DISCOVER CHANGES GRADUALLY: LLM-GUIDED PROGRESSIVE NETWORK FOR REMOTE SENSING CHANGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing remote sensing change detection (RSCD) method usually adopts one-shot prediction strategy, which is vulnerable to complex background and illumination variation, resulting in missed or false detection especially around ambiguous change boundary. Human visual perception follows a coarse-to-fine analysis process, progressively focusing from salient region to fine-grained detail. Inspired by this process, this paper formulates RSCD as a gradual reasoning process that progressively discovers change region. First, task-oriented question prompt is designed to guide large language model (LLM) to generate high-level semantic prior from bi-temporal image, including change region, boundary characteristic, and potential missed or false detection, etc. Additionally, shared dictionary–coefficient representation learning is utilized to project bi-temporal image into unified representation space, where coefficient-level visual difference information is extracted. Furthermore, text–image consistency is explored to transform global semantic prior into spatial-aware response, thereby adaptively enhancing visual difference feature.Subsequently, Segment Anything Model (SAM) is employed to achieve progressive change discovery strategy. It jointly considers change probability, prediction reliability, and remaining search region to generate multi-stage high-quality prompt, guiding SAM to progressively focus on unresolved change region. Moreover, the soft-union fusion is adopted to obtain final change result. Experiments on multiple datasets demonstrate that the proposed model achieves superior performance and exhibits stronger effectiveness in complex change identification.

1 INTRODUCTION

RSCD aims to identify and localize changed regions by comparing images acquired over the same geographic area at different times. It has been widely applied to urban expansion analysis, disaster assessment, military reconnaissance, vegetation monitoring, and other fields Zang et al. (2025b); Wang et al. (2025c); Zhang et al. (2026b).

Currently, different illumination conditions pose significant challenges to RSCD. Since bi-temporal images are typically acquired at different times or even in different seasons, their imaging conditions are more complex and less controllable than those in natural CD scenes. To tackle illumination disturbance, the existing CD methods can be broadly divided into two categories: feature extraction Wang et al. (2025a); Yu et al. (2026); Wang et al. (2026a) and domain adaptation Ngo et al. (2024); Du et al. (2025); Zhao et al. (2026). CD methods based on feature extraction can broadly be categorized into two approaches: both traditional and deep learningXu et al. (2025). Traditional CD methods can be broadly classified into three categories, e.g., image arithmetic, image transformation, and post-classification, etc. However, these methods tend to be less generalization and robustness. Owing to powerful feature extraction and non-linear representation, deep learning methods have been extensively applied to CD tasks and have demonstrated excellent performanceKim & Kim (2025). Representative deep neural network frameworks for RSCD including Convolutional neural network (CNN)Zang et al. (2025a), TransformerZhang et al. (2026a), MambaParanjape et al. (2025), Denoising diffusion probabilistic model (DDPM)Chaminda Bandara et al. (2025) and their hybrid variants. However, feature extraction-based deep CD networks often exhibit limited generalization

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

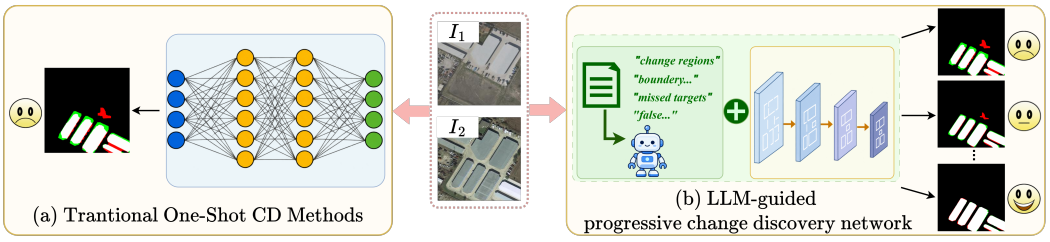


Figure 1: Comparison of different solution strategies for RSCD. (a) Traditional one-shot CD method directly predicts change region from bi-temporal image. (b) Proposed LGPcd integrates LLM-derived semantic prior with progressive change discovery to gradually refine unresolved change region and improve prediction completeness.

in cross-domain scenarios. Although domain adaptation methods improve model generalization to some extent, their performance still degrades substantially under complex domain shifts.

However, most existing RSCD methods usually adopt the one-shot prediction paradigm to directly generate CD results, which struggles to handle challenging change regions such as small objects and ambiguous boundaries. Furthermore, current approaches still rely primarily on visual cues for change discrimination. Thus, incorporating global semantic priors to complement or enhance visual features remains a critical issue for further improving the representation capability of change in complex scenes. Based on above analysis, this paper formulates RSCD as a progressive reasoning process that gradually discovers change from salient region to difficult residual region, and proposes an LLM-guided progressive change discovery network (i.e., LGPcd). It introduces task-related high-level semantic prior through LLM and combines shared dictionary–coefficient representation learning to enhance bi-temporal visual difference representation. Furthermore, a multi-stage residual change-region search mechanism based on SAM is developed to progressively focus on change region that remains insufficiently detected in previous stage.

The main contributions of this paper are summarized as follows:

(1) **A text-guided visual difference feature enhancement method is proposed.** Existing CD method lacks explicit high-level semantic understanding of change region, boundary characteristic, and potential missed or false detection. To introduce such task-related knowledge, task-oriented question prompt is designed to guide LLM in generating semantic prior containing change region, boundary characteristic, and potential missed and false detection region. Moreover, shared dictionary–coefficient representation learning maps bi-temporal image into unified representation space, where coefficient-level visual difference information is extracted. Through text–visual consistency modeling, global semantic prior is transformed into spatial-aware response, enabling adaptive enhancement of visual difference feature.

