

A Late-Stage Bitemporal Feature Fusion Network for Semantic Change Detection

Chenyao Zhou, Haotian Zhang, Han Guo,
Zhengxia Zou, *Member, IEEE*, and Zhenwei Shi*, *Senior Member, IEEE*

Abstract—Semantic change detection is an important task in geoscience and earth observation. By producing a semantic change map for each temporal phase, both the land use land cover categories and change information can be interpreted. Recently some multi-task learning based semantic change detection methods have been proposed to decompose the task into semantic segmentation and binary change detection subtasks. However, previous works comprise triple branches in an entangled manner, which may not be optimal and hard to adopt foundation models. Besides, lacking explicit refinement of bitemporal features during fusion may cause low accuracy. In this letter, we propose a novel late-stage bitemporal feature fusion network to address the issue. Specifically, we propose local global attentional aggregation module to strengthen feature fusion, and propose local global context enhancement module to highlight pivotal semantics. Comprehensive experiments are conducted on two public datasets, including SECOND and Landsat-SCD. Quantitative and qualitative results show that our proposed model achieves new state-of-the-art performance on both datasets.

Index Terms—Change detection, remote sensing, multi-task learning, feature fusion, semantic change detection.

I. INTRODUCTION

REMOTE sensing imagery interpretation plays an important role in geoscience and earth observation. As a fundamental task, semantic segmentation (SS) aims to classify pixels in remote sensing images into distinct land use land cover (LULC) categories for surface mapping. To better understand urbanization and its impact on environmental evolution, binary change detection (BCD) have been developed to monitor the changed regions among different temporal phases by predicting a binary mask [1, 2]. To further elevate the coarse-grained change occurrence mapping into fine-grained “from-to” semantic transition correspondence [3], semantic change detection (SCD) techniques are receiving increasing attention in recent literature. By generating a semantic mask for each temporal phase containing not only the change/no change information but also the detailed LULC semantics

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under the Grants 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. (*Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn)*)

Chenyao Zhou, Haotian Zhang, Han Guo, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

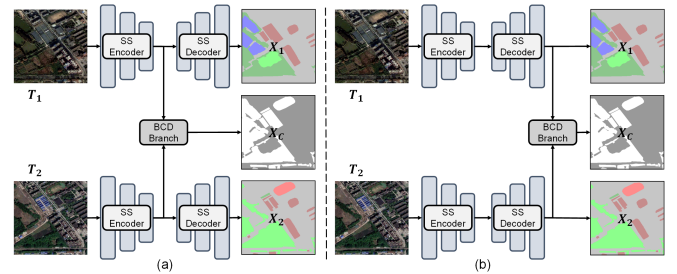


Fig. 1. Architecture comparison between previous works and our proposed model. (a) Previous works merge bitemporal SS branches from encoders. (b) Our proposed network fuse SS decoded features to achieve BCD.

within this particular temporal phase, richer change context can be demonstrated.

Before the prevalence of deep learning, traditional change detection methods adopt handcrafted features with the help of algebra, statistics and transformation [4]. With the intrinsic modeling ability of deep learning based algorithm, considerable improvements have been made mainly in the scope of bitemporal input SCD. By regarding SCD as a SS task for each temporal phase with additional “no-change” category, some CNN based and Transformer based siamese networks are implemented in an end-to-end manner [5–7]. However, without explicit constraint to regulate the change region mapping within each temporal branch, these methods struggle to suppress the changed regions discrepancy. To align together changed regions of different temporal phases, recently some multi-task learning based networks with triple branches are proposed to separately learning the LULC semantics within each temporal phase and the change location across time interval [3, 8–10]. In this scenario, two SS branches are developed to model the LULC semantics for bitemporal inputs, whilst a BCD branch is specifically designed to capture the change context. The predicted binary change mask is then utilized to filter out all the unchanged regions in predicted bitemporal semantic maps through dot product, resulting in the final predicted bitemporal semantic change maps.

