Hie-DinoMamba: Hierarchical DINOv3 and Mamba Architecture for Multi-Class Building Change Detection

Youngwoong Yoon[1], Jangwoo Cheon[2], Hwiyoung Kim[3], Impyeong Lee[4]

[1] Geospatial Team, Innopam, Seoul, Republic of Korea - yyw9969@innopam.com
[2] Geospatial Team, Innopam, Seoul, Republic of Korea - cjw@innopam.com
[3] Geospatial Team, Innopam, Seoul, Republic of Korea - hykim@innopam.com
[4] Dept. of Geoinformatics, University of Seoul, Seoul, Republic of Korea - iplee@uos.ac.kr

Keywords: Multi-Class Building Change Detection, Visual Foundation Models, DINOv3, Visual State Space Models, Low-Rank Adaptation

Abstract

Accurate multi-class building change detection in high-resolution aerial imagery is a critical task for urban analysis. However, it is hindered by two key challenges: severe class imbalance and the difficulty of obtaining robust, generalizable feature representations. While recent models show promise, encoders trained from scratch on aerial data remain limited in their representational capacity. Leveraging large-scale Visual Foundation Models (VFMs) offers a path to better features, but full fine-tuning is computationally prohibitive. To address this, we propose Hie-DinoMamba, a novel hierarchical architecture. We integrate a frozen 1.1B parameter DINOv3-L (SAT-493M) encoder, preserving its rich pre-trained knowledge. We efficiently adapt this encoder to the aerial domain using parameter-efficient Low-Rank Adaptation (LoRA). Furthermore, we design a new Hierarchical Mamba FPN decoder that uses Visual State Space Model (VSSM, Mamba) blocks to fuse and refine multi-scale feature pairs in a top-down manner. The model is optimized using a dual-loss strategy (Semantic and Boundary) to ensure both classification accuracy and precise boundary delineation. On the 4-class aerial building change detection benchmark, Hie-DinoMamba achieves state-of-the-art performance with an mIoU of 65.12%, a significant improvement of 2.1 percentage points over the strong VSSM-based baseline (ChangeMamba-MC). Qualitative analysis further demonstrates our model’s superior generalization, successfully detecting complex changes in unseen geographic regions where other models fail.

1. Introduction

The dynamic monitoring of urban environments is a critical task for sustainable development, infrastructure planning, and disaster management (Zhao et al., 2023, Saputra et al., 2025). Change detection—the process of identifying differences between co-registered images of the same area acquired at two time points (T1 and T2) to produce a per-pixel classification map—has long been a core technique in remote sensing for this purpose. However, binary change detection, which merely identifies the presence or absence of change, is insufficient for comprehensive urban analysis. Effective governance and planning require a multi-class approach that differentiates the nature of transformations (Zhu et al., 2024). This research addresses the challenge of 4-class building change detection, classifying each pixel into No Change, New Construction, Demolition, or Renovation, aiming to provide a robust and precise tool for understanding the nuanced dynamics of urban evolution.

Achieving this level of granularity presents formidable challenges. First, the task demands a high degree of semantic understanding. Models must not only detect pixel-level differences but also interpret the context of those differences to correctly classify the change type. This is particularly difficult for subtle categories like renovation, which may be texturally similar to no-change areas. Second, the severe class imbalance inherent in real-world aerial imagery, where unchanged regions vastly outnumber all change classes combined, poses a significant obstacle, biasing models toward the dominant no-change class (Aguilar-Ruiz and Michalak, 2024).

Many existing deep learning architectures for change detection, including those based on Convolutional Neural Networks (CNNs) or Transformers, are trained from scratch on specific, often limited, aerial datasets (Khelifi and Mignotte, 2020). This approach curtails their representational capacity, as they lack the broad visual understanding derived from large-scale, diverse pre-training. While recent Visual State Space Models (VSSMs) have shown promise in efficiently modeling long-range dependencies, they often share this same limitation when trained solely on domain-specific data (Chen et al., 2024).

To address this fundamental challenge in feature representation, we leverage large-scale Visual Foundation Models (VFMs). These models, pre-trained on massive, diverse datasets, possess an extraordinary ability to extract rich, generalizable, and robust visual features. We hypothesize that this powerful prior knowledge is key to achieving new performance levels in multi-class change detection. However, leveraging these colossal models, such as the 1.1-billion parameter DINOv3-Large (DINOv3-L), introduces a new, critical challenge: computational feasibility. Fully fine-tuning such a model for a specialized task is computationally prohibitive and risks catastrophic forgetting of its valuable pre-trained knowledge.

Our primary contribution is a novel framework, **Hie-DinoMamba**, that successfully integrates a large-scale VFM through a highly efficient adaptation strategy. We employ DINOv3-L as a frozen Siamese encoder, preserving its entire 1.1-billion parameter backbone to act as a powerful feature extractor. To bridge the significant domain gap between its general pre-training (satellite and ground-level imagery) and our target (aerial imagery), we integrate parameter-efficient Low-Rank Adaptation (LoRA) modules into its attention mechan-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-601-2026

isms. This allows the model to adapt to the new domain using only a minuscule fraction of trainable parameters, ensuring high computational efficiency.

With a robust encoder in place, our second major contribution is a novel Hierarchical Siamese Mamba Feature Pyramid Network (FPN) decoder. This decoder fuses multi-scale feature pairs from the frozen encoder in a top-down manner, integrating VSSM (Mamba) blocks for efficient spatio-temporal refinement to produce clean and accurate change masks.

