Exploring Generalizable Remote Sensing Change Detection via Low-Rank Exchange Adaptation of Vision Foundation Model

Mingwei Zhang[1, 2], **Jingtao Hu**[2], **Qiang Li**[2], **Qi Wang**[2*]

[1]School of Computer Science, Northwestern Polytechnical University
[2]School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University
{dlaizmw, liqmges, crabwq}@gmali.com, jthu@mail.nwpu.edu.cn

Abstract

Remote sensing change detection (CD) has achieved remarkable progress in recent years. However, little attention has been paid to generalizable change detection (GCD) methods that can effectively generalize to unseen scenarios or domains beyond the training distribution. The major challenges in GCD arise from domain diversity and bitemporal domain shifts in remote sensing images, caused by variations in imaging platforms, acquisition times, geographic regions, and observed events. To tackle these challenges, we propose GenCD, a GCD framework built upon vision foundation models (VFMs). Specifically, GenCD introduces two key components: (1) a Low-Rank Exchange Adaptation (LREA) strategy of VFMs that aligns bitemporal representations while preserving the generalization capacity of VFMs on single-temporal inputs; and (2) a Token-Guided Feature Refinement (TGFR) mechanism that leverages an input-independent token as a guide to refine difference features, improving the discrimination between changed and unchanged regions. We conduct extensive cross-dataset evaluations on eight diverse datasets across three binary CD tasks: land cover, land use, and building-only CD. The results consistently demonstrate the superior generalization of GenCD over SoTA methods, highlighting its effectiveness in GCD.

Code — https://github.com/ptdoge/LREA.

Introduction

Remote sensing Change Detection (CD) aims to monitor dynamic changes on the surface of the Earth by analyzing bitemporal images acquired at different times over the same geographic region. This task has wide application value in areas such as disaster response (Zheng et al. 2021), ecological protection (Willis 2015), and land resource management (Kennedy et al. 2009). It has achieved significant progress in recent years. However, Generalizable Change Detection (GCD) models remain limited. Following the generalizable dense prediction task setting, GCD targets detecting changes in unseen domains not present in the training data. Undoubtedly, GCD is crucial for real-world CD applications.

Remote sensing images, as illustrated in Figure 1, often exhibit substantial variations driven by multiple factors. Specifically, distinct imaging platforms typically lead to variations in spatial resolution. Acquisition times affect illumination conditions and seasonal characteristics. Geographic locations influence the spectral and structural properties of identical object classes. Additionally, external events give rise to diverse change patterns. These heterogeneous factors, spanning *spatial*, *temporal*, and *event* dimensions, collectively pose significant challenges for GCD. In existing studies, most are tailored for closed datasets, struggling to generalize to unseen scenarios. Although domain adaptive CD methods (Zhang et al. 2024) can mitigate this issue, they require access to target domain data during training, making them less practical for plug-and-play applications. Recently, Vision Foundation Models (VFMs) trained on large-scale datasets demonstrate strong representational generalizability (Awais et al. 2025). Meanwhile, Parameter-Efficient Fine-Tuning (PEFT) techniques have been shown to effectively unlock the potential of VFMs for various downstream tasks while preserving their superior generalization (Chen et al. 2025; Peng et al. 2025). For example, in the context of Domain-Generalized Semantic Segmentation (DGSS), effective fine-tuning of VFMs has significantly outperformed the approaches based on domain randomization or domain perturbation (Wei et al. 2024; Tang et al. 2025).

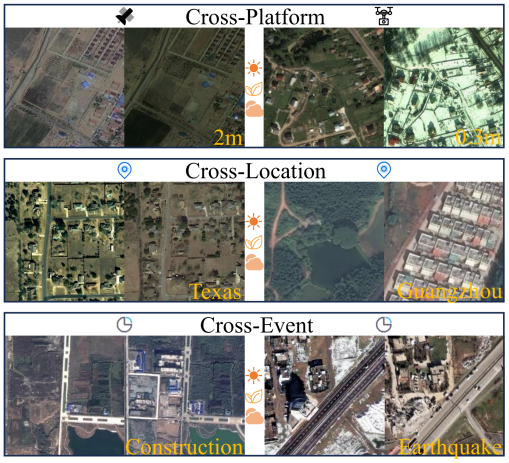

Figure 1: Illustration of domain diversities in remote sensing images across spatial, temporal, and event dimensions.

*Corresponding author.

Motivated by the above successes, we propose **GenCD**. GenCD explores two key issues: **(1) How to fine-tune VFMs towards GCD?** As discussed above, the core challenge in GCD lies in the diverse spatial, temporal, and event characteristics of remote sensing images, with uncertain bitemporal domain shifts further hindering generalization. To this end, we propose a Low-Rank Exchange Adaptation (LREA) strategy that equips VFMs with dual adaptation capabilities: one branch focuses on task-oriented tuning, while the other aligns bitemporal features through low-rank representation cross-temporal exchange adaptation. LREA introduces only a small number of trainable parameters and keeps the original VFM weights frozen, thereby preserving its strong generalization ability while enabling efficient task-specific and temporal alignment adaptation. **(2) How to generate discriminative differences in unseen scenarios with unfamiliar or novel changes?** For example, under the cross-platform setting, higher-resolution images in the test domain may reveal previously indistinguishable objects, such as vehicles, which are then annotated as changes. As such fine-grained patterns are unseen during training, the model may fail to detect them. To this end, we introduce a Token-Guided Feature Refinement (TGFR) mechanism to enhance the discrimination between changed and unchanged regions across different scenarios. In TGFR, an input-independent token without direct supervision acts as a global query to aggregate difference cues and refine local features, analogous to the [CLS] token in CLIP (Radford et al. 2021) but with a distinct role. Its independence from specific scenes makes it effective for improving change recognition in unseen contexts.