According to where these triple branches communicate with each other, the aforementioned methods can be classified into two types, i.e. early-stage fusion models and middle-stage fusion models [11]. Early-stage fusion models like HRSCD [8] implement three encoders to extract features for each branch. The feature map inside BCD branch is obtained from scratch, hence fails to make good use of the semantics within dual SS branches. To better utilize the semantics, middle-stage fusion networks like [3, 9, 10] capture change context on dual SS encoded features without specific BCD encoder. To construct

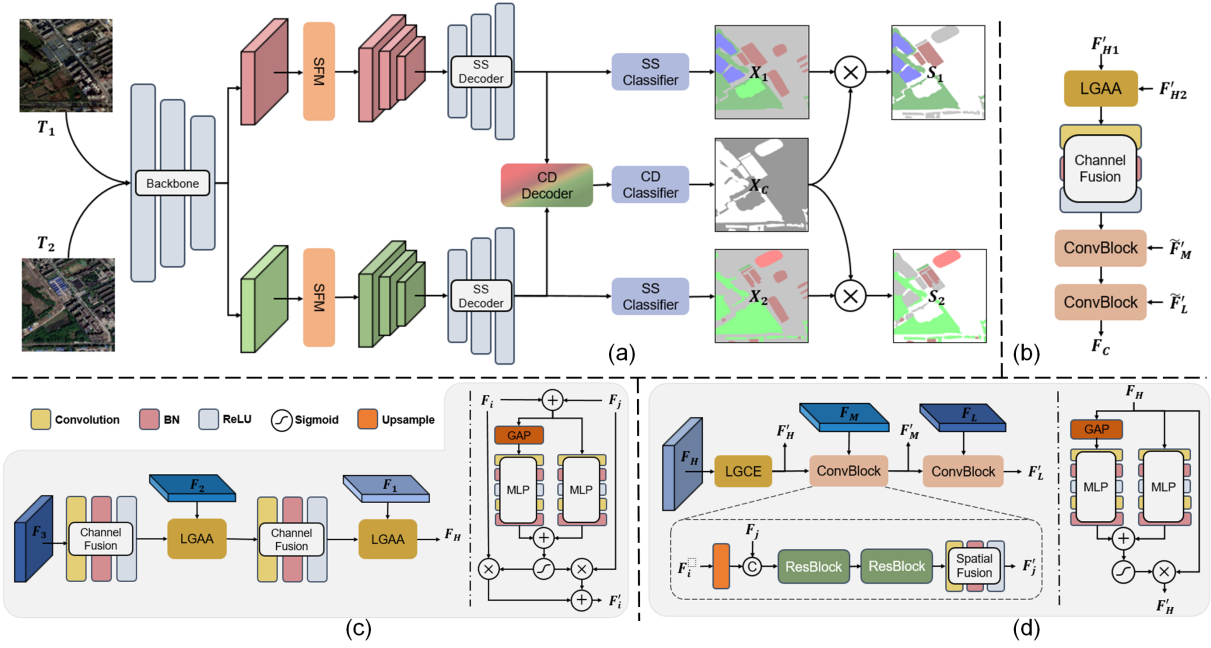


Fig. 2. Architectures of our proposed LSAFNet and its components. (a) Flowchart of LSAFNet. (b) Architecture of CD Decoder. (c) Architecture of SFM and detailed structure of LGAA. (d) Architecture of SS Decoder and detailed structure of LGCE, respectively.

dependencies between bitemporal images, these methods start correlating their triple branches from encoders. They either establish parallel BCD branch after SS encoders or jointly model semantic tokens for triple branches.

