UNIFYING OPTICAL FLOW AND OPTIMAL TRANSPORT FOR SEMANTIC CHANGE DETECTION UNDER MISREGISTRATION

Xinyang Song[1,2], *Yunhao Gao*[1,2,⋆], *Xiaofei Yin*[3], *Jianbu Wang*[3], *and Mengmeng Zhang*[1,2]

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2] Beijing Institute of Technology, Zhuhai, Guangdong 519088, China
[3] First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China
Emails: songxy9037@163.com; gaoyunhao@bit.edu.cn; yxf@fio.org.cn; wangjianbu@fio.org.cn; mengmengzhang@bit.edu.cn

ABSTRACT

Semantic change detection (SCD) identifies changed regions and land-cover transitions in bitemporal remote sensing imagery. Although deep models have advanced SCD, most assume accurate image registration and may confuse geometric displacement with semantic change when this assumption is violated. We propose UniFOT, a framework that combines optical flow and optimal transport for SCD under misregistration. Its Hybrid Alignment Matching Unit (HAMU) contains two complementary branches. The Flow Warp Module estimates local displacement fields and aligns bitemporal features, whereas Global Transport Attention establishes nonlocal semantic correspondence through spatially regularized optimal transport. A Geometry–Context Gated Mixer then adaptively fuses the alignment and association cues before multitask decoding. Experiments on SECOND and JL1 with simulated spatial perturbations show that UniFOT improves semantic transition recognition and reduces pseudo-changes caused by registration errors. These results support the complementary use of geometric alignment and semantic matching for SCD under controlled misregistration.

***Index Terms*—** Semantic change detection, remote sensing, image misregistration, optimal transport, optical flow, feature alignment

1. INTRODUCTION

Semantic change detection (SCD) jointly identifies changed regions and their land-cover transitions in bitemporal remote sensing images [1, 2]. Compared with binary change detection, SCD provides category-level transition information for applications such as urban planning and resource monitoring. Most existing studies, however, assume pixel-aligned image pairs. In practical remote sensing, variations in viewing geometry, platform pose, terrain, and preprocessing can leave residual spatial displacement [3]. Such misregistration makes unchanged objects appear at different locations and can therefore produce pseudo-changes that interfere with semantic transition recognition.

At the feature level, geometric displacement and semantic evolution are entangled in the same bitemporal discrepancy. An unchanged object with shifted boundaries can produce a response similar to a genuine transition, whereas an overly aggressive alignment operation may warp or suppress evidence from newly appearing and disappearing objects. The central difficulty is therefore not simply to register two feature maps, but to determine which discrepancies should be geometrically corrected and which should be retained for semantic interpretation.

Multitask Siamese networks have become a common solution for SCD [4]. Daudt et al. [1] established an early fully convolutional baseline that coupled change detection with bitemporal semantic segmentation. Subsequent CNN-based methods strengthened task interaction by reusing semantic features or reasoning over cross-temporal relationships [5, 6]. Transformer-based models further enlarged the effective context. ChangeMask uses a hierarchical Transformer to model multiscale semantics [7], while Pyramid-SCDFormer targets small changes and boundary details [8]. SCanNet explicitly models spatio-temporal dependencies and semantic transitions [9]. Despite these advances, these methods are primarily designed and evaluated under the same accurate-registration assumption.

Recent change-detection models have introduced adaptive sampling or feature alignment to reduce spatial discrepancies. Deformable convolution can learn data-dependent sampling offsets [10], and optical-flow estimation can provide displacement fields for feature warping [11]. BiFA explicitly aligns bitemporal features before comparison to suppress inconsistent responses [12]. These approaches improve local geometric adaptation, but a single alignment mechanism remains vulnerable under semantic changes. Flow estimation is reliable when corresponding content is preserved, yet it becomes ambiguous when an object disappears, emerges, or changes appearance. Conversely, global content matching can recover long-range correspondence but may associate semantically similar regions that are spatially unrelated. Robust SCD under misregistration therefore requires local geometric correction and global semantic association to be considered jointly.