(2) **Progressive residual refinement enables fine-grained change discovery.** Single-stage CD often struggles to accurately localize change region in complex area. Therefore, a SAM-based multi-stage residual change-region search mechanism is developed to progressively explore unresolved change information. Change probability, prediction uncertainty, and remaining search region are jointly considered to generate high-quality point, box, and mask prompts, guiding SAM to progressively focus on difficult pixel region unresolved in previous stage. Furthermore, input residual prompt and output residual constraint are employed to suppress repetitive prediction across different stages, achieving progressive completion from salient change region to fine-grained residual region.

(3) Extensive experiments on multiple public RSCD dataset demonstrate the effectiveness of integrating language semantic prior, shared representation learning, and progressive change discovery strategy for improving RSCD performance.

2 Related Work

A. Vision-Language Model for RSCD

Relying solely on visual feature makes it difficult to explicitly describe high-level change semantics, such as ”change-region morphology” and ”target likely to be missed or falsely detected”. With

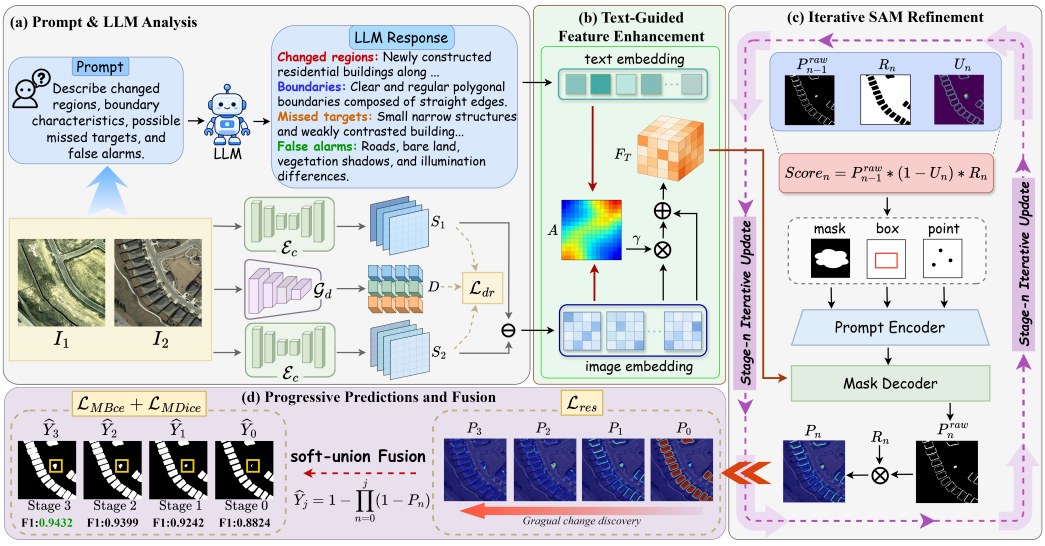


Figure 2: The pipeline of the proposed LGPcd for RSCD. (a) Text prompt and visual feature extraction. On the one hand, a LLM is used to generate word vectors $\boldsymbol{T}_v$ via specific prompts $\boldsymbol{Q}$; on the other hand, visual feature extraction is performed to obtain the shared dictionary $\boldsymbol{D}$ and coefficient representations $\boldsymbol{S}_i$ of bi-temporal images $\boldsymbol{I}_i$. (b) Text-Visual feature interaction. Global word vectors $\boldsymbol{T}_v$ are used to guide local image features $\boldsymbol{F}_\Delta$ for boosted semantic feature $\boldsymbol{F}_T$. (c) High-quality prompts for SAM. For multiple refinement stages, high-quality point, box, and mask prompts are generated by combining change probability $\boldsymbol{P}_{n-1}^{raw}$, prediction uncertainty $\boldsymbol{U}_n$, and residual search regions $\boldsymbol{R}_n$. (d) Progressive change result $\widehat{\boldsymbol{Y}}_n$ generation based on the soft-union fusion strategy.

the rapid development of vision-language pre-trained models, textual information has gradually been introduced into RSCD to provide additional semantic priors for conventional visual difference modeling (Wang et al., 2025b; Dong et al., 2024). ChangeCLIP (Dong et al., 2024) constructs a bi-temporal vision-language CD framework, where text prompts are used to enhance semantic change information. Change-LISA (Jia et al., 2026) performs language-conditioned change reasoning through a multimodal LLM and a SAM encoder. Existing method mainly focuses on category-level semantics or user-defined query condition, while task-specific knowledge required by CD itself remains insufficiently explored. Moreover, how to effectively transfer global textual description to local spatial difference feature remains a key challenge in vision-language RSCD.

B. SAM-Driven RSCD

SAM (Kirillov et al., 2023) establishes a unified promptable segmentation framework consisting of an image encoder, prompt encoder, and mask decoder, while supporting multiple prompt forms, including points, boxes, and masks. However, SAM itself lacks bi-temporal change discrimination capability. Its performance therefore largely depends on prompt quality and on how change-related feature is effectively incorporated into segmentation process. SAM-CD (Ding et al., 2024) adapts SAM to change-region identification from bi-temporal remote-sensing images. Subsequent studies further enhance the change awareness of SAM from the perspectives of change-feature modeling and structural adaptation. ChangeSAM (Jiang et al., 2026) introduces dedicated change-aware modeling to improve the identification of fine-grained change regions and boundaries. However, constructing change-aware prompt and discriminative difference representation remains critical for effectively adapting SAM to RSCD.