Though promising results have been achieved, we argue that their entangled design of triple branches may not be necessary and may not be optimal. For one thing, BCD can be interpreted as the “exclusive or” result of bitemporal semantic maps, thus it’s possible to capture change context directly from late-stage bitemporal semantics while achieving satisfying accuracy. From another perspective, entangled design makes it harder to adapt pretrained foundation models into downstream SCD task to transfer their modeling capability in a plug-and-play paradigm due to intermediate feature entanglement [12]. In this way, we propose a novel late-stage bitemporal feature fusion network with a shallow decoder interpreting change regions from SS decoded features. Fig. 1 shows the main difference between our proposed model and previous triple branches methods. Furthermore, [3, 5, 10] only apply naive fusion strategies like difference and concatenation when capturing change information without explicit change feature refinement. We argue that this is not sufficient for accurate change region localization and is vulnerable to irrelevant change semantics. To this end, we re-weight the primeval change features based on local and global context to boost representative ability.

The contribution of our work can be summarized as follows:

A novel SCD method LSAFNet is proposed with more decoupled architecture of two branches of SS and one BCD branch. With dual SS branches only interact in late-stage, our network achieve satisfying accuracy while being friendly to foundation model implantation.

We propose LGAA module and LGCE module to refine features based on local and global context for better

representative ability.

Comprehensive experiments are conducted on two public datasets, quantitative and qualitative studies show that our proposed LSAFNet outperforms state-of-the-art methods.

II. METHODS

A. Overall Architecture

As depicted in Fig. 2(a), the whole architecture of our proposed LSAFNet follows a multitask learning paradigm, where the SCD is decoupled into two SS branches and a BCD branch. Given two input remote sensing images T_1 and T_2 carrying different temporal information, in the early stage of LSAFNet, we first model the intra-temporal LULC semantics through encoder-decoder architecture separately without any cross-temporal interaction. By applying a visual backbone network, we extract a series of feature maps denoted as F_L , F_M , F_1 , F_2 , and F_3 , with channel dimension of 64, 64, 128, 256, and 512, respectively. Then, F_H with more pivotal semantic information is obtained from F_1 , F_2 and F_3 through semantic fusion module(SFM). F_L , F_M and F_H are further up-sampled and decoded layer by layer in the following SS decoder, and the corresponding semantic mask is predicted through its classifier. By now the semantic features from both temporal phases haven’t meet each other, until the intermediate feature maps from SS decoders are aggregated inside BCD decoder. We utilize local-global attentional aggregation module to highlight the semantic differences across time interval while suppressing the irrelevant changes, and adopt cascaded convolution blocks to connect varied stages. Ultimately, a change region binary mask is obtained through change detection classifier, and we take it as guidance to mask out unchanged areas in both semantic masks to achieve the final semantic change predictions.

B. Semantic Fusion Module

The key to achieve satisfying SCD result lies in identifying and matching every pixel's semantic category between the given two input images from different temporal phases. Due to the intrinsic nature of remote sensing images having rich background context and varied object scales, the raw features extracted by backbone network suffers from the perplexity of inter-class similarity and intra-class variability. The different imaging periods of multi-temporal image series further bring in interference factors such as irrelevant seasonal and illuminating changes [13]. Therefore, it's crucial to construct more representative features to facilitate downstream SS and BCD subtasks. To this end, motivated by [10], we propose semantic fusion module(SFM) to aggregate features F_1 , F_2 and F_3 into a more representative feature map F_H layer by layer. After channel reduction through pointwise convolution, the lower-level feature map is aggregated with its next level counterpart in LGAA module. The process can be expressed as follows:

$$F'_2 = \text{LGAA}(F_2, \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(F_3)))) \quad (1)$$

$$F_H = \text{LGAA}(F_1, \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(F'_2)))) \quad (2)$$

where BN represents the BatchNorm operation.