To address this requirement, we propose UniFOT, a modular framework built on a weight-sharing encoder and a lightweight multitask decoder, as shown in Fig. 1. Its core Hybrid Alignment Matching Unit (HAMU) is inserted at multiple encoder scales. Within each HAMU, the Flow Warp Module (FWM) estimates a dense residual displacement field and warps bitemporal features to correct local spatial shifts. In parallel, Global Transport Attention (GTA) formulates correspondence as an optimal transport problem and uses spatial regularization to constrain nonlocal semantic matching. Because the reliability of these cues varies across image pairs and spatial locations, the Geometry–Context Gated Mixer (GCGM) uses dynamic convolution and spatial gating to integrate them. The fused differ-

This work was supported by the National Natural Science Foundation of China (Grant No. 62501051) and the NSFC Projects of International Cooperation and Exchanges (Grant No. W2411055). ⋆Corresponding author: Yunhao Gao (gaoyunhao@bit.edu.cn).

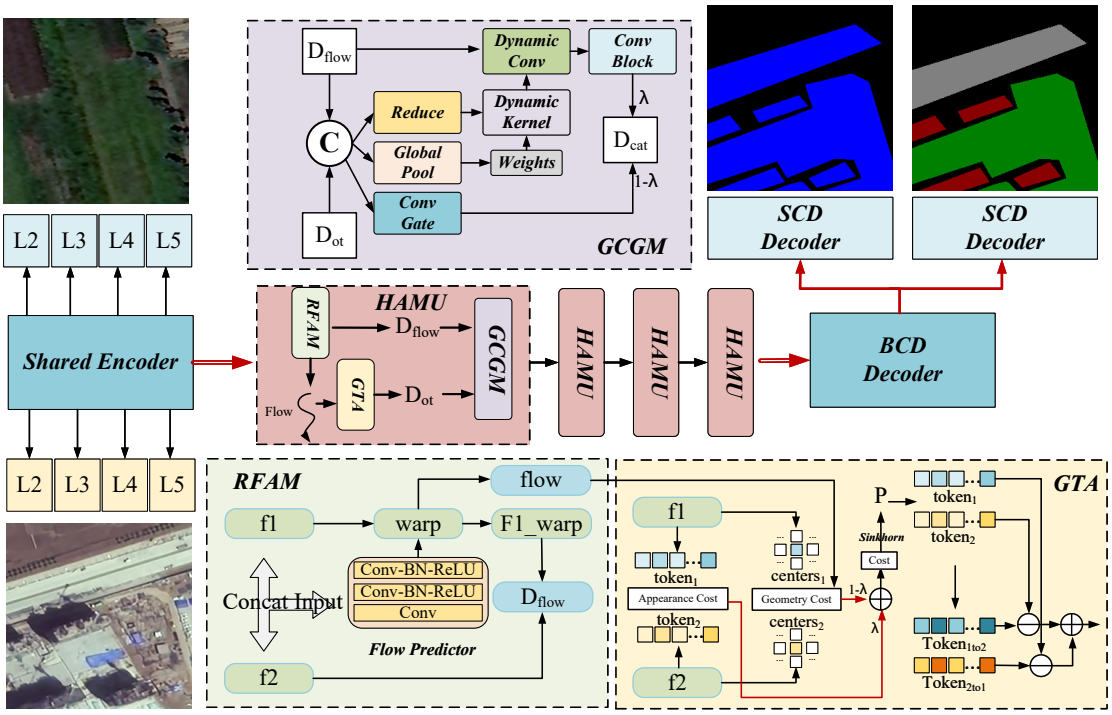


Fig. 1. Overview of UniFOT. HAMU combines flow-based local alignment (FWM) and transport-based global association (GTA), after which GCGM adaptively fuses the two difference cues for multitask decoding.

ence features are then injected into the semantic branches to coordinate semantic segmentation and binary change detection.

The main contributions are summarized as follows:

- We develop UniFOT for SCD under misregistration. The framework treats spatial displacement and semantic change as coupled correspondence problems within a standard encoder–decoder architecture.
- We introduce HAMU to combine FWM-based local alignment with GTA-based global semantic association. GCGM adaptively reconciles the two cues and reduces interference from unreliable correspondence.
- Experiments and component ablations on SECOND and JL1 under simulated affine perturbations show consistent improvements in F_{scd} and SeK over the compared SCD methods.