Beyond the proposed methods, we conduct comprehensive experiments within GenCD to explore the impact of different difference extraction and decoding schemes on performance, thereby establishing a solid benchmark to facilitate future research. In summary, the main contributions made by this work are as follows:

- We propose GenCD, a GCD framework built upon VFMs, alleviating the challenge of generalizing to unseen domains in remote sensing CD.
- We develop a LREA strategy to fine-tune VFMs towards GCD. It introduces dual adaptation branches for task-oriented tuning and bitemporal alignment, while preserving the generalization of VFMs.
- We present a TGFR mechanism that leverages an input-independent token to aggregate difference cues and refine local features, enhancing the discrimination between unchanged and changed regions while facilitating change recognition in unseen contexts.
- We conduct extensive cross-dataset evaluations on eight benchmarks across three binary CD tasks including land cover, land use, and the building-only, demonstrating the generalizability of GenCD.

Related Work

Remote Sensing Change Detection

CD is a highly popular topic in the remote sensing community due to its broad application value. Existing studies typically focus on developing high-performance models for specific, closed scenarios, where pseudo changes caused by bitemporal domain gaps pose a well-recognized challenge (Zhang et al. 2023). To address this issue, feature interaction alignment and image style diversification have emerged as primary approaches, such as FeaSpect (Zang et al. 2025b) and Changer (Fang, Li, and Li 2023). Although these methods are not specifically designed for GCD, they have provided valuable insights for advancing research in this direction. More recently, DonaNet has investigated the integration of domain generalization techniques into CD model design to enhance generalization in unseen domains (Zang et al. 2025a). Overall, considerable room for further GCD exploration still remains.

Domain Generalized Semantic Segmentation

DGSS aims to train models that perform well on unseen target domains differing from the source domains used during training (Pak et al. 2024). In earlier studies, most methods focused on learning domain-invariant and domain-specific features through techniques such as feature whitening or normalization, achieving promising results, e.g., IBN (Pan et al. 2018), RobustNet (Choi et al. 2021), and SAN-SAW (Peng et al. 2022). Meanwhile, data augmentation strategies emphasizing style diversity have also proven effective (Jia et al. 2024; Zhong et al. 2022). However, in the era of foundation models, these traditional methods appear outdated. Recently, a growing number of studies have explored the use of VFMs combined with dedicated PEFT strategies for DGSS, achieving superior performance over previous approaches (Wei et al. 2024). These advances motivate us to explore GCD within the VFM paradigm.

Methodology

Overall Framework

In this work, we propose GenCD, a GCD framework designed to advance plug-and-play deployment of CD models in real-world applications. The pipeline of GenCD is illustrated in Figure 2(a)-(d), which consists of four parts:

Bitemporal Feature Extraction. Bitemporal features are extracted from the input image pair $\{I_{t_1}, I_{t_2}\}$ using a Siamese encoder built upon an off-the-shelf VFM, such as DINOv2 (Oquab et al. 2023), SAM (Kirillov et al. 2023), or CLIP (Radford et al. 2021). The strong representation generalizability of VFMs make them well-suited for GCD. However, independently encoding each time point fails to eliminate the domain gap between bitemporal images, an acknowledged devil that often compromises CD accuracy. Therefore, we inject the proposed LREA into the frozen VFM. The resulting encoder, denoted as $\mathcal{E}_*$, is capable of generating task-relevant and temporally aligned features:

$$\{F_{t_1,i}, F_{t_2,i}\}_{i=1}^{4} = \mathcal{E}_*(I_{t_1}, I_{t_2}), \tag{1}$$

where $F_{t_k,i}$ denotes the feature at the i-th stage for time t_k.

Difference Extraction. Feature maps $F_{t_k,i}$ from four different stages of the VFM encoder are retained and rescaled for multi-scale difference extraction. We explore two typical