The proposed SFM distinguish from previous methods mainly on local-global attentional aggregation module(LGAA). Inspired by [14], we fuse adjacent levels of features with explicit per-channel re-weighting. By GAP, a global representative vector is obtained from the summation of two input feature maps. We further utilize two layers of Conv-BN-ReLU as the local channel context aggregator, and combine the local and global channel context through addition. The re-weighting vectors are subsequently applied to their corresponding feature maps and the results are added as the output of LGAA. Take F_i and F_j as the LGAA's inputs, the output F'_i can be expressed as follows:

$$w_1 = \text{Conv}(\text{BN}(\text{ReLU}(\text{Conv}(\text{BN}(\text{GAP}(F_i + F_j)))))) \quad (3)$$

$$w_2 = \text{Conv}(\text{BN}(\text{ReLU}(\text{Conv}(\text{BN}(F_i + F_j)))))) \quad (4)$$

$$F'_i = F_i \text{Sigmoid}(w_1 + w_2) + F_j (1 - \text{Sigmoid}(w_1 + w_2)) \quad (5)$$

where GAP represents the global average pooling operation.

C. Semantic Segmentation Decoder

As shown in Fig. 2(d), we use two parallel weight-sharing decoders for SS branches. The SS Decoder mainly comprises a local-global context enhancement module(LGCE) and two convolution blocks, gradually upsamples and aggregates adjacent levels of input feature maps for the final semantic map prediction and cross-temporal interaction.

Following the practice in [10], the ConvBlock consists of upsampling, concatenation and two cascaded ResBlocks to combine the two input features, and apply depthwise convolution to further fuse the spatial context. The high level feature containing pivotal semantics, denoted as F_H , is first processed in LGCE to distinguish interested semantics from interference

factors through channel attention. Similar to the calculation of LGAA, the output F'_H of LGCE can be formulated as follows:

$$w_1 = \text{Conv}(\text{BN}(\text{ReLU}(\text{Conv}(\text{BN}(\text{GAP}(F_H)))))) \quad (6)$$

$$w_2 = \text{Conv}(\text{BN}(\text{ReLU}(\text{Conv}(\text{BN}(F_H)))))) \quad (7)$$

$$F'_H = F_H \text{Sigmoid}(w_1 + w_2) \quad (8)$$

D. Change Detection Decoder

The above SS encoder-decoder branch only captures intra-temporal LULC categories within each temporal phase. To identify the changed region across two temporal phases and project intra-temporal LULC categories into cross-temporal change region semantics, we propose a simple yet efficient bridging decoder between two temporal branches in the late-stage of our network to achieve bitemporal interaction. The proposed CD Decoder, as depicted in Fig. 2(b), receives three levels of decoded features from both SS Decoders and generate feature map F_C related to change regions. We first implement the same LGAA module in Sec. II-B to merge high-level feature maps from both temporal branches. Then, we apply pointwise convolution to reduce its channel dimension and aggregates its semantic information with the spatial information from two subsequent lower-level feature maps layer by layer. For simplicity, we subtract one SS Decoder's output feature map from its counterpart of another SS Decoder and keep the absolute value as the lower-level features being processed in ConvBlocks. Given F'_{H1} , F'_{M1} , F'_{L1} from T_1 SS Decoder, and F'_{H2} , F'_{M2} , F'_{L2} from T_2 SS Decoder, the F_C can be calculated as follows:

$$\tilde{F}_H = \text{LGAA}(F'_{H1} + F'_{H2}) \quad (9)$$

$$\tilde{F}'_H = \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(\tilde{F}_H))) \quad (10)$$

$$\tilde{F}'_M = |F'_{M1} - F'_{M2}| \quad (11)$$

$$\tilde{F}'_L = |F'_{L1} - F'_{L2}| \quad (12)$$

$$F'_C = f(\tilde{F}'_H, \tilde{F}'_M) \quad (13)$$

$$F_C = f(F'_C, \tilde{F}'_L) \quad (14)$$

where $f(\cdot)$ represents ConvBlock described in Sec. II-C, $|\cdot|$ represents the absolute value operator.