2. METHODOLOGY

As illustrated in Fig. 1, UniFOT consists of a weight-sharing encoder, multiscale HAMUs, and a multitask decoder. Its design decomposes correspondence reasoning into two complementary hypotheses. FWM assumes that locally consistent content can be related through an explicit displacement field and produces a geometry-corrected residual. GTA relaxes this displacement assumption and searches globally for semantically compatible content through optimal transport. GCGM evaluates the two observations jointly and uses them to refine the local residual according to image-level context and spatially varying reliability. Applying HAMU at four encoder scales allows this reasoning to cover fine boundary displacement and broader object-level inconsistency. The decoder then aggregates the refined differences to predict two semantic maps and one binary change map, while difference injection couples change localization with bitemporal land-cover recognition.

2.1. Flow Warp Module

FWM performs preliminary alignment between bitemporal features. Given $F_1, F_2 \in \mathbb{R}^{C\times H\times W}$, the two features are concatenated and processed by a three-layer convolutional predictor Φ. The predictor regresses a dense flow field $V \in \mathbb{R}^{2\times H\times W}$:

$$V = \Phi(\mathrm{Concat}[F_1, F_2]), \tag{1}$$

where the two channels of V represent horizontal and vertical displacement. A bilinear warping operation $\mathcal{W}$ transforms F_1 toward the coordinate system of F_2:

$$F_1^{warp} = \mathcal{W}(F_1, V). \tag{2}$$

The absolute difference between the warped feature and the target feature is then computed as

$$D_{flow} = |F_1^{warp} - F_2|. \tag{3}$$

This operation converts an implicit coordinate discrepancy into an explicit feature residual. Where corresponding content is preserved, warping brings the same object into a common coordinate system and suppresses boundary responses caused only by displacement. The remaining activation is therefore more concentrated on semantic change and unresolved correspondence. This distinction is important

because FWM is not expected to explain every temporal difference: at newly appearing, disappearing, or strongly altered regions, the flow estimate can be ambiguous. Its output is consequently treated as a strong local geometric cue rather than a complete change representation.

2.2. Global Transport Attention

The receptive field and correspondence assumption of FWM limit its reliability under large displacement or substantial appearance change. GTA provides a second view of the discrepancy by establishing content-aware correspondence over the complete feature map. Given F_1 and F_2, we flatten them into token sequences $X_1, X_2 \in \mathbb{R}^{N\times C}$, where $N = HW$. Instead of regressing one displacement vector for each position, GTA formulates matching as a globally coupled optimal-transport problem. Its cost matrix combines semantic and spatial distances:

$$C_{ij} = 1 - \frac{x_{1,i}^{\top} x_{2,j}}{\|x_{1,i}\|\|x_{2,j}\|} + \lambda\|p_i - p_j\|_2^2, \tag{4}$$

where $x_{1,i}$ and $x_{2,j}$ denote feature tokens, p_i and p_j are their normalized coordinates, and λ balances semantic similarity and spatial proximity. The cosine term makes the assignment responsive to land-cover content, whereas the coordinate penalty discourages semantically plausible but spatially unrelated matches. Their combination preserves nonlocal search while retaining a weak geometric preference, which is more flexible than restricting correspondence to a fixed window.

An entropically regularized Sinkhorn solver computes the transport plan:

$$P = \mathrm{Sinkhorn}(C, \epsilon), \tag{5}$$

where $P \in \mathbb{R}^{N\times N}$ and ϵ controls the regularization strength. Because all token assignments are optimized jointly, P represents a globally normalized soft correspondence rather than a set of independent nearest-neighbor matches. This coupling distributes ambiguous evidence across compatible regions and reduces the dominance of a single locally attractive match. The plan transports features in both temporal directions:

$$\widetilde{X}_{1\to 2} = P^{\top} X_1, \qquad \widetilde{X}_{2\to 1} = P X_2. \tag{6}$$

After reshaping the transported sequences to the original spatial dimensions, GTA obtains the bidirectional transport difference

$$D_{ot} = |\widetilde{F}_{1\to 2} - F_2| + |\widetilde{F}_{2\to 1} - F_1|. \tag{7}$$

The bidirectional construction exposes discrepancies that may be weak or ambiguous in only one transport direction. Unlike flow warping, GTA does not require every position to be explained by a unique local displacement. It therefore supplies a complementary semantic correspondence cue when local geometry alone is insufficient, while the spatial term prevents unconstrained matching across repeated land-cover patterns.