Finally, we employ a dual-loss strategy that decouples semantic classification (via Focal Loss) from boundary delineation (via a geometry-focused Boundary Loss), ensuring both accurate change-type predictions and precise spatial boundaries. Combined with proven data-level strategies such as object-aware cropping, our framework learns effectively from all classes despite severe class imbalance. The contributions of this work are thus summarized as follows:

- A new paradigm for multi-class change detection is proposed that leverages a frozen 1.1B parameter DINOv3-L foundation model as an encoder. This approach is made computationally feasible and domain-specific via the integration of parameter-efficient LoRA.
- A novel Hierarchical Siamese Mamba FPN decoder is designed to fuse multi-scale feature pairs from the VFM encoder. This decoder utilizes VSSM blocks for efficient spatio-temporal refinement.
- A hierarchical dual-loss strategy is employed to decouple the optimization of semantic classification and boundary segmentation. This improves overall precision by optimizing for each task with a specialized head and loss function.
- The proposed Hie-DinoMamba sets a new state-of-the-art in 4-class building change detection, demonstrating superior accuracy, boundary definition, and minority class identification compared to existing models.

This paper is structured as follows: Section 2 reviews related work in change detection and model adaptation. Section 3 presents the Hie-DinoMamba architecture in detail. Section 4 outlines the experimental setup, and Section 5 provides a comprehensive analysis of the results. Section 6 concludes with a summary of our findings and potential directions for future research.

2. Related Work

2.1 Deep Learning for Change Detection

Early deep learning approaches for change detection were predominantly based on CNNs. Architectures, particularly those leveraging Siamese networks or U-Net variants, proved effective at extracting hierarchical spatial features. While adept at capturing local details, CNN-based models inherently struggle with modeling long-range dependencies due to their constrained receptive fields. Transformer-based architectures emerged as a powerful alternative, utilizing self-attention mechanisms to capture global contextual relationships across the entire image pair. However, this global modeling capability often incurs a significant computational cost, exhibiting quadratic complexity relative to input size (Khelifi and Mignotte, 2020). Furthermore, their data-hungry nature poses challenges for specialized remote sensing tasks where large-scale annotated datasets are scarce.

2.2 Multi-Class Change Detection and Class Imbalance

The majority of change detection literature has concentrated on binary tasks. Multi-class change detection, which our work addresses, remains a more complex and less-explored domain (Zhu et al., 2024). It requires a model to move beyond simple difference detection to perform fine-grained semantic classification for each of the four change types (new construction, demolition, renovation, and no change). A critical and persistent challenge in this area is the severe class imbalance inherent in datasets, where the no-change class and other dominant categories vastly underrepresent critical minority classes. To address this, research has bifurcated into data-level and model-level strategies. Data-level approaches include re-sampling techniques, such as object-aware cropping, to increase the prevalence of minority samples during training. Model-level strategies often involve designing sophisticated loss functions, such as focal loss or dynamic class-weighting schedulers, to focus the model’s learning on hard-to-classify examples (Aguilar-Ruiz and Michalak, 2024).

2.3 Visual Foundation Models and Parameter-Efficient Adaptation

Concurrently, the emergence of large-scale VFMs, pre-trained on web-scale datasets, has revolutionized computer vision. These models, such as DINOv3, possess exceptionally rich and generalizable feature representations learned from diverse data (Siméoni et al., 2025). However, their colossal size (often exceeding one billion parameters) makes full fine-tuning on downstream tasks computationally prohibitive and risks catastrophic forgetting of their powerful prior knowledge. Parameter-Efficient Fine-Tuning (PEFT) techniques have emerged as a solution. Among these, LoRA is particularly prominent (Hu et al., 2021), freezing the pre-trained weights and injecting small, trainable rank-decomposition matrices into specific layers, such as the attention blocks. This allows the model to adapt to a new domain or task using only a tiny fraction of trainable parameters, preserving the original model’s knowledge while maximizing computational efficiency. Recent work has begun exploring VFMs for change detection; for example, ChangeCLIP (Dong et al., 2024) leverages the CLIP vision-language model to extract robust semantic features for binary change detection. Our work differs by employing a satellite-specialized VFM (DINOv3-L) with LoRA adaptation for the more challenging multi-class setting.

2.4 Visual State Space Models (VSSM)

As an alternative to the quadratic complexity of Transformers, VSSMs, inspired by the Mamba architecture, have recently gained significant traction (Gu and Dao, 2023). VSSMs operate by encoding visual information into a latent state and applying a selective scan mechanism. This design allows them to model long-range dependencies with linear complexity, effectively combining the global context capabilities of Transformers with the high efficiency of CNNs. Their application in dense prediction tasks, including semantic segmentation and, more recently, remote sensing change detection, has demonstrated strong performance. VSSMs have shown potential in outperforming Transformer-based counterparts with significantly lower computational overhead, making them a compelling choice for efficient and powerful decoder designs. More recently, Wang et al. (Wang et al., 2025) proposed SPMNet, a Siamese Pyramid Mamba network that integrates an omni-directional selective scan module within a CNN-Mamba hybrid

backbone for binary change detection, demonstrating strong performance on WHU-CD and LEVIR-CD benchmarks. While SPMNet also leverages Mamba blocks within a pyramid structure, it is designed for binary change detection with an encoder trained from scratch. In contrast, our approach addresses the more challenging multi-class setting and leverages a frozen 1.1B-parameter VFM encoder with LoRA adaptation.