E. Loss Function

Our multitask schemed network produces three prediction maps in total, namely X_1 , X_2 , and X_C . The X_1 and X_2 serves as the LULC semantic maps correspond to each temporal phase, while X_C is a binary change mask, denoting the changed regions across time. In this work we supervise over X_1 , X_2 and X_C instead of semantic change maps. We choose the multi-class cross-entropy loss for semantic maps optimization and the binary cross-entropy loss for change region supervision. The formulation of \mathcal{L}_{SS} and \mathcal{L}_{BCD} can be expressed as

$$\mathcal{L}_{SS} = -\frac{1}{N} \sum_{i=1}^N y_i \log(p_i) \quad (15)$$

$$\mathcal{L}_{BCD} = -y_c \log(p_c) - (1 - y_c) \log(1 - p_c) \quad (16)$$

where N represents the number of categories in the semantic maps, y_i and p_i represents the groundtruth label index and the predicted probability of each category respectively, and y_c and p_c represents the groundtruth label index and the corresponding predicted probability of change region in the binary change map. We ignore the no-change class in the semantic change labels to maintain the semantic category consistency between semantic change labels and predicted masks. To better align the bitemporal SS subtask and BCD subtask, a semantic consistency loss \mathcal{L}_{SC} is proposed in [3] as

$$\mathcal{L}_{SC} = \begin{cases} 1 - \cos(x_1, x_2), y_c = 1 \\ \cos(x_1, x_2), y_c = 0 \end{cases} \quad (17)$$

where x_1 and x_2 signify the feature vectors of a pixel in X_1 and X_2 respectively. The total loss \mathcal{L} implemented through this paper is defined as follows:

$$\mathcal{L} = \mathcal{L}_{BCD} + 0.5 \times (\mathcal{L}_{SS1} + \mathcal{L}_{SS2}) + \mathcal{L}_{SC} \quad (18)$$

III. EXPERIMENTS

A. Datasets and Evaluation Metrics

To verify the effectiveness of our model, we conduct experiments on two publicly available SCD dataset, including SECOND [15] and Landsat-SCD [7]. SECOND dataset consists of 2968 pairs of bitemporal images of size 512×512 , with resolution ranging from 0.5m to 3.0m, including building, water, tree, low vegetation, ground and playground. Landsat-SCD dataset comprises 8468 pairs of bitemporal images of size 416×416 , with a consistent resolution of 30m, including water, farmland, building and desert.

For fair comparison, we keep the same scaling and partition strategy as previous work [3, 9, 10] throughout the whole experiment. To quantitatively measure the similarity between the predicted bitemporal semantic change probability maps and their corresponding labels, we introduce four well-established indicators, including mIoU and Avg evaluate the overall segmentation performance, as well as SeK and F_{scd} specifically focus on the semantic discrimination within changed regions.

B. Implementation Details

Experiments are conducted with PyTorch on two NVIDIA RTX4090 GPUs. We deploy ResNet-34 as backbone, initialize our network with Kaiming Initialization [16] and train it for 50 epochs with batch size of 8. SGD with weight decay of $5e-4$ and momentum factor of 0.9 is selected as the optimizer, with initial learning rate set to 0.07. Common data augmentation including random flipping and rotation is carried out during training. Code are publicly available at <https://github.com/STORMTROOPERRR/RSISCD>.

C. Comparison and Analysis

We compare our proposed model with other state-of-the-art methods for performance evaluation. Quantitative results are listed in Table I and Table II for each dataset, with best results marked in **bold**. Statistics show that our proposed