3 METHOD

3.1 TEXT-GUIDED VISUAL DIFFERENCE FEATURE ENHANCEMENT

Real changes in bi-temporal RS image are usually accompanied by variations in target structure, boundary morphology, and spatial distribution. However, complex background, illumination variation, and local appearance change may introduce missed detection and false alarm. To address

this issue, this paper introduces LLM to provide high-level semantic prior and adaptively enhance bi-temporal visual difference feature.

Text Prompt Considering the critical semantic information in RSCD, task-oriented question is manually designed to guide LLM to focus on the following aspects:
(1) potential change region;
(2) boundary characteristic of change region;
(3) latent missed detection region;
(4) potential false alarm region.
Given bi-temporal image pair $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, together with predefined question $\boldsymbol{Q}$, LLM generates change-related semantic description for current image pair:

$$\boldsymbol{T}_{desc} = \mathrm{LLM}(\boldsymbol{I}_1, \boldsymbol{I}_2, \boldsymbol{Q}). \tag{1}$$

Subsequently, the generated semantic description is further encoded into a fixed-dimensional text semantic vector $\boldsymbol{T}_v$. Since this description is generated from the complete bi-temporal image pair, it inherently provides global semantic prior rather than pixel-level supervision. This prior describes the change type, boundary pattern, and potential missed detection and false alarm region that deserve attention in current image pair, providing semantic complement for subsequent visual difference representation.

Visual Feature Direct cross-temporal visual difference comparison is easily affected by illumination variation, shadow, and others. Therefore, this paper introduces joint shared dictionary representation learning, which projects bi-temporal image into a unified and constrained representation space. Subsequently, visual feature is obtained by performing change difference in coefficient domain, enabling change reasoning to rely more on response difference within shared representation.

First, the bi-temporal image is concatenated along the channel dimension and fed into the dictionary generation network $\mathcal{G}_d(\cdot)$ to obtain the shared dictionary $\boldsymbol{D} = \mathcal{G}_d([\boldsymbol{I}_1; \boldsymbol{I}_2])$. Then, a parameter-shared coefficient encoder $\mathcal{E}_c(\cdot)$ is employed to separately extract coefficient representation $\boldsymbol{S}_i = \mathcal{E}_c(\boldsymbol{I}_i)$. The shared dictionary describes common structure and local pattern contained in bi-temporal image, while coefficient representation characterizes the response difference of different dictionary atom at each spatial location.

To preserve information completeness, the shared dictionary $\boldsymbol{D}$ and bi-temporal coefficient representations are utilized to reconstruct two temporal images $\widehat{\boldsymbol{I}}_i = \mathcal{R}(\boldsymbol{S}_i, \boldsymbol{D})$, where $\mathcal{R}(\cdot)$ denotes the reconstruction network. In addition, the dictionary representation loss $\mathcal{L}_{dr}$ is designed to improve the stability and consistency of dictionary-coefficient learning by minimizing the reconstruction errors of bi-temporal images together with sparse constraints on the coefficients.

$$\mathcal{L}_{dr} = \min_{\boldsymbol{D}, \boldsymbol{S}_i} \sum_{i=1}^{2} \left(\frac{1}{2} \left\| \boldsymbol{I}_i - \boldsymbol{D} * \boldsymbol{S}_i \right\|_F^2 + \left\| \boldsymbol{S}_i \right\|_1 \right). \tag{2}$$

Within the unified coefficient space, directional bi-temporal coefficient difference $\Delta \boldsymbol{S} = \boldsymbol{S}_2 - \boldsymbol{S}_1$ is further calculated and fed into difference encoding module to obtain visual change feature $\boldsymbol{F}_\Delta$, which serves as visual basis for subsequent semantic enhancement and progressive change-region discovery. Furthermore, the coefficient difference $\Delta \boldsymbol{S}$ is input into change decoder $\mathcal{G}_c(\cdot)$ to generate initial change probability map $\boldsymbol{P}_0 = \sigma(\mathcal{G}_c(\Delta \boldsymbol{S}))$.

Text-Guided Difference Feature Enhancement Text does not provide pixel-level annotation and therefore cannot directly indicate which pixel corresponds to real change. However, it can provide semantic judgment regarding change-related information, such as which region is likely to be falsely detected. Therefore, visual difference feature can be enhanced by modeling semantic consistency between visual feature and text prompt.

Cosine similarity between local visual feature $\boldsymbol{F}_\Delta$and text semantic vector $\boldsymbol{T}_v$ is calculated. and a relu function is applied to suppress negatively correlated response:

$$\boldsymbol{A} = \mathrm{ReLU}\left(\frac{\boldsymbol{F}_\Delta^\top \boldsymbol{T}_v}{\left\| \boldsymbol{F}_\Delta \right\|_2 \left\| \boldsymbol{T}_v \right\|_2} \right), \tag{3}$$

where $\boldsymbol{A}$ denotes the text–visual semantic consistency matrix. Although the input text provides global high-level semantics, position-wise matching with visual difference features at each spatial

location transforms the global semantic prior into spatial-aware responses, thereby characterizing the relevance between different spatial locations and the current textual description.

Finally, residual multiplicative modulation is employed to semantically enhance original visual difference feature:

$$\boldsymbol{F}_T = \boldsymbol{F}_\Delta + \gamma \boldsymbol{A} \odot \boldsymbol{F}_\Delta. \tag{4}$$

where γ is a learnable scaling parameter controlling influence intensity of textual information on visual feature. This formulation preserves original visual difference representation while providing additional enhancement only to spatial location highly consistent with text semantics, thereby preventing excessive interference of high-level textual prior with visual information.