3. Methodology

3.1 Overall Architecture of Hie-DinoMamba

The proposed Hie-DinoMamba architecture is designed to integrate the powerful representation capacity of a VFM with the efficiency of VSSMs, tailored specifically for the 4-class aerial change detection task. The complete framework, illustrated in Figure 1, is composed of two primary stages: a Siamese frozen encoder backbone and a hierarchical Mamba decoder. The process begins with a pair of pre-change (T1) and post-change (T2) images. Both images are fed into a weight-sharing, frozen DINOv3-L encoder, which has been adapted for the aerial domain using LoRA modules. This encoder extracts hierarchical feature maps from four distinct stages. These multi-scale feature pairs are then passed to the Hierarchical Siamese Mamba FPN decoder. Within the decoder, paired features from both time points are combined at each scale. These fused features are subsequently refined in a top-down manner, utilizing VSSM blocks to efficiently model spatio-temporal dependencies. Finally, the hierarchical decoder uses a dual-loss strategy, employing two separate heads (Semantic and Boundary) to generate the final precise 4-class change map.

3.2 Frozen Foundation Model Encoder (DINOv3-L)

The core of our feature extraction process is the DINOv3-L model, a state-of-the-art VFM with 1.1 billion parameters. We specifically utilize the checkpoint pre-trained on the SAT-493M dataset, which consists of 493 million RGB images sampled from Maxar satellite ortho-imagery (Siméoni et al., 2025). This pre-training on large-scale satellite data provides a significant advantage over standard web-data pre-training. It endows the model with a rich visual understanding of overhead-view characteristics, large-scale geographic patterns, and atmospheric conditions, making it highly suitable for remote sensing tasks.

However, despite this strong domain relevance, a crucial gap persists. The pre-training data consists of satellite imagery (e.g., 0.6m resolution Maxar data), which has distinctly different properties from our target aerial imagery (e.g., higher resolution, different sensor characteristics, and lower-altitude perspectives). Our framework's critical decision is the use of a frozen encoder strategy to manage this. The entire 1.1B parameter DINOv3-L backbone remains completely frozen during training. This strategy is twofold:

1. **Preservation of Prior Knowledge:** By freezing the encoder, we ensure that both its generalized knowledge (from large-scale web pre-training) and its specialized knowledge (from the SAT-493M satellite dataset) are preserved. This avoids catastrophic forgetting, where the model's valuable representations could be damaged by the narrow data distribution of a new task.
2. **Computational Feasibility:** Training or fine-tuning a 1.1B parameter model in a Siamese configuration is computationally prohibitive. A frozen strategy drastically reduces the training cost, as gradients are only computed for the decoder and the lightweight adaptation modules.

This approach allows us to leverage the full capacity of a remote-sensing VFM, while the subsequent adaptation (Section 3.3) bridges the remaining domain gap between satellite and aerial imagery.

3.3 Parameter-Efficient Adaptation via LoRA

While the frozen encoder preserves powerful pre-trained features, it cannot natively capture the nuances of our target aerial imagery. We therefore employ LoRA (Hu et al., 2021), a parameter-efficient fine-tuning technique that introduces a small number of trainable parameters without modifying the original encoder weights.

The mathematical principle of LoRA is to represent the weight update ΔW for a pre-trained weight matrix W_0 using a low-rank approximation. The full adapted weight W is formulated as:

$$W = W_0 + \Delta W \tag{1}$$

where W = full adapted weight matrix
W_0 = frozen pre-trained weight matrix
ΔW = matrix representing the learned change

LoRA's core innovation is the hypothesis that ΔW can be effectively approximated by factorizing it into two smaller, low-rank matrices, B and A:

$$\Delta W = BA \tag{2}$$

where W_0 has dimensions $d \times k$
B = trainable matrix with dimensions $d \times r$
A = trainable matrix with dimensions $r \times k$
r = rank ($r \ll d, k$)

The actual weight update is scaled by a factor of α/r, where α is a hyperparameter that controls the magnitude of the adaptation. A higher α relative to r amplifies the learned update, while a lower ratio keeps the model closer to its pre-trained state. In our Hie-DinoMamba framework, we inject these trainable LoRA modules into the query and value projection layers within the attention mechanisms of the frozen DINOv3-L encoder. The original pre-trained weights (W_0) remain frozen and untouched. Only the new, lightweight matrices B and A (which constitute ΔW) are updated during training. This strategy achieves two critical goals:

1. **Domain Adaptation:** The model learns to adapt its attention mechanisms specifically to the nuances of our high-resolution aerial imagery, effectively bridging the domain gap from satellite data.
2. **Efficiency:** The number of trainable parameters (determined by r) is minuscule compared to the full 1.1B backbone, dramatically reducing memory overhead and enabling the entire framework to be trained on standard high-end GPUs.

By using LoRA, we efficiently fine-tune the encoder's behavior without altering its core knowledge, creating a feature extractor that is both powerful and highly specialized for our 4-class change detection task.

