Urban Change Detection Using a Siamese DINOv3 Vision Transformer with Multi-Scale Feature Pyramid Decoding

Bekzod Kadirov s333564 Temurbek Karimov s333565
Politecnico di Torino

***Abstract*—Urban change detection from satellite imagery is challenging due to strong class imbalance and large visual variability across cities. The objective of this work is to design a robust deep learning model capable of accurately detecting urban changes in multi-temporal remote sensing images. We address this challenge using a Siamese Vision Transformer architecture with a DINOv3-pretrained backbone and a multi-scale Feature Pyramid Network decoder. Experiments are conducted on the ONERA Satellite Change Detection dataset using a city-level train/validation/test split and patch-based processing. The proposed model achieves high overall accuracy and substantially improves change accuracy compared to an alternative decoder and training strategy. These results demonstrate that combining self-supervised Vision Transformers with multi-scale decoding is effective for urban change detection.**

***Index Terms*—Change detection, remote sensing, Vision Transformer, DINOv3, FPN, OSCD.**

I. Introduction

Urban environments undergo continuous transformations due to construction, infrastructure development, and land-use changes. Detecting such changes from satellite imagery is essential for applications including urban planning, environmental monitoring, and damage assessment. However, urban change detection remains difficult due to severe class imbalance, acquisition/seasonal variation, and the need to generalize across cities.

The ONERA Satellite Change Detection (OSCD) dataset is a widely used benchmark for urban change detection. Early baselines rely on Siamese convolutional networks and often struggle to capture long-range context and to localize sparse changes precisely. In this project, we explore a Siamese architecture based on a self-supervised Vision Transformer backbone and multi-scale decoding.

The remainder of the paper is organized as follows: Section II reviews related work, Section III presents the method, 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 report high overall accuracy but can underperform on overlap-based metrics due to class imbalance and limited global context.

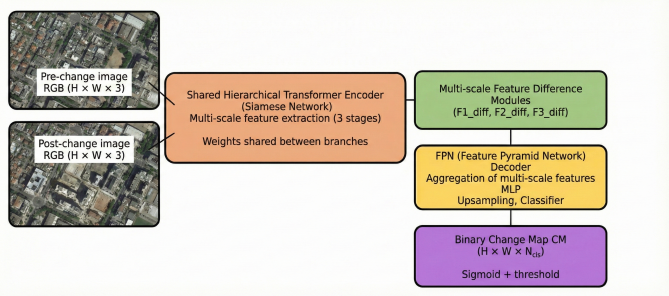


Fig. 1: Siamese DINOv3 + FPN pipeline (backbone treated as a black-box multi-scale encoder).

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 choice matters for localization. FPN-style decoders aggregate features across resolutions and have proven effective in many dense prediction tasks. Mask2Former provides a powerful masked-attention decoding strategy but can be more sensitive to optimization and data regime.

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. Architecture

We use a Siamese (weight-sharing) encoder: both images are passed through a shared Vision Transformer backbone producing multi-scale features. At each scale $s \in \{1,2,3\}$, we obtain a pair (F_s^{pre}, F_s^{post}) that is fused per scale and projected to a 256-channel representation. The resulting multi-scale fused features are decoded by an FPN decoder (top-down + lateral merges) to full resolution, followed by a 1×1 segmentation head producing logits of shape $[B,1,H,W]$.

C. Backbone

We use the pretrained model `vit_large_patch16_dinov3.sat493m`.

TABLE I: Decoder and training strategy comparison

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

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 + FPN (ours)	RGB	**94.84**	53.44	**97.22**

D. Training Strategy and Loss

Training uses a two-stage fine-tuning schedule: (1) freeze backbone for **20 epochs**, then (2) unfreeze only the **last 2 ViT blocks** (and final norm) with a smaller backbone learning rate. The final configuration uses **focal loss** to mitigate class imbalance. For comparison, we trained a Mask2Former-style decoder with **BCE loss** and a full-unfreeze schedule after 20 epochs.

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

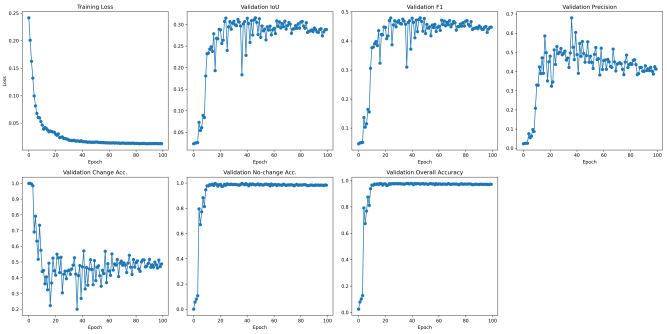
We tune the sigmoid threshold on the validation set by maximizing F1 across a grid of thresholds, then fix the chosen threshold for test evaluation to avoid test leakage.

E. Decoder and Training Strategy Ablation

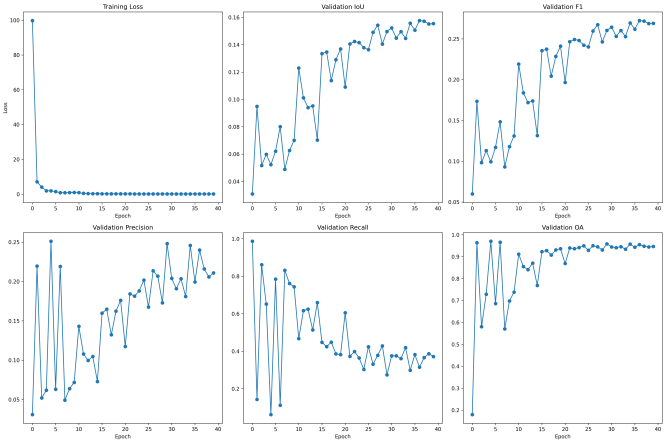
F. Comparison with OSCD Documentation (Accuracy-Based Metrics)

Table II compares our accuracy-based results 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.

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 lower change accuracy indicates a conservative decision boundary and reflects optimization under severe class imbalance, where background pixels dominate and overlap-based criteria (F1/IoU) are emphasized during validation threshold tuning.



(a) FPN-based model.



(b) Mask2Former-based model.

Fig. 2: Training and validation curves for both decoder configurations.

G. Training Dynamics

To avoid float-only pages in a two-column layout, we combine both training-curve plots into a single figure.

H. Take-Home Message

The FPN decoder with focal loss and partial unfreezing improves stability and change localization compared to the Mask2Former + BCE configuration. Compared to OSCD documentation baselines, our approach achieves higher OA and NoChangeAcc with RGB inputs, while ChangeAcc highlights a trade-off between background discrimination and sensitivity to sparse changes.

V. Conclusion

We presented a Siamese change detection system using a DINOv3-pretrained ViT encoder with multi-scale fusion and an FPN decoder. Experiments on OSCD show strong overall accuracy and strong background discrimination, and the ablation study demonstrates that focal loss and partial unfreezing are more stable than full fine-tuning with BCE in our setting. Future work includes per-city error analysis, alternative fusion operators, and extension to multi-class change detection.

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.