3.2 Progressive Residual Completion for Change Region

SAM is a class-agnostic foundation model for segmentation, and its segmentation accuracy highly depends on input prompt quality. For RSCD, directly generating prompt from high-response region of previous stage may cause subsequent stage to repeatedly focus on change target that has already been reliably detected. Meanwhile, some false-alarm region may also exhibit high change response. To address this issue, a progressive residual prompting mechanism is designed by jointly considering change probability, prediction reliability, and current remaining search region. Accordingly, point, box, and mask prompt are adaptively generated for different stage, guiding subsequent SAM stage to progressively focus on change region that has not been sufficiently detected in previous stage.

For residual refinement stage n, pixel-level uncertainty is first estimated from raw change probability $\boldsymbol{P}_{n-1}^{raw}$ of previous stage using entropy:

$$\widetilde{\boldsymbol{U}}_n = -\boldsymbol{P}_{n-1}^{raw} \log \boldsymbol{P}_{n-1}^{raw}. \tag{5}$$

where $n \in \{1, 2, \ldots, N\}$, $\boldsymbol{P}_0^{raw} = \boldsymbol{P}_0$, and $\boldsymbol{P}_0$ denotes initial change probability map. Subsequently, min–max normalization is applied to obtain uncertainty map of current n-th stage:

$$\boldsymbol{U}_n = \frac{\widetilde{\boldsymbol{U}}_n - min\, \widetilde{\boldsymbol{U}}_n}{max\, \widetilde{\boldsymbol{U}}_n - min\, \widetilde{\boldsymbol{U}}_n + \varepsilon}. \tag{6}$$

where ε is a small constant used to avoid zero denominator. A larger $\boldsymbol{U}_n$ indicates lower prediction reliability at corresponding spatial location.

To prevent subsequent stage from repeatedly processing change region that has already been reliably detected, a progressively updated remaining search region is further maintained. Change region reliably covered at stage j is $\boldsymbol{RC}_j = \mathbb{I}(\boldsymbol{P}_j > \tau_j)$, where τ_j denotes confidence threshold of j-th stage. Then, remaining region $\boldsymbol{R}_n$ available for SAM at n-th stage is uniformly defined as

$$\boldsymbol{R}_n = 1 - \mathrm{clip}\left(\sum_{j=0}^{n-1} \boldsymbol{RC}_j, 0, 1\right). \tag{7}$$

Accordingly, as preceding stage continuously discovers new high-confidence change region, $\boldsymbol{R}_n$ gradually shrinks, allowing subsequent stage to concentrate computation on region that has not yet been sufficiently detected.

However, restricting search region alone is still insufficient to guarantee high reliability of selected prompt point. Therefore, change likelihood, prediction reliability, and residual search range are jointly considered to construct candidate-point scoring function:

$$\boldsymbol{Score}_n = \boldsymbol{P}_{n-1}^{raw} \odot (1 - \boldsymbol{U}_n) \odot \boldsymbol{R}_n. \tag{8}$$

Subsequently, Top-K change anchor with highest score $\boldsymbol{Score}_n$ is selected as positive point prompt for SAM, and corresponding box prompt is generated according to these candidate points. In addition to discrete point and box prompt, response of previous-stage change probability within current remaining region is further utilized to construct continuous residual mask prompt:

$$\boldsymbol{M}_n^{prompt} = \boldsymbol{P}_{n-1}^{raw} \odot \boldsymbol{R}_n. \tag{9}$$

Notably, point prompt, box prompt, and mask prompt play different roles in this process. Point prompt provides reliable change anchor, box prompt specifies coarse spatial range of potential

change target, while mask prompt constrains SAM to focus on current residual search region. These three prompt types are jointly fed into SAM prompt encoder and combined with text-guided enhanced visual difference feature described above, guiding SAM to segment currently unresolved pixel and produce raw change prediction $\boldsymbol{P}_n^{raw}$.

Although residual prompt guides SAM toward remaining region, prediction may still produce response in region already detected by preceding stage. Therefore, after obtaining $\boldsymbol{P}_{n-1}^{raw}$, current remaining region is further used to constrain output:

$$\boldsymbol{P}_n = \boldsymbol{P}_n^{raw} \odot \boldsymbol{R}_n. \tag{10}$$

This operation preserves only valid prediction within current remaining region, thereby giving output of each SAM stage an explicit incremental meaning. Through this process, reliably detected region from preceding stage is continuously removed from subsequent search space, enabling N-th SAM stages to progressively shift from salient change region toward more difficult residual change region. To preserve newly discovered change response from different stage while avoiding excessive numerical amplification caused by direct probability accumulation, a soft-union strategy is adopted to fuse change prediction probability from previous j stages:

$$\widehat{\boldsymbol{Y}}_j = 1 - \prod_{n=0}^{j} \left(1 - \boldsymbol{P}_n\right). \tag{11}$$

3.3 Loss Function

To jointly optimize RSCD, multi-stage residual refinement, and shared dictionary representation learning, this paper designs a joint optimization objective consisting of CD loss, residual supervision loss, and dictionary representation loss.

