features to the original resolution, a technique that often results in a less detailed edge representation. Unlike BIT, our approach employs a hierarchical fusion of change-enhanced multiscale features, which are then relayed to the decoder, ensuring the preservation of fine details.

When evaluating the performance of DASNet, SUNet, and DTCD-SCN, as depicted in Fig. 4(g) and Fig. 3(c), these methods exhibit a more pronounced incidence of missed detections, particularly in larger-scale changes, which are marked in red. This shortcoming is attributed to their inability to integrate global and complementary local information, leading to inaccuracies in change detection. In stark contrast, Multi-scale Change-Aware Transformer maintains the internal compactness of change regions and the integrity of boundary detection through the strategic use of varying window sizes across different stages and the fusion of multiscale features.

Furthermore, our method demonstrates exceptional proficiency in suppressing false positives. As illustrated in Fig. 4(g), other methods are often misled by non-building areas with similar color profiles, resulting in misjudgments. Multi-scale Change-Aware Transformer, however, directly engages with change regions during the feature extraction phase and subsequently enhances them locally. This targeted interaction enables Multi-scale Change-Aware Transformer to effectively filter out irrelevant change regions, leading to more precise detection outcomes.

3 ABLATION STUDY

We add more ablation studies on different components. on the LEVIR dataset and also present the results on our MACD dataset. **Analysis of Dynamic Change-Aware Attention.** Fig. 6 offers a detailed analysis of the benefits of our proposed Dynamic Change-Aware Attention (DCAA) mechanism. It reveals that DCAA is particularly adept at discerning change regions and achieving precise focus on areas where alterations occur. This superior performance

Table 1: Experiments on the Impact of the Number of Stacked Dynamic Change-Aware Attention Modules at Each Stage.

L_1	L_2	L_3	L_4	F1	IoU
2	2	2	2	92.71	88.10
3	3	3	3	93.67	88.64
3	3	6	3	93.55	88.44
3	3	18	3	93.66	88.62
3	4	6	3	93.75	88.76
3	6	6	3	93.54	88.42

is a direct result of the module's innovative approach to feature extraction, which efficiently orchestrates interactions between feature maps. By prioritizing change areas for extraction and enhancement from the earliest stages, DCAA minimizes the computational resources allocated to static regions, thereby enhancing the overall efficiency and accuracy of the change detection process.

Effect of Local Window Sizes. Furthermore, we experimentally validate the choice of window size M_i for different stages. As shown in Fig. 5, the overall trend indicates that the performance significantly benefits from mixed-scale windows, outperforming single-scale windows. The richer the variety of window scales, the better the performance. For the four stages, the window sizes of (3, 3, 5, 7) performs the best. The progressive local window sizes that extracts local details with a small window for shallow features and captures the overall outline with a large-scale window for deep abstract features. This progressive strategy effectively captures features of various scales, providing ample room for multiscale aggregation in Change-Enhanced Multiscale Aggregator.

Effect of Numbers of Layer in Stage. To investigate the influence of network depth on model performance, we conduct experiments by varying the number of stacked Dynamic Change-Aware Attention modules at each stage and analyze the results on the LEVIR dataset. Throughout the experiments, we maintained the number of stacked layers mostly consistent, except for the first and last stages, where we altered the depth of the middle two stages to observe the outcomes.

As presented in Tab. 1, it is evident that the results are at their lowest when the number of layers is set to 2 for each stage. With an increase in network depth, both F1 and IoU values experience a slight improvement, reaching optimal performance with the combination (3, 4, 6, 3). Consequently, we selected (3, 4, 6, 3) as the configuration for stacked modules at each stage in our model.

REFERENCES

- [1] Chao-Peng Chen, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bor-Shiun Wang. 2023. SARAS-net: scale and relation aware siamese network for change detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14187–14195.
- [2] Hao Chen, Zipeng Qi, and Zhenwei Shi. 2021. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–14.
- [3] Hao Chen and Zhenwei Shi. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing* 12, 10 (2020), 1662.
- [4] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. 2020. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2020), 1194–1206.

- [5] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. 2018. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 4063–4067.
- [6] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. 2021. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geoscience and Remote Sensing Letters* 19 (2021), 1–5.
- [7] Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. 2023. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–15.
- [8] Danyang Hong, Chunping Qiu, Anzhu Yu, Yujun Quan, Bing Liu, and Xin Chen. 2023. Multi-Task Learning for Building Extraction and Change Detection from Remote Sensing Images. *Applied Sciences* 13, 2 (2023), 1037.
- [9] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K Nandi. 2023. Ultralightweight Spatial–Spectral Feature Cooperation Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.
- [10] Shujun Li and Lianzhi Huo. 2021. Remote sensing image change detection based on fully convolutional network with pyramid attention. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 4352–4355.
- [11] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. 2020. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters* 18, 5 (2020), 811–815.
- [12] Weibin Meng, Ying Liu, Yuheng Huang, Shenglin Zhang, Federico Zaiter, Bingjin Chen, and Dan Pei. 2020. A semantic-aware representation framework for online log analysis. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–7.