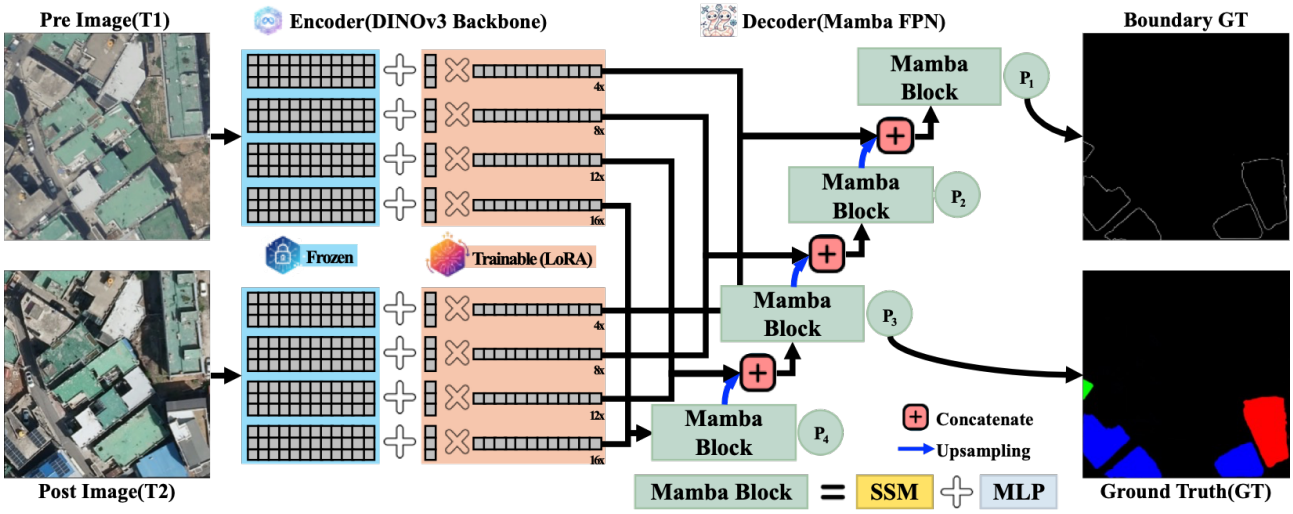

Figure 1. Overall architecture of Hie-DinoMamba. The Siamese DINOv3-L encoder (left) is frozen, with trainable LoRA modules adapting its hierarchical features. These features are processed by the Mamba FPN decoder (right), which fuses paired features from both time points and uses VSSM (Mamba) blocks to refine the maps in a top-down manner. The model is optimized with a dual-loss strategy, using a Semantic Head (P_3) and a Boundary Head (P_1). In the Ground Truth (GT) masks, colors represent: Black (No Change), Red (New Construction), Green (Demolition), and Blue (Renovation).

3.4 Hierarchical Siamese Mamba FPN Decoder

The decoder architecture is the central component for interpreting the hierarchical features provided by the frozen DINOv3-L encoder. Our decoder, the Siamese FPN, is designed to fuse and refine multi-scale feature pairs across hierarchical levels using VSSM blocks (Gu and Dao, 2023). The process begins by receiving the two lists of multi-scale features, F_{T1} and F_{T2}, from the Siamese encoder (four feature maps in each list, from F_1 to F_4). As depicted in our decoder design, the first step is to combine the feature representations from both time points at each of the four hierarchical levels. Specifically, the pre-change and post-change feature maps at each level i are concatenated along the channel dimension and then projected through a 1×1 convolutional layer to a uniform channel dimension $F_{fuse,i}$. This allows the model to learn the optimal comparison strategy for detecting changes from the paired features. Once the fused feature maps are prepared, they are processed through a top-down FPN structure. The process starts with the deepest, lowest-resolution feature map, $F_{fuse,4}$. This feature map is passed through the top-most processing block, which contains a VSSM block, to produce the initial change map P_4. The decoder then iterates upwards, fusing features from deeper layers with shallower ones. For each level i (from 3 down to 1):

1. The feature map P_{i+1} from the deeper layer is upsampled by a factor of 2.
2. The upsampled feature map is concatenated with the corresponding fused feature map from the encoder, $F_{fuse,i}$.
3. This fused tensor is passed through a subsequent lateral block, which also contains a VSSM block, to produce the refined change map P_i.

A key innovation of our decoder is the inclusion of VSSM blocks at every stage of the FPN. At their core, these blocks implement a State Space Model (SSM) defined by the continuous system $h'(t) = Ah(t) + Bx(t)$, $y(t) = Ch(t)$, which is discretized for sequence processing. The selective scan mechanism makes matrices A, B, and C input-dependent, enabling content-aware reasoning with linear $O(n)$ complexity—in contrast to the quadratic cost of self-attention. In our decoder, VSSM blocks process the fused feature maps at each FPN level using a 2D selective scan across four spatial directions, allowing the model to capture long-range spatial context and produce cleaner, more coherent segmentation masks. The final output is a list of four refined feature maps (P_1,P_2,P_3,P_4), which are passed to the dual heads for loss computation.