Let $\boldsymbol{Y}$ denote ground-truth change label, and $\widehat{\boldsymbol{Y}}_n$ denote cumulative change prediction at n-th stage. To constrain prediction of each stage to progressively approach ground-truth change region, $\mathcal{L}_{MBce}$ and $\mathcal{L}_{MDice}$ are adopted to construct multi-stage detection loss:

$$\mathcal{L}_{MBce} = \sum_{n=0}^{N} \alpha_n \, \mathcal{L}_{MBce}\left(\widehat{\boldsymbol{Y}}_n, \boldsymbol{Y}\right), \mathcal{L}_{MDice} = \sum_{n=0}^{N} \alpha_n \, \mathcal{L}_{MDice}\left(\widehat{\boldsymbol{Y}}_n, \boldsymbol{Y}\right). \tag{12}$$

where α_n denotes supervision weight for different stage. In this paper, $\boldsymbol{\alpha} = [0.2, 0.4, 0.7, 1.0]$ is adopted to progressively strengthen optimization effect of later residual stage on final change result.

Since SAM progressively complements change region that remains undetected in previous stage, rather than repeatedly predicting already determined change target, residual-region supervision is further introduced to strengthen complementary relationship among different stage during optimization. For stage n, residual label is constructed according to current remaining region $\boldsymbol{R}_n$:

$$\boldsymbol{Y}_{res}^{n} = \boldsymbol{Y} \odot \boldsymbol{R}_n. \tag{13}$$

Residual label is then used to supervise effective prediction of corresponding stage:

$$\mathcal{L}_{res} = \sum_{n=1}^{N} \mathcal{L}_{Bce}\left(\boldsymbol{P}_n, \boldsymbol{Y}_{res}^{n}\right). \tag{14}$$

Finally, overall training objective is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{MBce} + \lambda_1 \mathcal{L}_{MDice} + \lambda_2 \mathcal{L}_{res} + \lambda_3 \mathcal{L}_{dr}. \tag{15}$$

4 Experimental Results

4.1 Experimental Parameters

The LGPcd model is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. The SGD optimizer is used for model optimization, with momentum at 0.98, weight decay at 5e-4, and an initial learning rate of 0.0005. The learning rate of the optimizer is dynamically adjusted, decaying by 0.7 times every 10 training epochs. Additionally, each dataset undergoes a training period set to 150 epochs, with a batch size of 12. Furthermore, validation is conducted after each training cycle, and the best model on the validation set is used to evaluate the test set.

Table 1: Performance comparison on three benchmark datasets.

Method	LEVIR-CD	SYSU-CD	SECOND
	P / Re / F1 / OA / mIoU	P / Re / F1 / OA / mIoU	P / Re / F1 / OA / mIoU
BIT	92.16/70.68/80.01/98.20/82.40	80.52/72.55/76.16/89.33/74.32	69.42/64.37/66.80/**88.72**/68.71
ICIFNet	76.65/81.25/78.88/98.30/81.69	84.60/70.70/77.03/90.06/75.36	**78.21**/53.96/63.86/87.79/66.61
USSFC-Net	**94.85**/73.98/83.13/98.47/84.77	81.66/68.24/74.35/88.90/72.97	73.15/54.43/62.48/86.90/65.37
DMINet	79.23/77.61/78.41/98.33/81.39	86.07/71.13/77.89/90.48/76.17	74.18/58.50/65.41/87.63/67.30
ChangeMamba	94.16/75.81/84.01/98.53/85.44	79.55/73.91/76.63/89.37/74.62	69.29/63.19/66.10/87.04/67.26
STRobustNet	77.06/85.01/80.84/98.43/83.11	82.41/74.42/78.21/90.22/76.18	69.74/62.13/65.71/86.95/67.01
ConvFormer	76.98/85.99/81.24/98.45/83.40	**86.22**/73.49/79.35/**90.98**/77.43	69.03/63.28/66.03/87.01/67.20
FCdDNet	92.05/87.13/89.52/**98.96**/89.97	74.96/77.66/76.28/88.61/73.86	68.42/69.01/68.72/87.33/68.88
Ours	89.80/**89.58**/**89.69**/98.95/**90.11**	79.93/**82.40**/**81.15**/90.97/**78.53**	67.41/**70.70**/**69.02**/87.30/**68.95**