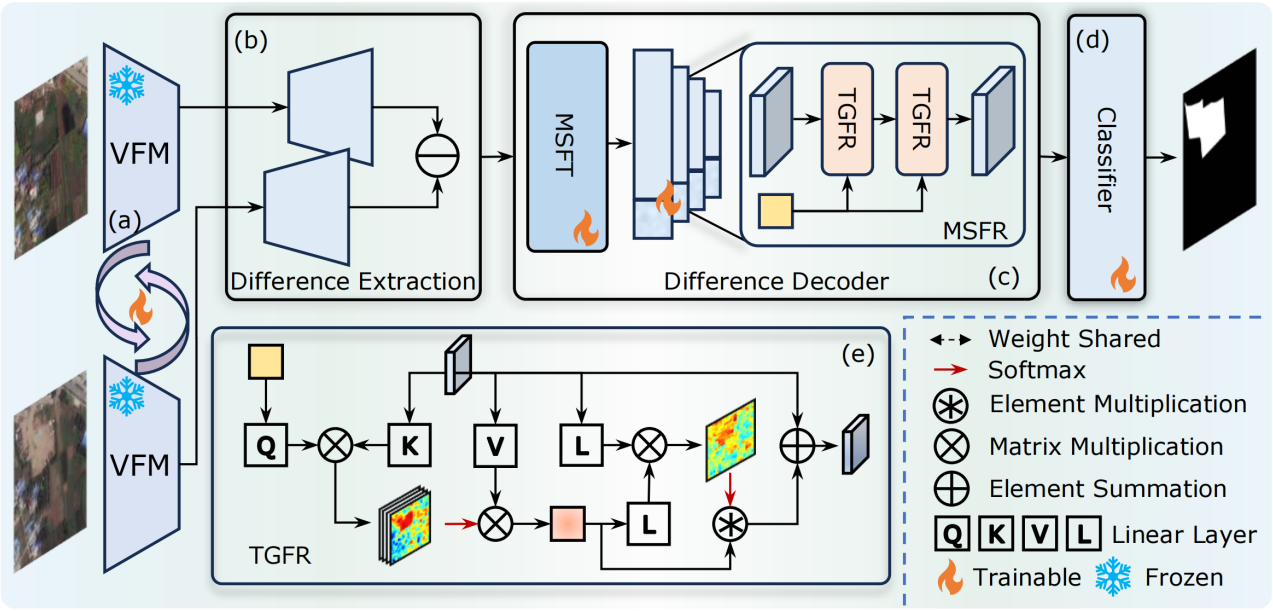

Figure 2: Overview of the GenCD framework: (a) VFM-based bitemporal feature extractor with LREA; (b) difference extraction; (c) difference decoder with Multi-Scale Feature Transformation (MSFT) and Multi-Scale Feature Refinement (MSFR); (d) final classifier; (e) the structure diagram of TGFR.

ways: (i) feature concatenation:

$$D_i = [F_{t_1,i}; F_{t_2,i}], \tag{2}$$

and (ii) absolute difference:

$$D_i = |F_{t_1,i} - F_{t_2,i}|. \tag{3}$$

Here, feature concatenation $[;]$ stacks the features from both time points, while absolute difference computes the element-wise difference magnitude. Thereby, bitemporal differences with four different scales are constructed.

Difference Decoder. In GenCD, the difference decoder consists of two steps: feature transformation and refinement. For the transformation step, two representative decoding heads are explored to process bitemporal differences: (i) applying a vanilla convolution operator (VCO) at each scale:

$$\hat{D}_i = \text{Conv}_{1\times1}(D_i), \tag{4}$$

and (ii) using a feature pyramid network (FPN) to fuse multi-scale features:

$$\{\hat{D}_i\}_{i=1}^{4} = \text{FPN}(\{D_i\}_{i=1}^{4}). \tag{5}$$

Then, the difference features $\{\hat{D}_i\}_{i=1}^{4}$ are refined via the TGFR mechanism to enhance the distinction between changed and unchanged regions:

$$\tilde{D}_i = \text{TGFR}_{\times K,i}(\hat{D}_i). \tag{6}$$

where K indicates the number of times TGFR is applied, and the $\tilde{D}_i$ denotes the final output of the decoder at scale i.

Classifier. Finally, the refined difference features at the four different scales are concatenated:

$$D_o = [\tilde{D}_1; \tilde{D}_2; \tilde{D}_3; \tilde{D}_4], \tag{7}$$

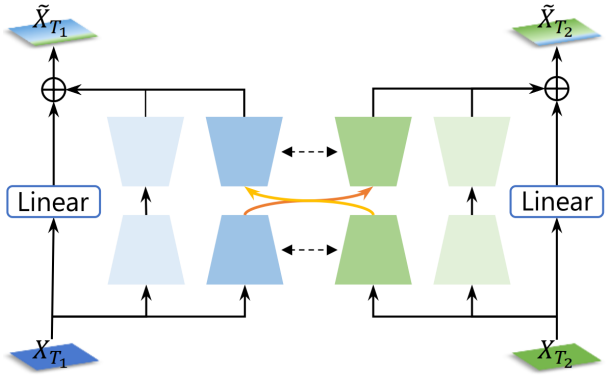

Figure 3: The structure diagram of LREA.

and further passed through a two-layer convolutional module to predict the probability maps of changed and unchanged regions:

$$P = \text{softmax}(\text{Conv}_{1\times1}(\text{Conv}_{1\times1}(D_o))). \tag{8}$$

Loss Function. The GenCD is trained using the cross-entropy loss function, which supervises the predicted change map against the ground-truth (GT) binary labels.

Low Rank Exchange Adaptation

Before introducing LREA, we first review LoRA (Hu et al. 2022), a widely adopted method for fine-tuning pretrained models. LoRA introduces low-rank matrices $\Delta W \in \mathbb{R}^{n\times m}$ to update model weights during fine-tuning, motivated by the observation that deep neural networks tend to optimize within a subspace whose dimensionality is much smaller

than the full parameter space. The updated weights W' at LoRA-injected layers are formulated as:

$$W' = W + \Delta W = W + sBA, \tag{9}$$

where W denotes the original pretrained weights, and ΔW is the update term, expressed as the product of two low-rank matrices $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times m}$, with $r \ll \min(n, m)$. By optimizing only A and B while keeping W frozen, LoRA enables efficient task-specific adaptation.