3.5 Hierarchical Dual-Loss Strategy

To effectively optimize the hierarchical decoder, we employ a dual-loss strategy that decouples the task of semantic classification from that of precise boundary delineation. This approach recognizes that different levels of the FPN are specialized for different tasks. First, a Semantic Head is attached to a deep, lower-resolution feature map (P_3, the third decoder output). This feature map is rich in contextual and semantic information, making it ideal for accurately classifying the type of change (e.g., new construction vs. demolition). This head is trained using a primary classification loss, such as Focal Loss, to handle class imbalance at the semantic level (Lin et al., 2017). Second, a Boundary Head is attached to the shallowest, highest-resolution feature map (P_1, the first decoder output). This feature map retains the most precise spatial and edge details, making it optimal for delineating the exact boundaries of the changed objects. This head is trained with a dedicated Boundary Loss, which is a geometry-focused compound loss. This Boundary Loss is a weighted combination of Focal Tversky Loss and Dice Loss, designed to be robust to the extreme imbalance of boundary-pixel classification (Salehi et al., 2017, Milletari et al., 2016). The Focal Tversky Loss is used to balance the precision and recall of the boundary detection, while the Dice Loss provides stable gradients by maximizing the overlap between predicted and ground-truth boundary pixels. By training the model to optimize these two distinct objectives at different hierarchical levels simultaneously, Hie-DinoMamba learns to produce change maps that are both semantically accurate and spatially precise.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-601-2026 | © Author(s) 2026. CC BY 4.0 License.

4. Experimental Setup

4.1 Dataset

This study utilizes a large-scale aerial building change detection dataset from AI-Hub (National Information Society Agency, 2023), managed by the National Information Society Agency (NIA) of South Korea. The dataset consists of 41,548 image pairs from two different time points (T1 and T2), captured over urban areas in Seoul, South Korea. The dataset is annotated for multi-class change detection with four distinct classes: No Change (0), New construction (1), Demolition (2), and Renovation (3). The images are provided as 512×512 pixel patches at 0.1 m (10 cm) spatial resolution. For our experiments, we follow the official data split: 36,940 patches for training and 4,608 patches for testing.

4.2 Class Imbalance Mitigation

As discussed in the Introduction, the dataset exhibits a severe class imbalance, with the ’No Change’ class dominating the pixel distribution. To mitigate this foundational problem, we adopt two key strategies from prior work, which are also implemented in our data loader. First, we employ Object-Aware Cropping during training (Mishra et al., 2021). Instead of standard random cropping, this method ensures that a high percentage of the training crops are centered on pixels belonging to the minority change classes (New construction, Demolition, Renovation). This significantly increases the representation of these critical classes during the training process. Second, we apply standard data augmentation techniques, including geometric augmentations (random flips, rotations) and color augmentations (color jitter, noise).

4.3 Evaluation Metrics

To provide a comprehensive and robust evaluation of our model’s performance on the 4-class task, we utilize a suite of standard semantic segmentation metrics, as implemented in our evaluation module and presented in our results (Section 5). The performance is assessed using Precision, Recall, F1-score, and Intersection over Union (IoU), which are calculated for each of the four classes (No Change, New construction, Demolition, Renovation). To summarize overall performance, we use two primary aggregate metrics:

- **mean Intersection over Union (mIoU):** The main metric for overall segmentation quality, averaged across all four classes.
- **Cohen’s Kappa:** A statistical measure that assesses the agreement between the model’s predictions and the ground truth, correcting for chance agreement.

4.4 Baseline Models

To validate the effectiveness of Hie-DinoMamba, we compare its performance against several state-of-the-art (SOTA) change detection models, including:

- **SNUNet:** A deeply supervised network with channel attention (Fang et al., 2021).
- **BAN:** A bitemporal attention network (Li et al., 2023).
- **ChangeFormer:** A Transformer-based model for change detection (Chen et al., 2022).
- **ChangeMamba-MC:** A VSSM-based baseline model for multi-class change detection that relies on an encoder trained from scratch (Chen et al., 2024).

4.5 Implementation Details

All models are trained under the same conditions for a fair comparison. The Hie-DinoMamba model is implemented using PyTorch. The DINOv3-L encoder is initialized with the SAT-493M pre-trained weights. The LoRA modules are applied with a rank (r) of 8 and an alpha of 16. For optimization, we use the AdamW optimizer with an initial learning rate of 1e-4, cosine learning-rate decay, and a batch size of 32. The hierarchical dual-loss is configured as described in Section 3.5 (Focal Loss for the semantic head and a combination of Focal Tversky and Dice losses for the boundary head, with a Dice weight of 0.5).

5. Results and Discussion

5.1 Quantitative Results

We conduct a comprehensive quantitative evaluation of our proposed Hie-DinoMamba against four established baseline models: SNUNet, BAN, ChangeFormer, and our VSSM-based baseline, ChangeMamba-MC. The detailed per-class and overall results for this 4-class comparison are presented in Table 1. As shown in Table 1, Hie-DinoMamba achieves a new state-of-the-art performance on the benchmark. Our model obtains an overall mIoU of 65.12% and a Kappa of 75.77%. This represents a significant improvement of 2.1 percentage points in mIoU over the strongest baseline, ChangeMamba-MC (63.02% mIoU).

These results validate the effectiveness of our proposed architecture, demonstrating that leveraging a pre-trained DINOv3-L VFM with LoRA adaptation provides a substantial advantage over a VSSM encoder trained from scratch. In addition to the superior overall performance, the per-class IoU scores highlight our model’s specific strengths. Hie-DinoMamba achieves the highest IoU scores in three of the four categories: No Change (97.09%), New Construction (56.17%), and Renovation (61.38%). Notably, the substantial 4.97 percentage points improvement in the challenging ‘Renovation’ class (61.38% vs. 56.41% from ChangeMamba-MC) underscores the superior feature representation provided by the DINOv3-L encoder. While the ‘Demolition’ class (45.84% IoU) remains a significant challenge for all models—with ChangeFormer achieving the highest IoU (48.44%) in this category—our model’s overall performance demonstrates the effectiveness of the proposed architecture.