All values are reported in percentage (%). Color convention: **best**, 2-nd best

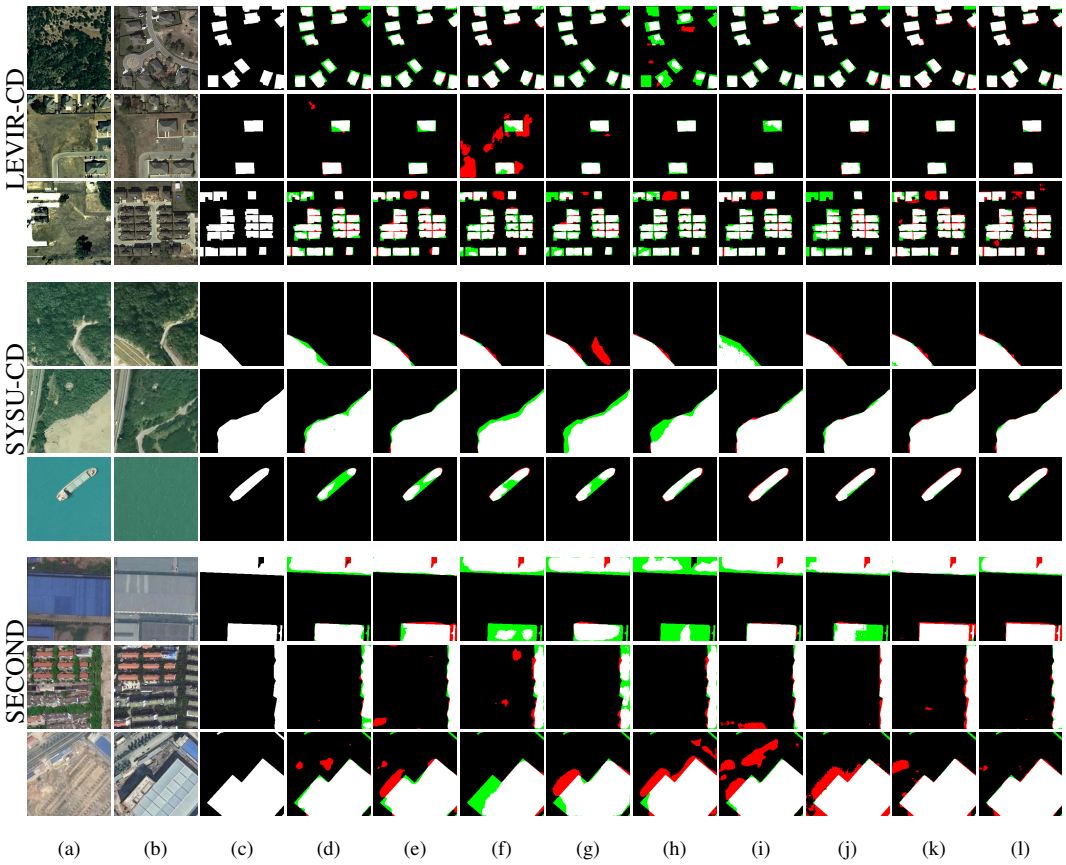


Figure 3: Qualitative results of different CD methods on LEVIR-CD, SYSU-CD and SECOND datasets. Columns: (a) $\boldsymbol{I}_1$ image; (b) $\boldsymbol{I}_2$ image; (c) Ground Truth; (d) BIT; (e) ICIFNet; (f) USSFC-Net; (g) DMINet; (h) ChangeMamba; (i) STRobustNet; (j) ConvFormer; (k) FCdDNet; (l) ours.

4.2 Performance Comparison

1) Objective Indicator To validate the effectiveness of LGPcd, we conduct a comprehensive comparison with several representative CD models, including BIT (Chen et al., 2022), ICIFNet (Feng et al., 2022), USSFC-Net (Lei et al., 2023), DMINet (Feng et al., 2023), ChangeMamba (Chen et al., 2024), STRobustNet (Zhang et al., 2025), ConvFormer (Lin et al., 2023), and FCdDNet (Wang et al., 2026b). To ensure a fair comparison, all competing methods are implemented using source code released by respective authors, with hyperparameters configured according to original publications.

Table 1 presents quantitative metrics of different models on three datasets. The proposed model achieves outstanding performance in comprehensive indicators such as F1-score, OA, and mIoU

across multiple datasets. For instance, it attains an F1-score of 89.69 on the LEVIR-CD dataset, which may benefit from the integration of language semantic prior, shared representation learning, and progressive residual segmentation, enabling more accurate localization of change regions.

2) Visual Comparison As shown in Fig. 3, competing methods suffer from missed detections on small targets and boundary errors on large changed regions. LGPcd produces cleaner change maps with fewer false alarms and missed detections.

4.3 Ablation Experiments

1) Parameter Analysis To validate the efficacy of each loss function and determine their optimal configurations, we experimentally tune regularization terms on SYSU-CD dataset. As shown in Table 2, the optimal parameter combination for $\boldsymbol{\lambda}$ is [1, 0.5, 0.5, 0.1]. Subsequently, we extend this configuration to the other two datasets, where it continues to yield favorable quantitative results as shown in Table 1. This demonstrates our LGPcd robust generalization and indicates that it exhibits little over-sensitivity to the regularization coefficients.

Table 2: Ablation Experiment of Regular Term Coefficient

λ_1	λ_2	λ_3	P	Re	F1	OA	mIoU
0.5	1	0.1	79.10	82.31	80.67	90.70	78.03
0.1	1	0.1	77.47	83.33	80.30	90.35	77.54
1	1	0.1	76.43	**84.63**	80.32	90.22	77.45
0.5	0.5	0.1	79.93	82.40	**81.15**	**90.97**	**78.53**
0.5	0.1	0.1	**81.69**	79.03	80.34	90.88	77.96
0.5	0.5	0.5	81.43	76.39	78.83	90.32	76.62
0.5	0.5	1	77.91	80.91	79.38	90.09	76.78

2) The Ablation for Progressive Stage Number Table 3 evaluates the effect of progressive stage number N. Performance gradually improves as N increases from 1 to 3, with the best F1, OA, and mIoU achieved at $N = 3$. When N further increases to 4, performance slightly decreases, indicating that excessive refinement stages may introduce redundant information.