Recent work, VFMSeg (Tang et al. 2025) demonstrates that directly applying LoRA to VFMs is effective for single-input dense prediction tasks under domain generalization settings. Motivated by this insight, we adopt LoRA to perform task-oriented tuning of the VFM, adapting it effectively for CD. During this process, the pretrained weights remain frozen, thereby preserving the inherent representation generalizability of the VFM. However, LoRA alone is insufficient to handle the uncertain domain shifts between bitemporal images, a well-known challenge that significantly hampers the generalization and accuracy of CD. To address this limitation, the Low-Rank Exchange Adaptation (LREA) is developed. As illustrated in Figure 3, LREA consists of two coordinated branches: one mirrors the original LoRA structure, while the other facilitates exchange adaptation. We formalize the operations of both branches as follows:

$$\begin{aligned} h_{t_1} &= (W + sB_{t_1,1}A_{t_1,1})\, x_{t_1} + (sB_{t_1,2}A_{t_2,2})\, x_{t_2}, \\ h_{t_2} &= (W + sB_{t_2,1}A_{t_2,1})\, x_{t_2} + (sB_{t_2,2}A_{t_1,2})\, x_{t_1}, \end{aligned} \tag{10}$$

where low-rank matrices $B_{t_k,j} \in \mathbb{R}^{n\times r}$ and $A_{t_k,j} \in \mathbb{R}^{r\times m}$, with $r \ll \min(n, m)$, are weight-shared for bitemporal inputs. Intuitively, the exchange branch aims to learn domain-related representations from the pretrained features via low-rank subspace encoding, such as illumination, texture, or seasonal variations that are irrelevant to actual scene changes, thereby alleviating temporal domain discrepancies through cross-temporal modulation.

Token-Guided Feature Refinement

In the context of GCD, although the adapted VFM produces aligned and generalized representations, the resulting difference features still exhibit inherent domain discrepancies due to persistent variations in geography, sensors, and event-specific conditions. This necessitates a dedicated refinement mechanism to enhance the discriminability of difference features, particularly in unseen scenarios.

To address this, we introduce TGFR, as illustrated in Figure 2(e). Given a difference feature map $\hat{D}_i \in \mathbb{R}^{H\times W\times C}$, we first reshape it into a sequence of spatial tokens $\hat{D}_{i,s} \in \mathbb{R}^{N\times C}$. A learnable and input-independent token $\mathbf{q} \in \mathbb{R}^{1\times C}$ then serves as the query to extract global difference cues $\mathbf{z} \in \mathbb{R}^{1\times C}$ via multi-head cross-attention, where $\hat{D}_{i,s}$ serves as both key and value. $\mathbf{z}$ captures a global, scene-invariant representation that may reflect either changed or unchanged regions, depending on the patterns learned by the token $\mathbf{q}$.

To exploit this representation for refinement, we first compute its affinity with the spatial tokens:

$$\mathbf{A} = (\hat{D}_{i,s}W_f)(\mathbf{z}W_z)^\top, \tag{11}$$

where $W_f, W_z \in \mathbb{R}^{C\times d}$ are projection matrices. The refined features are then obtained as:

$$\tilde{D}_{i,s} = \hat{D}_{i,s} + \text{softmax}(\mathbf{A}/\sqrt{d}) \circledast \mathbf{z}, \tag{12}$$

where $\circledast$ denotes element-wise multiplication between the attention weights and $\mathbf{z}$, broadcast along the spatial dimension. The refined sequence $\tilde{D}_{i,s}$ is reshaped back to the original spatial layout $\tilde{D}_i \in \mathbb{R}^{H\times W\times C}$.

The Principle of TGFR. In the training dataset, change regions often cover multiple object categories or exhibit intra-class variation. TGFR leverages a single, input-agnostic token across distinct samples to aggregate difference cues, enabling it to learn a generalized representation that captures either changed or unchanged patterns. Thereby, with the effect of the token, scene-invariant difference cues can be extracted. Meanwhile, TGFR effectively utilizes them to improve the discriminability of difference features, thus facilitating the change recognition in novel environments.

Experiments

Datasets and Evaluation Metric

Datasets. Following the standard DGSS setting, the GCD is validated through cross-dataset generalization, where a model trained on one dataset is evaluated on another. Accordingly, we assess the proposed method via cross-dataset testing across eight datasets spanning three binary CD tasks: land cover, land use, and building-only CD. For land cover CD, three datasets with precise annotations are used: JL1-CD (JL) (Liu et al. 2025), SVCD (SV) (Lebedev et al. 2018), and SECOND (SE) (Yang et al. 2021), where JL1-CD serves as the training set, and SVCD and SECOND are utilized for testing. For the land use CD task, the FUSU dataset (Yuan et al. 2024), with images collected from two distinct locations, Xi'an (FX) and Jiaxing (FJ), is used. In the building-only CD experiments, four datasets are employed: LEVIR-CD+ (LE) (Chen and Shi 2020), WHU-CD (WH) (Ji, Wei, and Lu 2018), GZ-CD (GZ) (Peng et al. 2021), and Türkiye-CD (TK) (Shen et al. 2024), where LEVIR-CD+ is used for training, and others for testing.

Evaluation Metric. We adopt mean Intersection over Union (mIoU), mean F1-score (mF1), and overall accuracy (Acc) as evaluation metrics to assess model performance across both changed and unchanged categories comprehensively.

Implementing Details

The proposed GenCD is implemented using PyTorch and the MMSegmentation framework. During training, we adopt the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay of 0.05 to fine-tune the model parameters. Meanwhile, a polynomial learning rate scheduler with a power of 0.9 is applied throughout the training process. To facilitate efficient training, the total number of iterations is set to 10k, with a batch size of 4 and input images cropped to 512×512. These settings allow the model to be trained on a single RTX 3090 GPU. Besides, data augmentation includes random brightness and contrast adjustments, as well as horizontal and vertical flipping. The last checkpoint is