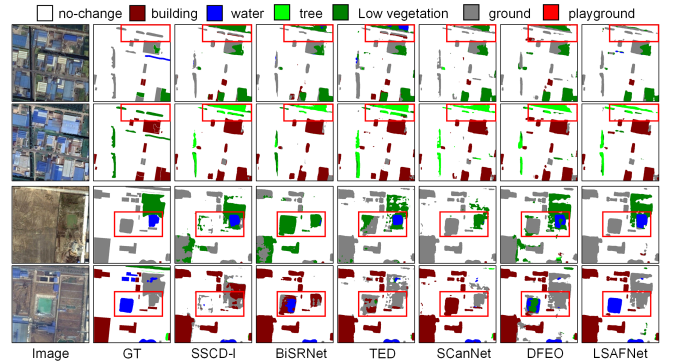


Fig. 3. Qualitative comparisons of the results on SECOND dataset. First two rows and last two rows contain different bitemporal image pairs, respectively.

model achieve new SOTA results on both dataset, especially on changed regions across time. For more intuitive demonstration, we select two pairs of bitemporal images from both datasets to qualitatively evaluate different models' performance in Fig. 3 and Fig. 4. Note that for simplicity, we only display the best six of all competing methods. We further highlight key areas where different models perform most diversely with red box. Fig. 3 shows that our model can achieve high intra-category consistency with strong semantic capture ability. Fig. 4 reveals our model's promising capability of modeling changed regions with various scales and delicate contour. Though being more decoupled, our proposed BCD branch succeeds in capture the change context between bitemporal images, while two SS branches maintain high accuracy in modeling each temporal phase's LULC semantics. The local global context aggregation guides SS encoders to extract representative features, ensuring the satisfying performance for both SS and BCD subtasks.

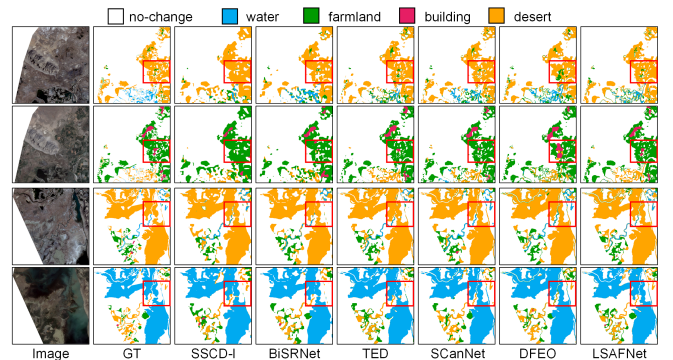


Fig. 4. Qualitative comparisons of the results on Landsat dataset. First two rows and last two rows contain different bitemporal image pairs, respectively.

D. Ablation Study

To quantitatively measure the improvements brought by each core component of our proposed model, we further conduct a series of ablation study on two datasets. We define our base model as the proposed LSAFNet without LGAA module and LGCE module. To deactivate LGAA module, we first concatenate two input features and apply a MLP to perform channel reduction. We replace LGCE module with identity when needed. Experimental results in Table III suggest

TABLE I
NUMERICAL RESULTS OF DIFFERENT MODELS ON SECOND

Method	mIoU(%)	Avg(%)	SeK(%)	Fscd(%)
FC-Siam-conv	68.86	86.92	16.61	56.45
FC-Siam-diff	68.96	86.86	16.50	56.23
HRSCD	71.15	86.62	18.80	58.21
SCDNet	70.95	87.29	19.75	59.77
SSCD-1	72.60	87.19	21.86	61.22
Bi-SRNet	73.38	87.48	22.43	61.62
TED	73.05	87.20	22.37	61.23
SCanNet	73.20	87.46	23.34	62.59
DEFO-MLTSCD	73.76	87.80	23.73	62.73
LSAFNet (Ours)	74.01	87.66	24.32	63.20

that our proposed LGAA module and LGCE module both have its own contribution to the overall performance of our proposed LSAFNet. LGAA module is utilized both in SS encoders and BCD decoder, thus having a major impact on elevating our proposed model’s capability.