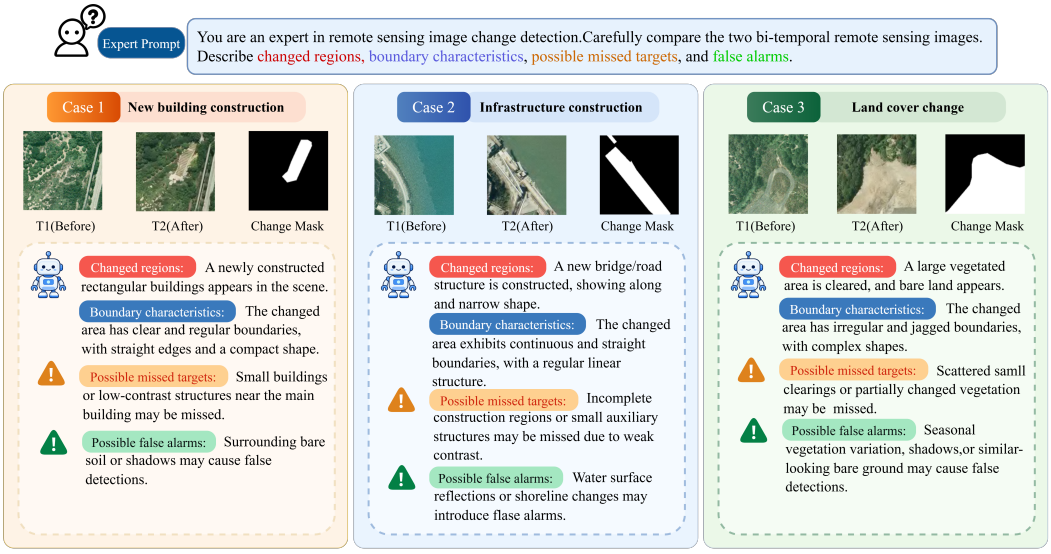


Figure 4: Examples of LLM-generated task-oriented semantic descriptions, including changed regions, boundary characteristics, missed targets, and false alarms.

4.4 Qualitative Analysis of LLM-Generated Semantic Prompt

To evaluate the effectiveness of task-oriented question prompting, the semantic descriptions generated by LLM under different change scenarios are visualized in Fig. 4. The generated descriptions contain change regions, boundary characteristics, potential missed targets, and false-alarm regions.

For different scenes, LLM adapts the semantic descriptions according to image content, capturing structural characteristics and potential interference factors.

These results demonstrate that task-oriented prompting enables LLM to generate change-related semantic priors rather than generic scene descriptions, providing complementary information for subsequent text–visual consistency modeling.

4.5 VISUALIZATION

1) Visualization for pixel classification To analyze pixel-level feature discrimination, T-SNE is employed to visualize changed and unchanged features. LGPcd achieves clearer feature separation with less confusion than other methods, leading to sharper semantic boundaries and consistent improvement in quantitative results.

Table 3: Ablation on the number of progressive stages

N	P	Re	F1	OA	mIoU
2	78.06	**84.36**	81.09	90.72	78.30
3	**79.93**	82.40	**81.15**	**90.97**	**78.53**
4	79.59	80.75	80.17	90.58	77.63

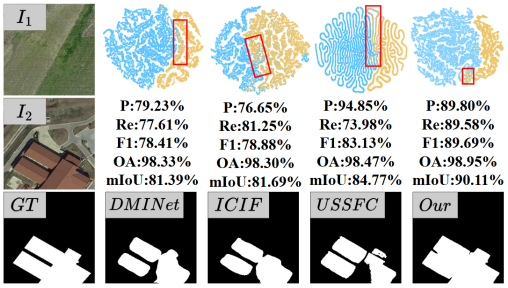


Figure 5: Pixel-level classification visualization of different models.

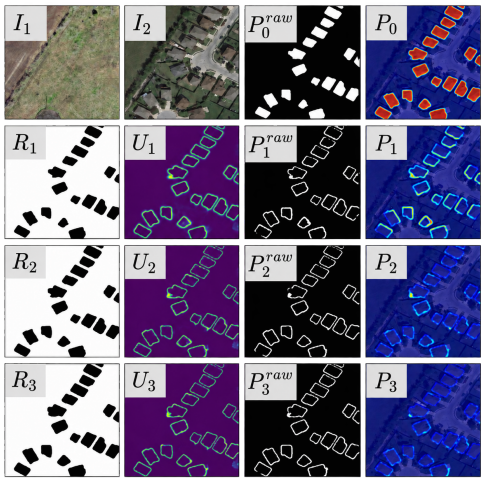


Figure 6: Visualization of the progressive change discovery process. $\boldsymbol{R}_n$, $\boldsymbol{U}_n$, $\boldsymbol{P}_{n-1}^{raw}$, and $\boldsymbol{P}_n$ denote the remaining search region, uncertainty map, raw prediction, and residual-constrained prediction at stage n, respectively.

2) Visualization for Progressive Change Discovery To further analyze the progressive change discovery process, intermediate results of different stages are visualized in Fig. 6. The initial prediction $\boldsymbol{P}_0$ first identifies salient change regions, while subsequent stages progressively explore unresolved regions guided by the remaining search region and uncertainty information. With residual refinement, redundant responses are suppressed and missed regions are gradually complemented. Finally, the soft-union fusion integrates multi-stage predictions to obtain a more complete change result.

5 CONCLUSION

This paper addresses the limitations of existing RSCD methods that rely heavily on one-shot prediction, have difficulty fully exploiting high-level semantic information, and are vulnerable to complex background, illumination variation, and boundary uncertainty, which may lead to missed detection and false alarm. To overcome these issues, an LLM-guided progressive residual change detection network is proposed. The proposed method reformulates RSCD as an easy-to-hard progressive reasoning process. Specifically, LLM is employed to generate change-related semantic prior, while shared dictionary–coefficient representation learning is further introduced to enhance the representation capability of bi-temporal difference feature. On this basis, a progressive residual-guided strategy drives SAM to gradually focus on change region that remains insufficiently determined in previous stage, enabling stepwise completion and refinement of complex change target. Experimental results demonstrate that the proposed method achieves superior detection performance compared with existing state-of-the-art methods, validating the effectiveness of integrating high-level semantic prior, shared representation learning, and progressive residual reasoning for improving RSCD performance.

AI USE STATEMENT

The authors did not use generative AI tools in the preparation of this manuscript.

REFERENCES

Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M. Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5250–5262, 2025. doi: 10.1109/WACV61041.2025.00513.

