Urban Change Detection Using a Siamese DINOv3 Vision Transformer with Multi-Depth Fusion and a Lightweight Refinement Decoder

Bekzod Kadirov (s333564) Temurbek Karimov (s333565)
Politecnico di Torino

***Abstract*—Urban change detection from satellite imagery is a challenging task due to strong class imbalance, visual variability across acquisition times, and the need to generalize across different urban environments. In this work, we propose a deep learning framework for binary change detection based on a Siamese Vision Transformer architecture. Pre-change and post-change images are processed using a shared DINOv3-pretrained backbone, and multi-depth feature representations are fused to capture both high-level semantic information and fine-grained spatial details. A lightweight refinement decoder is employed to aggregate the fused features and produce dense pixel-wise change predictions. The model is trained and evaluated on the ONERA Satellite Change Detection dataset using a city-level split and patch-based processing. Experimental results demonstrate that the proposed approach achieves strong overall performance and robust background discrimination while effectively detecting sparse urban changes.**

***Index Terms*—Change detection, remote sensing, Vision Transformer, DINOv3, OSCD, refinement decoder, focal loss.**

I. Introduction

Urban environments undergo continuous transformations due to construction, infrastructure development, and land-use change. Detecting such changes from satellite imagery is essential for urban planning, environmental monitoring, and damage assessment. However, urban change detection remains difficult due to (i) severe foreground-background imbalance, (ii) seasonal and radiometric variation, and (iii) the requirement to generalize across cities.

The ONERA Satellite Change Detection (OSCD) dataset is a widely used benchmark for urban change detection. Earlier baselines rely on Siamese convolutional networks and can struggle to capture long-range context and to localize sparse changes precisely. In this project, we investigate a Siamese architecture based on a self-supervised Vision Transformer backbone and a decoder tailored to the fixed-resolution nature of ViT features.

The remainder of the paper is organized as follows: Section II reviews related work, Section III presents the proposed method and explored variants, Section IV describes experiments and results, and Section V concludes.

II. Related Work

Daudt et al. introduced fully convolutional Siamese networks for OSCD with early/late fusion variants. These CNN-based models can report high overall accuracy, yet may underperform on overlap-based metrics due to class imbalance and limited global context.

Vision Transformers have become strong feature extractors for dense prediction tasks. Self-supervised learning (e.g., DINO-style pretraining) improves transferability when labeled data is limited, which is typical in remote sensing.

Decoder design strongly impacts localization. FPN-style decoders aggregate features across multiple spatial resolutions and work best when the encoder naturally provides a spatial pyramid (as in CNNs). Mask2Former uses masked-attention decoding and has shown strong segmentation performance, but it is comparatively heavy and often benefits from large-scale training. When encoder features share the same spatial resolution (as in standard ViTs with fixed patch grids), lightweight refinement decoders that fuse multi-depth features can be better aligned with the representation structure and computational budget.

III. Method

A. Problem Formulation

Given a pre-change image $I^{pre} \in \mathbb{R}^{H\times W\times 3}$ and a post-change image $I^{post} \in \mathbb{R}^{H\times W\times 3}$, the goal is to predict a binary change mask $M \in \{0,1\}^{H\times W}$.

B. Base Architecture: Siamese DINOv3 + Multi-Depth Fusion + Refinement Decoder

We use a Siamese (weight-sharing) encoder: both images are passed through a shared Vision Transformer backbone. From multiple depths of the Transformer, we extract feature maps (F_s^{pre}, F_s^{post}) for $s \in \{1,2,3\}$. In our implementation, all extracted feature maps have the same spatial resolution due to the fixed patch-token grid of the ViT (e.g., $256/16 = 16$ tokens per side).

For each extracted depth s, we compute a fusion representation using concatenation of $(F_s^{pre}, F_s^{post}, |F_s^{pre} - F_s^{post}|)$ and apply a learnable projection to a fixed 256-channel space. All projected depth-wise fused features are then concatenated and passed into a lightweight refinement decoder consisting of convolutional refinement blocks and a 1×1 prediction head. Finally, logits are upsampled to the original image resolution.

C. Backbone

We use the pretrained model `vit_large_patch16_dinov3.sat493m`.

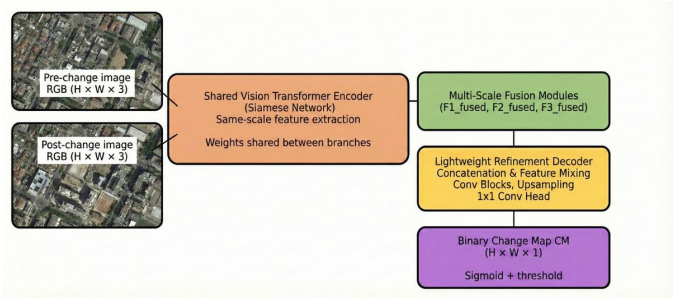

Fig. 1: Final model: Siamese DINOv3 backbone with multi-depth fusion and a lightweight refinement decoder.