2.3. Geometry–Context Gated Mixer

The reliability of D_{flow} and D_{ot} varies with displacement, appearance change, and local content. A fixed summation implicitly assumes that both cues are equally reliable everywhere and can therefore amplify conflicting correspondence. GCGM instead performs hierarchical conditioning through dynamic convolution and spatial gating. The concatenated cues are first summarized by global average pooling (GAP), and a multilayer perceptron predicts mixture weights $\mathbf{w} \in \mathbb{R}^K$ for K learnable basis kernels:

$$\mathbf{w} = \mathrm{Softmax}(\mathrm{MLP}(\mathrm{GAP}(\mathrm{Concat}[D_{flow}, D_{ot}]))) . \tag{8}$$

The instance-specific dynamic kernel is constructed as

$$\mathcal{K}_{dyn} = \sum_{k=1}^{K} w_k \mathcal{K}_k. \tag{9}$$

The basis kernels encode a set of reusable transformations, while $\mathbf{w}$ selects an image-pair-specific combination according to the joint geometry–semantic context. A bottleneck with reduction ratio r limits the computational cost. The dynamically enhanced feature is

$$F_{dyn} = \mathrm{Expand}(\mathcal{K}_{dyn} * \mathrm{Reduce}(D_{flow})) . \tag{10}$$

GCGM then predicts a spatial gate $G \in \mathbb{R}^{1\times H\times W}$ from both correspondence cues:

$$G = \sigma(\mathrm{Conv}_{1\times 1}(\mathrm{Concat}[D_{flow}, D_{ot}])) . \tag{11}$$

The fused difference representation is defined as

$$D_{fused} = G \odot F_{dyn} + (1 - G) \odot D_{flow}, \tag{12}$$

where $\odot$ denotes element-wise multiplication. This formulation deliberately avoids directly adding D_{ot} to D_{flow}. Instead, the transport cue conditions both the dynamic transformation and the spatial gate, allowing global semantic association to indicate where the local residual should be recalibrated without replacing its geometric reference. The global kernel mixture captures image-level correspondence characteristics, whereas G resolves reliability at individual spatial locations. GCGM thus retains stable flow-aligned evidence in well-matched regions and applies context-dependent correction where the two branches reveal inconsistent correspondence.

2.4. Encoder and Multitask Decoder

A pretrained ResNet-34 serves as the weight-sharing encoder and extracts feature pyramids at four scales. Weight sharing places the two dates in a common representation space, while multiscale HAMUs address displacement at different spatial extents. The decoder contains one binary change-detection branch and two weight-sharing semantic segmentation branches, with each block using bilinear upsampling followed by convolution. At every scale, D_{fused} is concatenated with the corresponding encoder feature before semantic decoding. This difference injection prevents the semantic branches from reasoning independently of the detected change evidence and encourages the two land-cover predictions to remain compatible with the binary change map.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Implementation Details

UniFOT is implemented in PyTorch 2.4.1 on Ubuntu 20.04 and trained on an NVIDIA RTX 4090 GPU. AdamW uses an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-3}, and batch size 16. Training lasts 50 epochs with random cropping, rotation, and flipping. The checkpoint with the best validation performance is retained, and all experiments use a fixed random seed.

Table 1. Component ablation on SECOND and JL1. Inj.: difference injection.

Idx.	Components				Metrics			
	FWM	GTA	GCGM	Inj.	OA	F_{scd}	mIoU	SeK
					SECOND			
1				✓	86.37	58.60	70.37	18.45
2	✓			✓	86.16	59.37	70.40	19.00
3		✓		✓	86.25	58.48	69.98	18.16
4	✓	✓		✓	86.54	59.32	70.60	19.14
5	✓	✓	✓		86.79	58.62	70.92	18.68
6	✓	✓	✓	✓	**86.94**	**60.63**	**71.42**	**20.46**
					JL1			
1				✓	89.68	80.17	79.43	42.61
2	✓			✓	90.42	81.92	80.59	45.53
3		✓		✓	89.98	81.29	79.55	43.84
4	✓	✓		✓	90.36	81.76	80.34	45.07
5	✓	✓	✓		89.88	80.53	80.08	43.54
6	✓	✓	✓	✓	**91.00**	**82.94**	**81.62**	**47.61**

To evaluate robustness against registration errors, a random affine transformation is applied to the second image T_2 and its label with a fixed probability, while T_1 remains unchanged. The perturbation includes translation, rotation within $\pm 5^{\circ}$, and scaling from 0.95 to 1.05. This protocol provides a controlled evaluation of geometric misregistration rather than covering every source of real-world displacement.