H. Chen, Z. Qi, and Z. Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. doi: 10.1109/TGRS.2024.3368168.

Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. doi: 10.1016/j.isprsjprs.2024.01.004.

Yuheng Du, Sheng Yang, Lingxuan Wang, Zhenghua Hou, Chengying Cai, Zhitao Tan, Mingxia Chen, Shi-Sheng Huang, and Qiang Li. Rtmap: Real-time recursive mapping with change detection and localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28021–28030, 2025. doi: 10.1109/ICCV51701.2025.02601.

Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng. Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

Y. Feng, J. Jiang, H. Xu, and J. Zheng. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

Xiangyu Jia, Zhibo Chen, Shengyi Zhang, and Xiaojing Xue. Change-lisa: Language-guided reasoning for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026. doi: 10.1109/TGRS.2026.3684817.

Kaixuan Jiang, Chen Wu, Zhenghui Zhao, Bo Du, and Liangpei Zhang. From segmentation to change: Releasing segment anything model for remote sensing change detection. *Pattern Recognition*, 180:114605, 2026. doi: 10.1016/j.patcog.2026.114605.

Jae-Woo Kim and Ue-Hwan Kim. Towards generalizable scene change detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24463–24473, 2025. doi: 10.1109/CVPR52734.2025.02278.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

T. Lei, X. Geng, H. Ning, Z. Lv, M. Gong, Y. Jin, and A. K. Nandi. Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

X. Lin, Z. Yan, X. Deng, C. Zheng, and L. Yu. Convformer: Plug-and-play cnn-style transformers for improving medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 642–651. Springer Nature Switzerland, 2023.

Ba Hung Ngo, Nhat-Tuong Do-Tran, Tuan-Ngoc Nguyen, Hae-Gon Jeon, and Tae Jong Choi. Learning cnn on vit: A hybrid model to explicitly class-specific boundaries for domain adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28545–28554, 2024. doi: 10.1109/CVPR52733.2024.02697.

Jay N. Paranjape, Celso De Melo, and Vishal M. Patel. A mamba-based siamese network for remote sensing change detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1186–1196, 2025. doi: 10.1109/WACV61041.2025.00123.

Bin Wang, Xiaohu Jiang, Pinle Qin, and Jianchao Zeng. Dsfdcd: Joint distribution sampling and feature decoupling deep network for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–12, 2025a. doi: 10.1109/TGRS.2025.3585229.

Bin Wang, Sichao Qiao, Ziyv Zhao, Pinle Qin, and Jianchao Zeng. Uafdcd: Uncertainty-aware feature decoupling network for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5625815–5625815, 2026a. doi: 10.1109/TGRS.2026.3701174.

Bin Wang, Bo Wang, Pinle Qin, and Jianchao Zeng. Fcddnet: Feature cross-domain decoupling network for remote sensing change detection. *Pattern Recognition*, 177:113259, 2026b. doi: 10.1016/j.patcog.2026.113259.

J. Wang, F. Liu, L. Jiao, H. Wang, S. Li, L. Li, P. Chen, X. Liu, and W. Ma. Change knowledge-guided vision-language remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025b.

Zhipan Wang, Zijun Yang, and Qingling Zhang. Ia-cdnet: Change detection in adverse remote sensing image conditions with an advanced image-adaptive method. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025c. doi: 10.1109/TGRS.2024.3518568.

Dazhi Xu, Ming Li, Yan Wu, Peng Zhang, and Xinyue Xin. Statistic-guided difference enhancement graph transformer for unsupervised change detection in polsar images. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):11667–11684, Nov 2025. ISSN 1558-2205. doi: 10.1109/TCSVT.2025.3575082.

Ruteng Yu, Lukun Wang, and Jiaming Pei. Gsag-cdgan: A generalizable small-sample attention-guided gan for remote sensing change detection (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):41452–41454, Mar. 2026. doi: 10.1609/aaai.v40i48.42303. URL https://ojs.aaai.org/index.php/AAAI/article/view/42303.

Qi Zang, Dong Zhao, Shuang Wang, Dou Quan, and Zhun Zhong. Feature spectrum learning for remote sensing change detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12647–12657, 2025a. doi: 10.1109/CVPR52734.2025.01180.

Qi Zang, Dong Zhao, Shuang Wang, Dou Quan, and Zhun Zhong. Feature spectrum learning for remote sensing change detection. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12647–12657, 2025b. URL https://api.semanticscholar.org/CorpusID:280148271.

H. Zhang, Y. Teng, H. Li, and Z. Wang. Strobustnet: Efficient change detection via spatial–temporal robust representations in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Mingwei Zhang, Jingtao Hu, Qiang Li, and Qi Wang. Exploring generalizable remote sensing change detection via low-rank exchange adaptation of vision foundation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12663–12671, 2026a. doi: 10.1609/aaai.v40i15.38262.

Mingwei Zhang, Jingtao Hu, Qiang Li, and Qi Wang. Exploring generalizable remote sensing change detection via low-rank exchange adaptation of vision foundation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12663–12671, Mar. 2026b. doi: 10.1609/aaai.v40i15.38262. URL https://ojs.aaai.org/index.php/AAAI/article/view/38262.

Jiaqi Zhao, Jianpeng Xie, Yong Zhou, Wen-Liang Du, Hancheng Zhu, and Rui Yao. Causal decoupling domain generalization for remote sensing change detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):13135–13143, Mar. 2026. doi: 10.1609/aaai.v40i15.38314. URL https://ojs.aaai.org/index.php/AAAI/article/view/38314.