Method	Backbone	Trainable Params (M)	JL2SV	JL2SE	FJ2FX	FX2FJ	LE2TK	LE2WH	LE2GZ
ChangeFormer	MiT-b0	3.85	49.06	58.13	52.82	48.88	50.89	83.86	59.46
Changer	MiT-b0	3.46	55.80	57.49	50.15	47.75	50.55	80.39	52.16
TinyCDv2	N/A	0.12	55.41	55.53	47.61	47.37	49.26	76.48	52.55
MambaBCD	VMamba-S	54.00	48.45	59.25	54.03	28.62	51.42	84.22	64.25
Change3D	X3D-L	7.01	60.36	57.88	49.17	36.96	48.73	85.41	56.86
TTP	SAM-L	6.21	53.64	60.56	51.98	43.15	51.54	84.22	55.60
BAN-CF$_1$	CLIP-L	4.47	51.11	59.11	53.14	52.16	53.92	84.99	<u>75.63</u>
BAN-BiT	CLIP-L	3.81	52.39	60.46	52.38	51.30	52.63	87.17	73.47
BAN-CF$_2$	RemoteCLIP-L	4.47	52.92	59.21	50.49	51.63	52.88	82.33	73.54
Rein	DINOv2-L	327.77	55.22	58.56	55.66	55.60	53.92	84.34	59.48
FADA	DINOv2-L	326.82	61.45	62.86	54.89	<u>58.48</u>	54.89	85.57	61.59
EarthAdapter	DINOv2-L	334.35	58.95	63.33	<u>56.37</u>	56.47	58.06	87.63	71.37
GenCD-FPN (Our)	DINOv2-L	9.19	**67.96**	<u>64.17</u>	55.79	**58.57**	**65.90**	**89.50**	73.92
GenCD-VCO (Our)	DINOv2-L	6.83	<u>66.90</u>	**64.19**	**57.39**	57.18	<u>65.58</u>	<u>89.07</u>	**77.49**

Table 1: Quantitative comparison between our GenCD and existing SoTA methods using mIoU. (%)

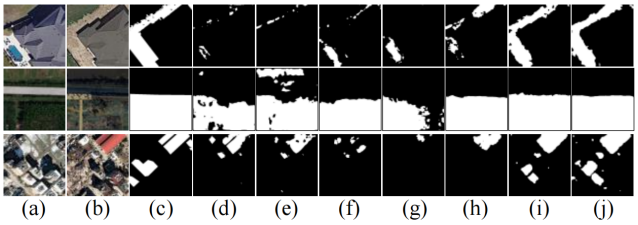

Figure 4: Visual Comparison. (a) I_{t1}, (b) I_{t2}, (c) GT, (d) BAN-CF$_1$, (e) BAN-BiT, (f) Rein, (g) FADA, (h) EarthAdapter, (i) GenCD-VCO, (j) GenCD-FPN.

used to evaluate the performance of GenCD for GCD, ensuring no test data leakage during training. Without instructions, the DINOv2 is used for bitemporal feature extraction while LREA is applied to its qkv projection layers, with the rank set to 8. The number of TGFR K is 2 at each scale.

Comparison with State-of-The-Art Methods

We compare our approach against eight state-of-the-art (SoTA) CD methods: ChangeFormer (Bandara and Patel 2022), Changer (Fang, Li, and Li 2023), TinyCDv2 (2023), MambaBCD (Chen et al. 2024a), Change3D (Zhu et al. 2025), TTP (Chen et al. 2024b), BAN-CF (Li, Cao, and Meng 2024), and BAN-BiT (Li, Cao, and Meng 2024). Among them, BAN and TTP are built upon VFMs, while the others adopt lightweight backbones including MiT-b0 (Xie et al. 2021), VMamba-S (Liu et al. 2024), and X3D-L (Feichtenhofer 2020). To establish a comprehensive benchmark, we also include three recent DGSS methods, Rein (Wei et al. 2024), FADA (Bi et al. 2024), and EarthAdapter (Hu et al. 2025). EarthAdapter is tailored for remote sensing images. These DGSS models are adapted for CD by incorporating the absolute difference for difference extraction.

Quantitative Results. We report the quantitative results of the compared models in Table 1. First, GenCD achieves a competitive trainable parameter count, benefiting from the PEFT strategy and the lightweight decoder that is more compact than the Mask2Former (Cheng et al. 2022) architecture used in Rein, FADA, and EarthAdapter. Second, we report the performance of all methods. The best results are highlighted in bold, and the second-best are underlined.

In land cover CD (JL2SV and JL2SE), GenCD-VCO and GenCD-FPN achieve the best performance. In particular, JL2SV involves cross-platform generalization scenarios, and the bitemporal images in SVCD exhibit significant domain shifts due to seasonal variations. Our methods surpass the previous SoTA on JL2SV by 6.51% and 5.45% in mIoU, respectively. For land use CD (FJ2FX and FX2FJ), which tests cross-location generalization, both variants remain competitive. In building-only CD (LE2TK, LE2WH, and LE2GZ), GenCD-FPN and GenCD-VCO achieve the best results on the challenging LE2TK, characterized by shifts in both geolocation and event domains, outperforming EarthAdapter by 7.84% and 7.52% in mIoU, respectively. Overall, both GenCD variants show superior performance across diverse benchmarks. Meanwhile, benchmark results suggest that land use and land cover CD tasks pose greater generalization challenges in unseen domains than the building-only CD. Furthermore, regarding decoder heads, the results indicate that FPN does not consistently improve performance across all datasets, suggesting that multi-scale feature fusion is not universally beneficial.