3.2. Datasets

We evaluate UniFOT on SECOND [6] and JL1. SECOND contains 4,662 RGB image pairs of 512×512 pixels with spatial resolutions from 0.5 to 3 m. It provides annotations for six land-cover classes and exhibits severe class imbalance across semantic transitions.

JL1 is a cropland change-detection competition dataset collected by the Jilin-1 satellite. It contains 6,000 image pairs of 256×256 pixels at 0.75 m spatial resolution. Five land-cover classes define nine annotated "from-to" transition categories.

3.3. Ablation Study

Table 1 evaluates FWM, GTA, GCGM, and difference injection on both datasets. The baseline in row 1 contains the multitask decoder and difference injection but excludes the three HAMU components. Adding FWM alone (row 2) improves F_{scd} and SeK on both datasets, showing that local feature warping reduces errors induced by displacement. GTA alone (row 3) is less effective on SECOND but improves the primary metrics on JL1. This difference indicates that global semantic association is sensitive to correspondence ambiguity when it lacks an explicit local alignment cue.

Directly summing FWM and GTA (row 4) does not consistently outperform FWM alone. Their cues can conflict when flow warping and global matching assign different correspondences. Replacing static summation with GCGM (row 6) improves all metrics over row 4 on both datasets. The result supports adaptive rather than uniform fusion of local geometric and global semantic evidence.

Removing difference injection from the full model (row 5) reduces all four metrics on both datasets. Thus, passing D_{fused} to the semantic branches is important for coordinating change localization with bitemporal land-cover prediction.

Table 2. Comparison on SECOND and JL1. Best and second-best values are bold and underlined.

Method	OA	F_{scd}	mIoU	SeK
	SECOND			
HGINet	84.60	51.73	66.37	11.83
CDSC	85.57	53.59	67.01	13.49
Bi-SRNet	86.84	59.73	71.06	19.57
BT-SCD	86.82	60.13	70.67	<u>20.36</u>
SCanNet	86.61	<u>60.39</u>	70.72	19.90
SSCDl	**87.00**	60.26	<u>71.28</u>	20.04
UniFOT	<u>86.94</u>	**60.63**	**71.42**	**20.46**
	JL1			
HGINet	83.23	67.79	69.39	24.46
CDSC	85.55	70.77	73.67	29.44
Bi-SRNet	89.75	80.71	79.82	43.62
BT-SCD	90.40	81.31	80.82	45.05
SCanNet	89.86	<u>81.57</u>	79.64	44.44
SSCDl	<u>90.54</u>	81.54	<u>81.35</u>	<u>45.95</u>
UniFOT	**91.00**	**82.94**	**81.62**	**47.61**

3.4. Comparison with State-of-the-Art Methods

We compare UniFOT with six representative SCD methods: BT-SCD [13], CDSC [14], HGINet [15], SCanNet [9], Bi-SRNet, and SSCDl [6]. All results in Table 2 are obtained under the same perturbed evaluation setting.

On SECOND, UniFOT achieves the best F_{scd}, mIoU, and SeK. Its F_{scd} and SeK exceed the second-best results by 0.24 and 0.10 percentage points, respectively. SSCDl obtains a slightly higher OA, but its lower F_{scd} and SeK indicate weaker recognition of semantic transitions. This distinction is important because unchanged pixels dominate OA on class-imbalanced SCD data. The gains in the primary metrics are consistent with HAMU reducing displacement-induced false transitions while retaining semantic change evidence.

On JL1, UniFOT ranks first in all four metrics. It exceeds the second-best method by 1.37 points in F_{scd} and 1.66 points in SeK. HGINet and CDSC show larger degradation under the perturbed setting, whereas Bi-SRNet, SCanNet, and SSCDl remain more competitive through stronger temporal or task interaction. UniFOT further improves their results by explicitly combining local alignment and nonlocal association. The larger gain on JL1 is consistent with the value of this complementary correspondence design, although the current protocol does not isolate every source of dataset difficulty.