5.2 Ablation Study

To quantify the individual contribution of each key component, we conduct an ablation study by systematically removing one component at a time. Table 2 presents the per-class IoU and overall results for all configurations trained for 100K iterations.

Among the three components, removing LoRA causes the largest overall degradation (-4.17 percentage points in mIoU), confirming that domain adaptation of the frozen encoder is the single most critical component. The DINOv3-L backbone, pre-trained on 0.6 m Maxar satellite imagery, produces features that are highly relevant but not perfectly aligned with our higher-resolution aerial data. Without LoRA, performance drops uniformly across all change classes: New Construction IoU falls from 56.17% to 49.37% (-6.80), Demolition from 45.84% to 40.81% (-5.03), and Renovation from 61.38% to 57.63%

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-601-2026 | © Author(s) 2026. CC BY 4.0 License.

Baseline Models	Class	Recall	Precision	F1 score	IoU	mIoU	Kappa
SNUNet	No Change	95.39	97.36	96.37	92.99	44.31	52.90
	New Construction	44.93	47.38	46.12	29.97		
	Demolition	22.37	52.73	31.41	18.63		
	Renovation	60.18	46.65	52.56	35.65		
BAN	No Change	91.38	**98.54**	94.82	90.16	54.61	68.24
	New Construction	**75.61**	50.43	60.5	43.37		
	Demolition	**66.70**	61.73	64.12	47.18		
	Renovation	70.22	44.93	54.80	37.74		
ChangeFormer	No Change	97.88	97.62	97.75	95.60	60.36	69.01
	New Construction	60.30	72.37	65.79	49.02		
	Demolition	56.51	**77.24**	**65.27**	**48.44**		
	Renovation	62.33	68.37	65.21	48.38		
ChangeMamba-MC	No Change	98.31	97.93	98.12	96.31	63.02	70.71
	New Construction	64.99	**74.07**	69.23	52.94		
	Demolition	63.61	63.17	63.39	46.40		
	Renovation	72.24	72.03	72.13	56.41		
Hie-DinoMamba	No Change	**98.75**	98.29	**98.52**	**97.09**	**65.12**	**75.77**
	New Construction	72.74	71.14	**71.93**	**56.17**		
	Demolition	57.46	69.39	62.86	45.84		
	Renovation	**73.29**	**79.07**	**76.07**	**61.38**		

Table 1. Quantitative performance comparison against established baseline models. The evaluation includes per-class metrics, overall mIoU, and Kappa scores. Bold values indicate the best performance in each category.

Configuration	No Change	New Const.	Demolition	Renovation	mIoU	Kappa
Full model	**97.09**	**56.17**	45.84	**61.38**	**65.12**	**75.77**
w/o LoRA	95.99	49.37	40.81	57.63	60.95	71.80
w/o Mamba (Conv)	96.62	53.05	**48.33**	60.24	64.56	74.62
w/o Boundary Loss	95.57	53.12	46.64	56.22	62.89	71.26

Table 2. Ablation study results. Each row removes one component from the full model. Per-class IoU (%) and overall metrics are reported. Bold values indicate the best in each column.

(-3.75). The disproportionately large drop in New Construction suggests that LoRA is particularly important for distinguishing newly appeared structures from the existing urban fabric. These results validate our hypothesis that a frozen VFM requires targeted domain adaptation for optimal downstream performance.

The dual-loss strategy contributes an improvement of 2.23 percentage points in mIoU over using only the semantic classification loss. The impact is most pronounced on the Renovation class, which drops from 61.38% to 56.22% (-5.16) when the boundary objective is removed, consistent with the nature of renovation changes that involve modifications to building facades without altering the overall footprint. While the w/o Boundary variant achieves a marginally higher Demolition IoU (46.64% vs. 45.84%), this comes at the cost of a substantial overall mIoU degradation (-2.23), demonstrating that the full model's dual-loss configuration provides the best balance across all classes. The No Change class also drops notably (-1.52 to 95.57%), indicating that the boundary head helps suppress false positives along object edges.

Replacing Mamba blocks with standard residual convolutional blocks results in the smallest overall change (-0.56 percentage points in mIoU), suggesting that the representations from DINOv3-L already encode substantial global context. A per-class analysis reveals that Mamba's contribution is class-dependent. Renovation IoU decreases from 61.38% to 60.24% (-1.14), consistent with the need for long-range spatial dependencies when segmenting spatially distributed renovation areas. Conversely, the w/o Mamba variant achieves a higher Demolition IoU (48.33% vs. 45.84%), likely because demolition events involve localized structural removal where the local receptive field of convolutional blocks is more effective. However, this per-class advantage again does not translate to overall superiority, as the full model maintains the highest mIoU (65.12%) and Kappa (75.77%) across all configurations.

In summary, the ablation results reveal a clear hierarchy of component importance. LoRA-based domain adaptation is the foundation, providing the largest single contribution. The boundary loss adds essential spatial precision for boundary-sensitive classes, and Mamba blocks provide incremental but meaningful refinement. Critically, only the full model achieves the best overall performance, as no single-component variant matches its combined mIoU and Kappa despite occasional per-class advantages.

5.3 Qualitative Analysis

Figure 2 provides a qualitative comparison of Hie-DinoMamba against the baseline models on several challenging test examples. These visual results complement the quantitative data in Table 1, offering insights into the model's practical performance, particularly in complex scenarios and on difficult-to-classify objects.