Qualitative Results. Figure 4 presents the visual comparison of several advanced methods with ours. The three pairs of images correspond to the land cover CD (JL2SV), land use CD (FJ2FX), and building-only CD (LE2TK) tasks, respectively. The first case includes diverse land cover types, such as vehicles and a swimming pool, that are converted into bare land. Notably, only GenCD successfully captures these changes. In the second case, forest land is transformed into a blank area. Both EarthAdapter and GenCD exhibit su-

Method		VCO			FPN		
		mIoU	mF1	Acc	mIoU	mF1	Acc
$[F_{t_1}; F_{t_2}]$	Full	45.97	54.34	81.92	44.98	49.71	86.57
	Frozen	49.30	55.86	89.90	48.51	54.65	88.87
	+VPT	54.73	64.20	89.00	47.55	53.15	88.57
	+Adapter	49.55	56.38	88.93	46.95	52.46	87.97
	+LoRA	50.27	58.05	88.00	47.72	53.31	88.81
	+LREA	66.82	77.47	91.81	60.10	70.29	91.22
$\lvert F_{t_1} - F_{t_2} \rvert$	Full	47.36	56.68	81.82	45.92	56.82	77.11
	Frozen	58.00	69.30	86.81	61.73	72.23	91.10
	+VPT	57.72	68.60	87.81	60.34	70.78	90.61
	+Adapter	58.25	68.55	89.73	59.21	69.16	91.31
	+LoRA	60.21	71.24	88.80	64.56	74.92	92.57
	+LREA	66.12	77.01	90.90	67.05	77.50	92.58
	++TGFR	**66.90**	**77.53**	**91.91**	**67.96**	**78.34**	**92.85**

Table 2: Ablation studies on various configurations of GenCD and our proposed components.

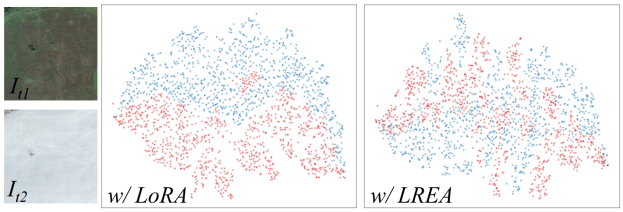

Figure 5: t-SNE visualization illustrating the effect of LREA on bitemporal feature alignment.

perior performance, which is consistent with the quantitative results. The primary challenge in this case arises from the fact that, although the land cover appears locally unchanged, the overall land use type has shifted. The third example includes newly constructed and earthquake-damaged collapsed buildings. While most models detect the former, only GenCD-VCO and GenCD-FPN accurately identify the latter. This highlights the effectiveness of our method in cross-event generalization, given that the training set (LEVIR-CD+) contains no such damage scenarios. Overall, these visual results confirm the effectiveness of GenCD.

Ablation Studies

Given the significant domain diversity of the bitemporal images in SVCD, which reflect the complexity of real-world changes, comprehensive ablation studies are conducted using cross-dataset evaluation from JL1-CD to SVCD.

The effectiveness exploration of various variants of GenCD. Table 2 presents the quantitative results comparing feature concatenation and absolute difference as difference extraction schemes, under various fine-tuning methods and feature decoders (VCO and FPN). The absolute difference consistently outperforms concatenation across previous fine-tuning approaches, Adapter (Chen et al. 2022), VPT (Jia et al. 2022), and LoRA, confirming its better generalization capabilities. Intuitively, concatenation simply merges

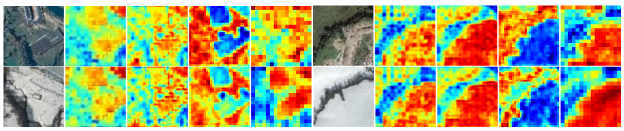

Figure 6: The affinity maps $\mathbf{A}$ visualization in TGFR.

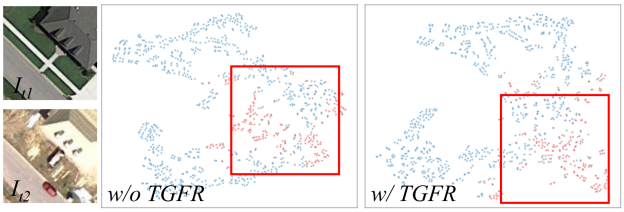

Figure 7: t-SNE visualization illustrating the effect of TGFR on feature separation.

the features without explicitly encoding temporal changes, forcing the model to implicitly learn change patterns. This increases the risk of overfitting to bitemporal domain gaps in the training data, and its asymmetric structure may hinder generalization to time-reversed inputs.

Notably, introducing LREA leads to a substantial performance boost under the concatenation setting. Compared to LoRA, LREA improves mIoU by 16.55% and 12.38% with VCO and FPN as the decode heads, respectively, strongly validating the effectiveness of the proposed LREA. For the absolute difference strategy, LREA also yields mIoU gains of 5.91% and 2.49% under the two decoders, with further improvements achieved through the integration of TGFR.

In addition, we provide qualitative evidence to illustrate the effectiveness of LREA and TGFR. Figure 5 shows a case with significant seasonal variation and visualizes the last-layer bitemporal feature distributions of DINOv2 fine-tuned with LoRA and LREA, respectively. In the LoRA case, a clear decision boundary emerges despite the absence of actual changes, as seasonal differences lead to significant feature shifts. If the model learns to predict unchanged regions under such conditions, the risk of misclassifying real changes as unchanged in other scenarios increases. With LREA, the bitemporal features form an inseparable distribution, indicating effective alignment and further confirming the effectiveness of LREA.

The qualitative analysis of TGFR is illustrated in Figures 6 and 7, with the FPN as the decoder head. Figure 6 displays the affinity maps $\mathbf{A}$ produced by the TGFR modules applied to each scale from two examples with varied change types. It can be confirmed that the learnable token at each scale effectively captures scene-agnostic difference cues for distinguishing changed from unchanged patterns. Furthermore, Figure 7 illustrates the feature distributions of changed and unchanged regions based on the final-layer features used for classification in GenCD-FPN, confirming that the incorporation of TGFR effectively improves feature separability. Notably, the illustrated example covers multiple types of changes, including previously unseen cars.