D. Training Schedule and Loss

We adopt a two-stage fine-tuning schedule: (1) freeze the backbone for **20 epochs**, then (2) unfreeze only the **last 2 ViT blocks** (and final norm) with a smaller backbone learning rate. We use **focal loss** to mitigate class imbalance and reduce the dominance of easy background pixels.

E. Explored Alternatives (Decoders, Fusion, Losses)

Beyond the final configuration, we explored multiple decoder, fusion, and loss variants to identify the most robust design for OSCD.

a) Mask2Former Decoder: We evaluated Mask2Former as a drop-in decoder using BCE loss and full backbone unfreezing. Although Mask2Former achieved high overall accuracy, training was less stable and the model tended to suppress sparse foreground changes, leading to low overlap-based scores (IoU/F1). This behavior is consistent with the combination of extreme imbalance and the higher capacity of masked-attention decoding, which can require stronger regularization and more data to generalize.

b) FPN Decoder: We also implemented an FPN-style decoder operating on multi-depth ViT features. While FPN improved over Mask2Former in overlap-based metrics, it is conceptually designed for backbones that provide a true spatial pyramid. In our setting, all extracted ViT feature maps share the same patch-grid resolution; thus, the benefits of top-down multi-resolution fusion are reduced. This mismatch increases complexity without clearly outperforming the refinement decoder.

c) Single-Scale Fusion: As a baseline, we tested single-scale decoding using only the final transformer block. This configuration underperformed multi-depth fusion, indicating that intermediate representations carry complementary semantic cues useful for localizing small changes.

d) Loss Functions: BCE vs Dice+BCE vs Focal: We compared three loss formulations: (i) BCE, (ii) Dice+BCE, and (iii) Focal loss. BCE converged quickly but strongly favored background pixels, often producing conservative masks. Dice+BCE improved sensitivity to changes (higher recall) but introduced noisier predictions and less stable optimization. Focal loss provided the best balance by down-weighting easy negatives, yielding more stable training and improved overlap-based metrics after threshold tuning.

IV. Experiments and Results

A. Dataset Split and Patch Processing

We follow a city-level train/validation/test split to measure generalization across unseen cities. Images are processed in overlapping patches with:

$$\texttt{patch\_size} = 256, \quad \texttt{stride} = 128.$$

B. Augmentation

We apply paired augmentations (same transform to pre/post) during training: random flips, rotations, and rescaling/cropping. No augmentation is used for validation/testing.

C. Optimization Setup

We train for **100 epochs** with batch size **4** using AdamW. Learning rates:

$$\texttt{lr\_decoder} = 3 \times 10^{-5}, \quad \texttt{lr\_backbone} = 5 \times 10^{-7}.$$

D. Threshold Tuning

We tune the sigmoid threshold on the validation set by maximizing F1 across a grid of thresholds, then fix the chosen threshold for test evaluation to avoid test leakage.

E. Decoder and Training Strategy Ablation

TABLE I: Decoder and training strategy comparison. Mask2Former is a masked-attention decoder, FPN is a pyramid-style decoder (less aligned with fixed-resolution ViT features), and Refinement is our lightweight depth-fusion decoder.

Method	OA	IoU	F1	Prec.	Rec.
Mask2Former (BCE, full unfreeze)	0.9405	0.2039	0.3388	0.4282	0.2803
FPN (Focal, partial unfreeze)	0.9484	0.3602	0.5297	0.5250	0.5344
Refinement (Focal, partial unfreeze)	0.9439	0.3536	0.5225	0.4863	0.5644

Table I reports the quantitative comparison of decoder architectures and training strategies. Mask2Former achieves high overall accuracy but substantially lower overlap-based metrics (IoU and F1), indicating strong bias toward background predictions. Both FPN and refinement decoders significantly improve overlap-based performance. FPN yields the highest IoU and F1, while the refinement decoder attains slightly lower overlap scores but higher recall, suggesting improved sensitivity to sparse changes.

F. Why We Chose the Refinement Decoder (Final Design)

The ablation results in Table I indicate that decoder complexity alone does not guarantee better change detection performance. Although Mask2Former is powerful for large-scale semantic segmentation, it proved less stable in our setting and biased predictions toward background regions, resulting in poor overlap-based metrics. The FPN decoder improves localization performance but is conceptually designed for multi-resolution encoder pyramids, whereas our Vision Transformer backbone produces features on a fixed patch grid. The refinement decoder is better aligned with ViT representations, as it directly fuses complementary depth-wise semantic information with low computational overhead. In combination with focal loss and partial backbone unfreezing, this design achieves a strong

balance between background discrimination and sensitivity to sparse urban changes, motivating its selection as the final architecture.

After depth-wise fusion and projection to a shared 256-channel space, the refinement decoder aggregates the resulting change-aware features by channel-wise concatenation followed by a 1×1 mixing layer. A small stack of 3×3 convolutional refinement blocks then enforces spatial consistency on the fixed ViT grid, acting as a lightweight denoising and boundary-regularization stage. A final 1×1 head produces change logits, which are upsampled to the input resolution.