Row (A) displays a demolition example. While the 'Demolition' class shows a modest IoU score (45.84%) in our quantitative results, the qualitative analysis reveals that Hie-DinoMamba identifies the demolished buildings accurately. It produces clean segmentation masks with significantly fewer false positives or missed detections compared to other methods like SNUNet. This suggests an excellent detection capability for this class that may not be fully captured by the IoU metric alone.

Row (B) highlights the model's superior semantic understanding and context differentiation. The T1 and T2 images show

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-601-2026 | © Author(s) 2026. CC BY 4.0 License.

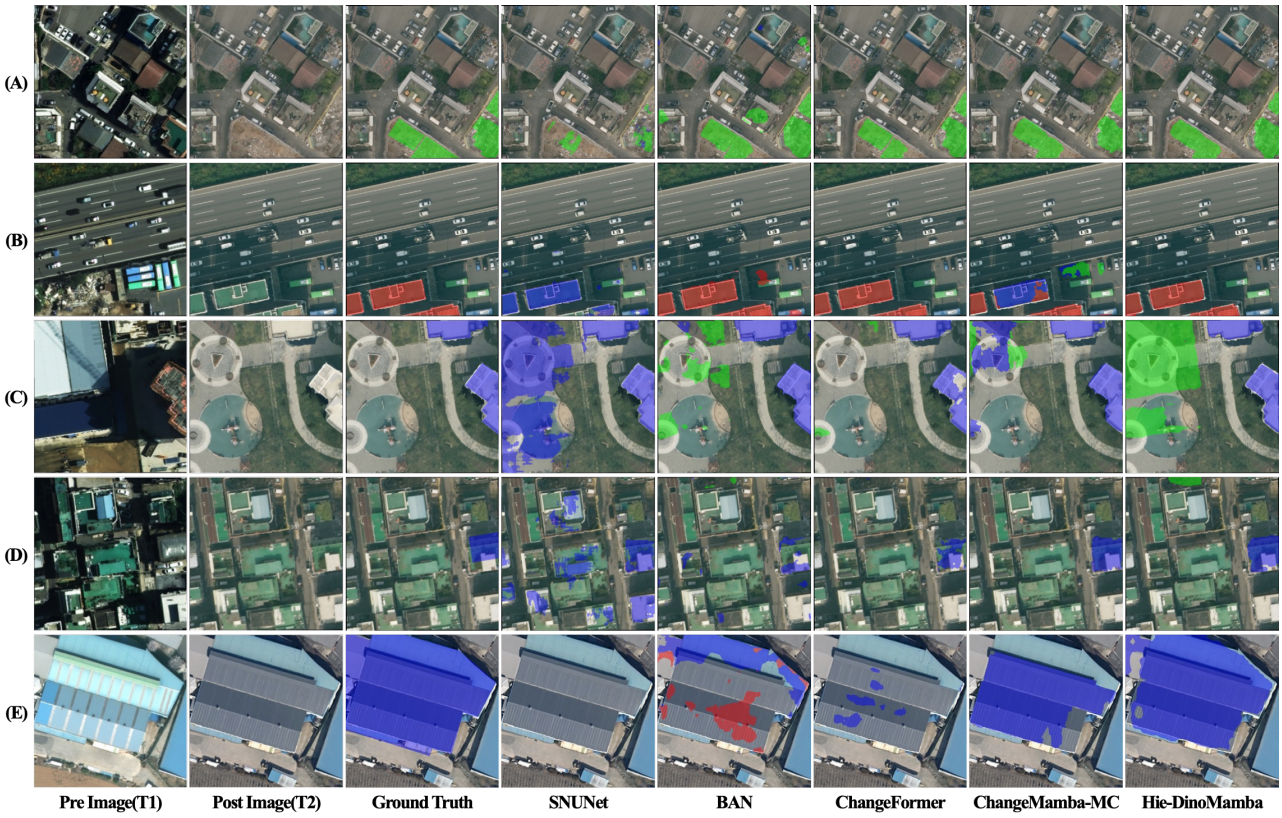

Figure 2. Qualitative comparison of Hie-DinoMamba against baseline models. Rows (A) to (E) visualize results across various challenging scenarios, including demolition and new construction. The visualization colors correspond to: Black (No Change), Red (New Construction), Green (Demolition), and Blue (Renovation).

multiple vehicles, including buses that can be easily confused with building rooftops. Hie-DinoMamba robustly distinguishes these non-building objects and correctly identifies only the ’New Construction’ (red) pixels, demonstrating a very high level of precision in complex urban environments.

Rows (C) and (D) illustrate our model’s superior sensitivity. These cases suggest that some label omissions may exist in the ground truth annotations. In Row (C), Hie-DinoMamba is the only model to successfully detect a large demolished area (green) that other models, including the baseline ChangeMamba-MC, failed to identify. Similarly, in Row (D), our model uniquely detects a single demolition object in addition to the main target. This superior sensitivity in correctly identifying real-world changes indicates that the model’s true performance is likely higher than the quantitative metrics reflect, as the model is penalized during evaluation for correctly identifying changes that the ground truth itself missed.

Finally, Row (E) demonstrates the model’s ability to handle a completely unseen geographic domain. We use an extra test sample from Asan-si, a city outside the training area and not included in any of the official splits. The scene features a large, complex renovation object that most baseline models fail to detect or only partially identify. Hie-DinoMamba, however, produces a detection map that closely aligns with the real-world change. This result provides strong empirical evidence that the pre-trained DINOv3-L features enable effective generalization to novel geographic areas, overcoming the limitations of models trained only on a specific domain.