3.5. Qualitative Visualization

Figure 2 compares representative predictions. In the SECOND examples, CDSC and HGINet produce fragmented building regions, while UniFOT recovers more contiguous shapes and boundaries. In the JL1 examples, several methods miss narrow cropland structures or produce irregular boundaries after displacement. UniFOT better preserves these structures and shows fewer spatially shifted errors. The visual observations agree with the improvements in F_{scd} and SeK, but they are illustrative rather than a substitute for the full quantitative evaluation.

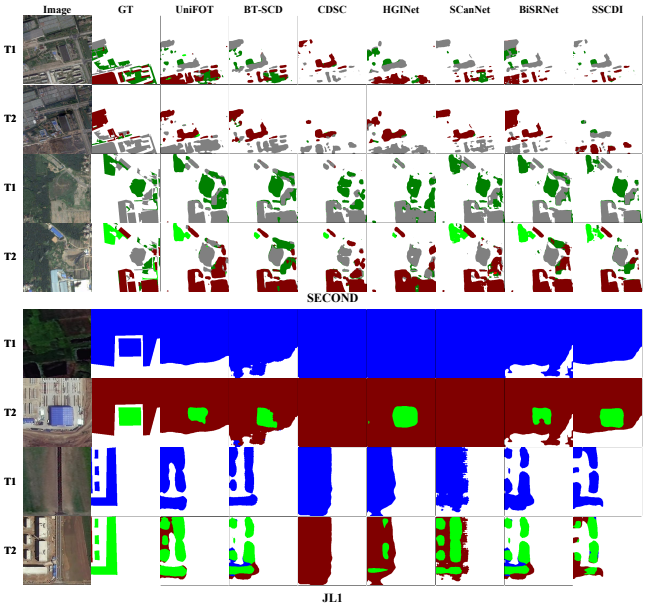


Fig. 2. Qualitative comparison on SECOND (top) and JL1 (bottom). Columns show the input image, ground truth, UniFOT prediction, and predictions from the compared methods.

4. CONCLUSION

We presented UniFOT for semantic change detection under misregistration. HAMU combines FWM-based local spatial correction with GTA-based global semantic association, and GCGM adaptively integrates the resulting cues before multitask decoding. Experiments on perturbed SECOND and JL1 data show consistent improvements in F_{scd} and SeK, while the ablations support the contributions of adaptive fusion and difference injection. The current evidence is limited to controlled affine perturbations on two datasets. Future work will examine naturally misregistered multisensor imagery and reduce the cost of global transport for large-scale processing.

5. REFERENCES

[1] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, pp. 102783, 2019.

[2] Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional lstm networks for urban change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7651–7668, 2021.

[3] Grace Mugambi, Richard Rimiru, Michael Kimwele, Charles Muriuki, Consolata Gakii, and Esther Mukoya, "Spatial temporal fusion based features for enhanced remote sensing change detection," *Scientific Reports*, vol. 15, no. 1, pp. 1–13, 2025.

[4] Daifeng Peng, Lorenzo Bruzzone, Yongjun Zhang, Haiyan Guan, and Pengfei 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, pp. 102465, 2021.

[5] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15193–15202.

[6] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo 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.

[7] Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang, "Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 228–239, 2022.

[8] Panli Yuan, Qingzhan Zhao, Xingbiao Zhao, Xuewen Wang, Xuefeng Long, and Yuchen 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.

[9] Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[10] Liye Mei, Haonan Yu, Zhaoyi Ye, Chuan Xu, Cheng Lei, and Wei Yang, "Admnet: adaptive deformable convolution large model combining multi-level progressive fusion for building change detection," *Geo-spatial Information Science*, pp. 1–18, 2025.

[11] Sheng Fang, Kaiyu Li, and Zhe Li, "Changer: Feature interaction is what you need for change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[12] Haotian Zhang, Hao Chen, Chenyao Zhou, Keyan Chen, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi, "Bifa: Remote sensing image change detection with bitemporal feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[13] Yingjie Tang, Shou Feng, Chunhui Zhao, Yongqi Chen, Zhiyong Lv, and Weiwei Sun, "A semantic change detection network based on boundary detection and task interaction for high-resolution remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 9, pp. 17184–17198, 2025.

[14] Qi Wang, Wei Jing, Kaichen Chi, and Yuan Yuan, "Cross-difference semantic consistency network for semantic change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[15] Jiang Long, Mengmeng Li, Xiaoqin Wang, and Alfred Stein, "Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 318–335, 2024.