G. Comparison with OSCD Documentation (Accuracy-Based Metrics)

Table II compares accuracy-based metrics with values reported in the OSCD documentation. We report Overall Accuracy (OA), ChangeAcc, and NoChangeAcc. The OSCD documentation includes CNN baselines under different input channel settings; for a fair RGB comparison we include the 3-channel baselines, and we also report the best multi-spectral configuration (10-channel Early Fusion) from the table.

TABLE II: Accuracy-based comparison with OSCD documentation results

Method	Input	OA (%)	ChangeAcc (%)	NoChangeAcc (%)
OSCD Siamese CNN (3 ch.)	RGB	84.13	78.57	84.43
OSCD Early Fusion CNN (3 ch.)	RGB	83.63	82.14	83.71
OSCD Early Fusion CNN (10 ch.)	Multi-spec.	89.15	82.75	89.50
Siamese DINOv3 + Refinement (ours)	RGB	**94.39**	56.44	**96.57**

Despite using only RGB inputs, our model achieves substantially higher overall accuracy and no-change accuracy than the CNN baselines reported in the OSCD documentation. The change accuracy remains lower than the CNN baselines, reflecting a conservative decision boundary under severe imbalance and the trade-off induced by optimizing overlap-based metrics and validation-based threshold tuning.

H. Final Test Metrics (Refinement Decoder)

For completeness, the final test evaluation of the refinement decoder model is: IoU=0.3536, F1=0.5225, Precision=0.4863, OA=0.9439, ChangeAcc=0.5644, and NoChangeAcc=0.9657.

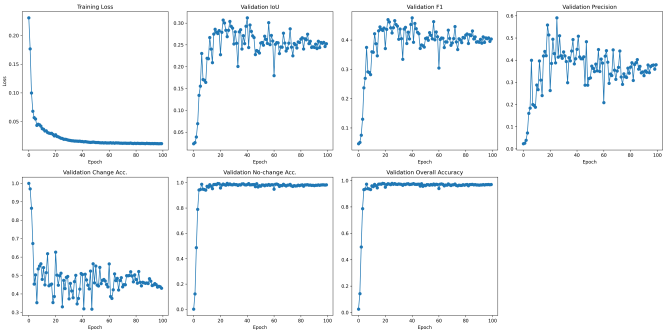

Fig. 2: Final model training curves.

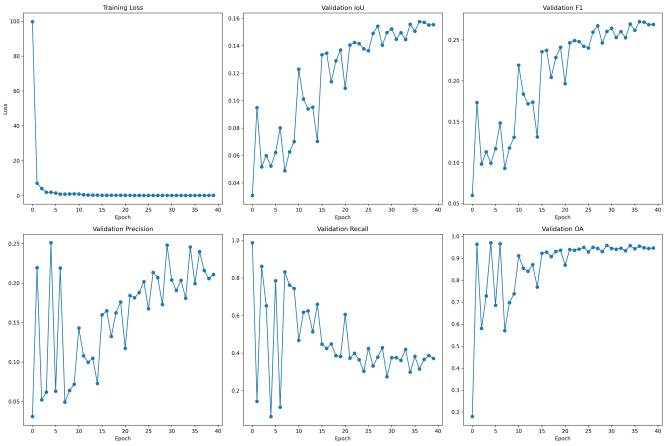

Fig. 3: Mask2Former training curves.

V. Conclusion

We presented a Siamese change detection system using a DINOv3-pretrained ViT encoder with multi-depth fusion and a lightweight refinement decoder. Compared to Mask2Former, the refinement decoder provides a simpler and more stable alternative in the setting where extracted ViT features share the same spatial resolution and the dataset is limited and highly imbalanced. Compared to OSCD documentation baselines, our model achieves higher overall accuracy and no-change accuracy using RGB inputs, while change accuracy highlights an ongoing trade-off between sensitivity to sparse changes and background discrimination. Future work includes per-city error analysis, exploration of hierarchical backbones, and extension toward multi-class change detection.

VI. Future Work

Future work may explore alternative decoder architectures that further exploit transformer representations, including hybrid convolution–transformer decoders and attention-based refinement modules. Additional fusion strategies, such as adaptive depth selection or temporal attention mechanisms, could be investigated to better capture subtle urban changes. Extending the framework to multi-class change detection and incorporating multi-spectral inputs are also promising directions. Finally, more detailed per-city analysis and domain adaptation techniques could be studied to further improve generalization across diverse urban environments.

References

[1] R. C. Daudt, B. Le Saux, and A. Boulch, "Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks," in *IGARSS*, 2018.

[2] T.-Y. Lin *et al.*, "Feature Pyramid Networks for Object Detection," in *CVPR*, 2017.

[3] T.-Y. Lin *et al.*, "Focal Loss for Dense Object Detection," in *ICCV*, 2017.

[4] A. Kirillov *et al.*, "Mask2Former: Masked-attention Transformer for Universal Image Segmentation," in *CVPR*, 2022.

[5] M. Caron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," in *ICCV*, 2021.