6. Conclusion

In this paper, we addressed the significant challenge of multi-class (4-class) building change detection in high-resolution aerial imagery. We identified a key limitation in prior work: encoders trained from scratch on specific aerial datasets lack the rich, generalizable representations of large-scale Visual Foundation Models (VFMs).

To overcome this, we proposed Hie-DinoMamba, a novel architecture that successfully integrates a 1.1B parameter, pre-trained DINOv3-L encoder with an efficient Mamba FPN decoder. We demonstrated that by keeping the VFM backbone frozen and adapting it to the aerial domain using parameter-efficient LoRA, we can leverage its powerful feature extraction capabilities without incurring prohibitive computational costs. Our hierarchical decoder, optimized with a dual-loss strategy, proved effective in translating these features into precise, multi-class change maps.

Our experimental results demonstrate the effectiveness of this approach. Hie-DinoMamba achieved a new state-of-the-art mIoU of 65.12% on the 4-class benchmark, representing a significant 2.1 percentage points improvement over the strong VSSM-based baseline, ChangeMamba-MC (63.02% mIoU). Our ablation study further validated the contribution of each component: LoRA adaptation proved most critical (-4.17 percentage points when removed), confirming that bridging the domain gap between satellite pre-training and aerial imagery is essential. The dual-loss strategy and Mamba decoder each provided complementary improvements, with the boundary loss particularly benefiting spatially precise classes like Renovation.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-601-2026 | © Author(s) 2026. CC BY 4.0 License.

Furthermore, our qualitative analysis highlighted the model's superior generalization capabilities, showing its ability to successfully detect complex changes in geographically unseen regions (Asan-si) where other models failed.

Despite these strong results, limitations remain. The 'Demolition' class, while qualitatively strong, still presents a significant challenge in quantitative metrics (45.84% IoU). As suggested by our analysis, this may be partially related to potential label omissions in the ground truth data, which warrants further investigation. Future work could focus on refining the detection of challenging minority classes like 'Demolition', perhaps by integrating more advanced boundary-refinement techniques or exploring training methods robust to label noise. Additionally, applying the Hie-DinoMamba framework to other dense remote sensing tasks, such as land-cover change, presents a valuable direction for future research.

In conclusion, Hie-DinoMamba provides an effective and efficient framework for multi-class change detection, successfully bridging the gap between large-scale foundation models and specialized aerial analysis.

7. Acknowledgements

This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant R&D program of Digital Land Information Technology Development funded by the Ministry of Land, Infrastructure and Transport (MOLIT) (Grant RS-2022-00142501).

This research used datasets from ‘The Open AI Dataset Project (AI-Hub, S. Korea)’. All data information can be accessed through ‘AI-Hub (www.aihub.or.kr/)’.

References

Aguilar-Ruiz, J. S., Michalak, M., 2024. Classification performance assessment for imbalanced multiclass data. *Scientific Reports*, 14(1), 10759.

Chen, H., Song, J., Han, C., Xia, J., Yokoya, N., 2024. Changemamba: Remote sensing change detection with spatiotemporal state space model. *arXiv preprint arXiv:2404.03425*.

Chen, Z., Zhang, W., Lu, T., Chen, H., Li, J., 2022. A Transformer-based Siamese network for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

Dong, S., Wang, L., Du, B., Meng, X., 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 53–69.

Fang, S., Li, K., Shao, Z., Li, Z., 2021. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5.

Gu, A., Dao, T., 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Khelifi, L., Mignotte, M., 2020. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access*, 8, 126385–126400.

Li, K., Cao, X., Meng, D., 2023. A New Learning Paradigm for Foundation Model-based Remote Sensing Change Detection. *arXiv preprint arXiv:2312.01163*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2980–2988.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, IEEE, 565–571.

Mishra, S., Shah, A., Bansal, A., Jagannatha, A., Anjaria, J., Sharma, A., Krishnan, D., 2021. Object-aware cropping for self-supervised learning. *arXiv preprint arXiv:2112.00319*.

National Information Society Agency, 2023. AI-Hub: Aerial Building Change Detection Dataset. https://aihub.or.kr/. Accessed: 2025-11-01.

Salehi, S. S. M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *International Workshop on Machine Learning in Medical Imaging*, Springer, 379–387.

Saputra, H., Helmi, R. R. A., Ghazali, M. D., Sumartini, W. O., 2025. Urban Resilience through IoT-Based Disaster Preparedness and Infrastructure Monitoring: A Systematic Literature Review. *Natural Hazards Research*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P., 2025. DINOv3. *arXiv preprint arXiv:2508.10104*.

Wang, J., Song, J., Zhang, H., Zhang, Z., Ji, Y., Zhang, W., Zhang, J., Wang, X., 2025. SPMNet: A Siamese Pyramid Mamba Network for Very-High-Resolution Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 4410314.

Zhao, X., Xia, N., Li, M., 2023. Dynamic monitoring of urban renewal based on multi-source remote sensing and POI data: A case study of Shenzhen from 2012 to 2020. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103586.

Zhu, Q., Guo, X., Li, Z., Li, D., 2024. A review of multi-class change detection for satellite remote sensing imagery. *Geospatial Information Science*, 27(1), 1–15.