The generalizability of LREA on different VFMs. Table 3

VFM	Backbone	Method	mIoU	mF1	Acc
SAM	ViT-L	LoRA	58.34	69.03	88.71
		LREA	60.20	70.68	90.41
DINOv2	ViT-B	LoRA	60.31	70.84	90.32
		LREA	64.27	74.91	91.57
Panopticon	ViT-B	LoRA	54.81	63.76	90.26
		LREA	55.24	64.33	90.35
CLIP	ViT-L	LoRA	53.74	64.44	85.29
		LREA	54.50	65.92	84.14
RemoteCLIP	ViT-L	LoRA	55.22	66.48	85.09
		LREA	57.22	67.97	87.85

Table 3: The generalizability of LREA on different VFMs.

Rank r	VCO			FPN		
	mIoU	mF1	Acc	mIoU	mF1	Acc
4	63.88	74.70	90.84	62.36	72.92	91.19
8	66.90	77.53	91.91	67.96	78.34	92.85
16	**67.69**	**78.19**	**92.38**	**68.41**	**78.73**	**93.06**
32	59.69	69.93	90.81	66.68	77.24	92.18

Table 4: The impact of the rank in LREA.

summarizes the effectiveness of LREA on five VFMs, using FPN as the decoder head. LREA consistently surpasses LoRA, validating the generalizability of LREA. Additionally, the DINOv2, despite utilizing the smaller ViT-B backbone, surpasses all other models, including those based on the larger ViT-L backbone. In particular, Panopticon (Waldmann et al. 2025) and RemoteCLIP (Liu et al. 2024) are specialized VFMs pretrained on remote sensing data. This highlights the strong superiority of DINOv2 for GCD.

The hyperparameter analysis of LREA. Table 4 reports the impact of the rank in LREA on the model performance. As the rank increases, the performance first improves and then degrades, achieving the best result at rank 16. Notably, when the rank reaches 32, the performance of GenCD with the VCO decoder drops. A plausible explanation is that lower ranks may result in insufficient capacity to capture discriminative information. In contrast, higher ranks may retain redundant or noisy components, disrupting the alignment effect of exchange adaptation. To balance representation capacity and noise suppression in LREA, the rank is chosen as 8 in GenCD. Besides, Table 5 compares the impact of inserting the LREA module at different positions. The results show that the (q, v) and (k, v) configurations perform worse than full (q, k, v) and only v integration, which can be attributed to the distribution mismatch between query and key features. Integrating LREA into all three components (q, k, v) consistently achieves the best performance in terms of mIoU, mF1, and Acc on both decoders.

The hyperparameter analysis of TGFR. We investigate the sensitivity of GenCD to the number of learnable tokens used in TGFR, with results presented in Table 6. As the token length increases, performance gradually degrades, indicating that a single input-independent token is better suited for capturing generalized difference cues across diverse scenarios and change types. In contrast, using multiple tokens may lead each to partially encode specific characteristics of the training data, limiting generalization.

LREA	VCO			FPN		
	mIoU	mF1	Acc	mIoU	mF1	Acc
v	64.85	75.66	91.08	62.94	73.52	91.36
q, v	63.51	74.21	91.48	64.53	75.11	91.81
k, v	61.89	72.83	89.81	61.76	72.36	90.84
q, k, v	66.90	77.53	91.91	67.96	78.34	92.85

Table 5: The impact of LREA insertion positions.

N	VCO			FPN		
	mIoU	mF1	Acc	mIoU	mF1	Acc
1	66.90	77.53	91.91	67.96	78.34	92.85
10	64.99	75.69	91.52	66.51	77.00	92.46
20	63.99	74.63	91.50	66.69	77.29	92.00

Table 6: The impact of the number of tokens used in TGFR.

Limitations

The limitations of GenCD mainly lie in two major categories. First, model design limitations include the sensitivity of LREA to the feature distribution of VFMs, despite its demonstrated improvements across distinct VFMs, and the difficulty for TGFR to refine all real-world changes, particularly those with complex patterns, even though its effectiveness has been validated. Second, evaluation limitations comprise the insufficient exploration of model selection strategies (Yu et al. 2024) and limited robustness evaluation under image degradations (Li et al. 2025).

Conclusion

GCD aims to train models that generalize well to unseen domains, making it both challenging and practically valuable. In this study, we conduct a preliminary exploration and propose GenCD, a method based on VFMs. GenCD includes two key components: the LREA strategy tailored for GCD, which effectively mitigates the domain gap between bitemporal images and improves generalization compared to LoRA; and the TGFR module, which enhances the distinction between changed and unchanged regions. With its input-agnostic design, TGFR remains effective in unseen scenarios. Extensive experiments across three binary CD tasks and eight datasets confirm the superior performance of GenCD. We also analyze the impact of different component configurations within GenCD, offering insights for future improvements. In summary, GCD is an important task that warrants further in-depth investigation.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62471394 and U21B2041, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025091.

References

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bandara, W. G. C.; and Patel, V. M. 2022. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, 207–210. IEEE.

Bi, Q.; Yi, J.; Zheng, H.; Zhan, H.; Huang, Y.; Ji, W.; Li, Y.; and Zheng, Y. 2024. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37: 94047–94072.

Chen, H.; and Shi, Z. 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10): 1662.