TABLE II
NUMERICAL RESULTS OF DIFFERENT MODELS ON LANDSAT-SCD

Method	mIoU(%)	Avg(%)	SeK(%)	Fscd(%)
FC-Siam-conv	79.31	90.79	36.11	76.04
FC-Siam-diff	77.68	88.53	32.75	73.89
HRSCD	78.51	91.47	32.90	73.20
SCDNet	80.14	93.62	40.05	75.17
SSCD-1	79.33	92.36	41.43	75.84
Bi-SRNet	82.19	93.16	40.09	76.01
TED	84.22	93.98	45.60	78.47
SCanNet	85.19	94.07	49.33	80.52
DEFO-MLTSCD	87.49	94.32	49.26	81.39
LSAFNet (Ours)	87.60	94.46	49.94	81.66

TABLE III
ABLATION STUDY ON TWO DATASETS

Method	SECOND		Landsat-SCD	
	mIoU(%)	SeK(%)	mIoU(%)	SeK(%)
Base	73.59	23.37	86.09	48.67
Base + LGAA	73.82	23.90	87.45	49.57
Base + LGCE	73.69	23.69	87.44	49.29
LSAFNet (Ours)	74.01	24.32	87.60	49.94

IV. CONCLUSION

In this letter, to design a multi-task learning based SCD network in a more disentangled manner, we propose a novel late-stage bitemporal feature fusion network LSAFNet that only bridge bitemporal features in decoder stage. To extract more representative features in dual SS encoders, we proposed LGAA module to refine feature maps through aggregated local and global context re-weighting, and further utilize it to highlight change context across time while suppressing irrelevant changes. We further propose LGCE module to enhance the high-level features in SS decoders to boost LULC semantics modeling. Experiments on two public datasets verify our model’s effectiveness, and ablation study confirms each component’s contribution. We will harness our proposed architecture’s disentanglement strengths to adapt pretrained foundation models into SCD field in our future work.

REFERENCES

- [1] L. Bruzzone and F. Bovolo, “A novel framework for the design of change-detection systems for very-high-resolution remote sensing images,” *Proceedings of the IEEE*, vol. 101, no. 3, pp. 609–630, 2012.
- [2] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, “A survey on deep learning-based change detection from high-resolution remote sensing images,” *Remote Sensing*, vol. 14, no. 7, p. 1552, 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 Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [4] E. J. Parelus, “A review of deep-learning methods for change detection in multispectral remote sensing images,” *Remote Sensing*, vol. 15, no. 8, p. 2092, 2023.
- [5] R. C. Daudt, B. Le Saux, and A. Boulch, “Fully convolutional siamese networks for change detection,” in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 4063–4067.
- [6] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, “Scdnet: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102465, 2021.
- [7] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, “A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images,” *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1506–1525, 2022.
- [8] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, “Multi-task learning for large-scale semantic change detection,” *Computer Vision and Image Understanding*, vol. 187, p. 102783, 2019.
- [9] L. Ding, J. Zhang, H. Guo, K. Zhang, B. Liu, and L. Bruzzone, “Joint spatio-temporal modeling for semantic change detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [10] Z. Li, X. Wang, S. Fang, J. Zhao, S. Yang, and W. Li, “A decoder-focused multi-task network for semantic change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [11] G. Cheng, Y. Huang, X. Li, S. Lyu, Z. Xu, Q. Zhao, and S. Xiang, “Change detection methods for remote sensing in the last decade: A comprehensive review,” *arXiv preprint arXiv:2305.05813*, 2023.
- [12] K. Li, X. Cao, and D. Meng, “A new learning paradigm for foundation model-based remote-sensing change detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.
- [13] H. Zhang, H. Chen, C. Zhou, K. Chen, C. Liu, Z. Zou, and Z. Shi, “Bifa: Remote sensing image change detection with bitemporal feature alignment,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [14] Y. Dai, F. Giesecke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560–3569.
- [15] K. Yang, G.-S. Xia, Z. Liu, B. Du, W. Yang, M. Pelillo, and L. Zhang, “Semantic change detection with asymmetric siamese networks,” *arXiv preprint arXiv:2010.05687*, 2020.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.