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

Chen, K.; Liu, C.; Li, W.; Liu, Z.; Chen, H.; Zhang, H.; Zou, Z.; and Shi, Z. 2024b. Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, 8581–8584. IEEE.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Chen, Z.; Zeng, Y.; Chen, Z.; Gao, H.; Chen, L.; Liu, J.; and Zhao, F. 2025. VFM-Adapter: Adapting Visual Foundation Models for Dense Prediction with Dynamic Hybrid Operation Mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2385–2393.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11580–11590.

Codegoni, A.; Lombardi, G.; and Ferrari, A. 2023. TINYCD: A (not so) deep learning model for change detection. *Neural Computing and Applications*, 35(11): 8471–8486.

Fang, S.; Li, K.; and Li, Z. 2023. Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 203–213.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, X.; Gong, Z.; Wang, Y.; Jia, Y.; Luo, G.; and Yang, X. 2025. Earth-Adapter: Bridge the Geospatial Domain Gaps with Mixture of Frequency Adaptation. *arXiv preprint arXiv:2504.06220*.

Ji, S.; Wei, S.; and Lu, M. 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1): 574–586.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Jia, Y.; Hoyer, L.; Huang, S.; Wang, T.; Van Gool, L.; Schindler, K.; and Obukhov, A. 2024. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, 91–109. Springer.

Kennedy, R. E.; Townsend, P. A.; Gross, J. E.; Cohen, W. B.; Bolstad, P.; Wang, Y.; and Adams, P. 2009. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote sensing of environment*, 113(7): 1382–1396.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lebedev, M.; Vizilter, Y. V.; Vygolov, O.; Knyaz, V. A.; and Rubis, A. Y. 2018. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42: 565–571.

Li, K.; Cao, X.; and Meng, D. 2024. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12.

Li, X.; Tao, Y.; Zhang, S.; Liu, S.; Xiong, Z.; Luo, C.; Liu, L.; Pechenizkiy, M.; Zhu, X. X.; and Huang, T. 2025. RE-OBench: Benchmarking Robustness of Earth Observation Foundation Models. arXiv:2505.16793.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16.

Liu, Z.; Zhu, R.; Gao, L.; Zhou, Y.; Ma, J.; and Gu, Y. 2025. JL1-CD: A New Benchmark for Remote Sensing Change Detection and a Robust Multi-Teacher Knowledge Distillation Framework. *arXiv preprint arXiv:2502.13407*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Pak, B.; Woo, B.; Kim, S.; Kim, D.-h.; and Kim, H. 2024. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, 37–54. Springer.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, 464–479.

Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; and Huang, X. 2021. SemiCDNet: A Semisupervised Convolutional Neural Network for Change Detection in High Resolution Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 5891–5906.

Peng, D.; Lei, Y.; Hayat, M.; Guo, Y.; and Li, W. 2022. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2594–2605.

Peng, Z.; Xu, Z.; Zeng, Z.; Huang, Y.; Wang, Y.; and Shen, W. 2025. Parameter-efficient Fine-tuning in Hyperspherical Space for Open-vocabulary Semantic Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15009–15020.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shen, J.; Zhang, C.; Zhang, M.; Li, Q.; and Wang, Q. 2024. Learning remote sensing aleatoric uncertainty for semi-supervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Tang, P.; Zhang, X.; Yang, C.; Yuan, H.; Sun, J.; Shan, D.; and Yang, Z. J. 2025. Unleashing the Power of Visual Foundation Models for Generalizable Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20823–20831.

Waldmann, L.; Shah, A.; Wang, Y.; Lehmann, N.; Stewart, A.; Xiong, Z.; Zhu, X. X.; Bauer, S.; and Chuang, J. 2025. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2204–2214.

Wei, Z.; Chen, L.; Jin, Y.; Ma, X.; Liu, T.; Ling, P.; Wang, B.; Chen, H.; and Zheng, J. 2024. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 28619–28630.

Willis, K. S. 2015. Remote sensing change detection for ecological monitoring in United States protected areas. *Biological Conservation*, 182: 233–242.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Yang, K.; Xia, G.-S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; and Zhang, L. 2021. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Yu, H.; Zhang, X.; Xu, R.; Liu, J.; He, Y.; and Cui, P. 2024. Rethinking the evaluation protocol of domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21897–21908.

Yuan, S.; Lin, G.; Zhang, L.; Dong, R.; Zhang, J.; Chen, S.; Zheng, J.; Wang, J.; and Fu, H. 2024. FUSU: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. *Advances in Neural Information Processing Systems*, 37: 132417–132439.

Zang, Q.; Wang, S.; Zhao, D.; Quan, D.; Hu, Y.; and Jiao, L. 2025a. Generalization-aware Remote Sensing Change Detection via Domain-agnostic Learning. *arXiv preprint arXiv:2504.00543*.

Zang, Q.; Zhao, D.; Wang, S.; Quan, D.; and Zhong, Z. 2025b. Feature Spectrum Learning for Remote Sensing Change Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12647–12657.

Zhang, M.; Li, Q.; Miao, Y.; Yuan, Y.; and Wang, Q. 2023. Difference-guided aggregation network with multiimage pixel contrast for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Zhang, X.; Huang, X.; Li, J.; Yang, J.; Wang, L.; and Xie, X. 2024. Enhancing inter-class discrimination for domain adaptation of change detection. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, 8513–8517. IEEE.

Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; and Zhang, L. 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265: 112636.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35: 338–350.

Zhu, D.; Huang, X.; Huang, H.; Zhou, H.; and Shao, Z. 2025. Change3D: Revisiting Change Detection and Captioning from A Video Modeling Perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24011–24022.