Environmental change detection WITH DINOv3
(Using OSCD Dataset)

Karimov Temurbek
Kadirov Bekzod

Project Proposition Value

- Our project proposes an automated change detection system that compares satellite images from different years to identify how landscapes, infrastructure, and urban areas have evolved over time.
- This solution enables analysts, researchers, and local authorities to efficiently monitor long-term changes without manual inspection.

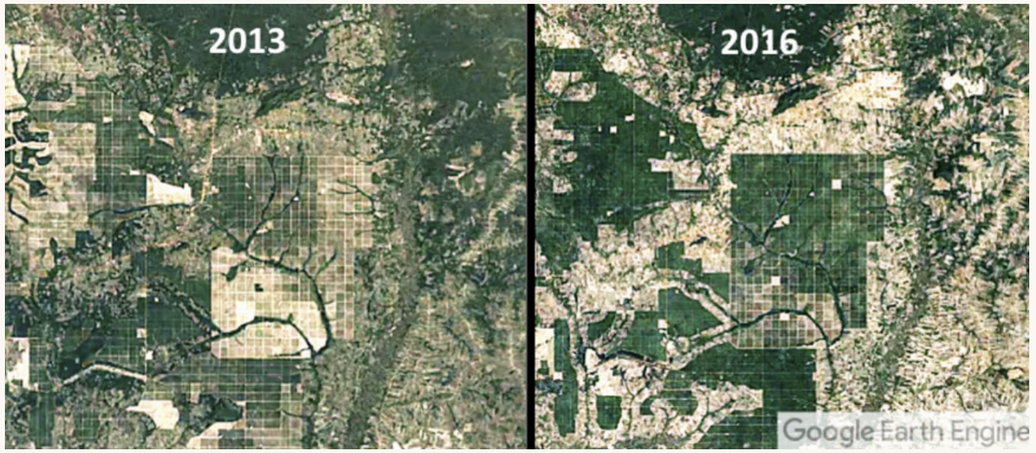
Value Provided

- Automated detection of year-to-year changes in buildings, vegetation, land use, and infrastructure
- Improved accuracy using transformer-based features, including DINOv3
- Faster analysis compared to manual comparison of satellite images
- Scalable monitoring for large geographic regions
- Supports data-driven decision-making for urban planning, environmental studies, and disaster prevention

Why This Problem Matters ?

Urban planning: Cities expand, infrastructure evolves, new buildings appear.

Environmental monitoring: Forest loss, agricultural expansion, water-level changes.



Infrastructure or landscape transformation over years

Primary Goal: Develop an automated end-to-end pipeline for multi-temporal satellite image change detection using the state-of-the-art DINOv3 vision transformer.

Technical Objectives:

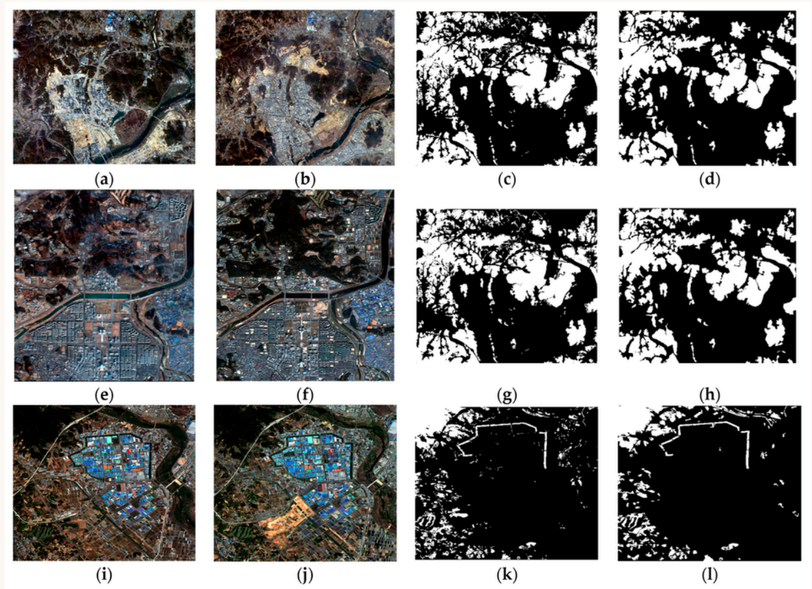
- Feature Extraction: Leverage self-supervised pre-trained DINOv3 features to capture subtle environmental and structural transformations.
- Accuracy Enhancement: Achieve higher precision in identifying urban expansion and deforestation compared to traditional CNN-based methods.
- Scalability: Ensure the model can process large-scale datasets (like OSCD) efficiently.

Impact Objective: Provide a robust tool for urban planners and environmental researchers to monitor land-use changes with minimal manual intervention.

Research Questions & Hypotheses

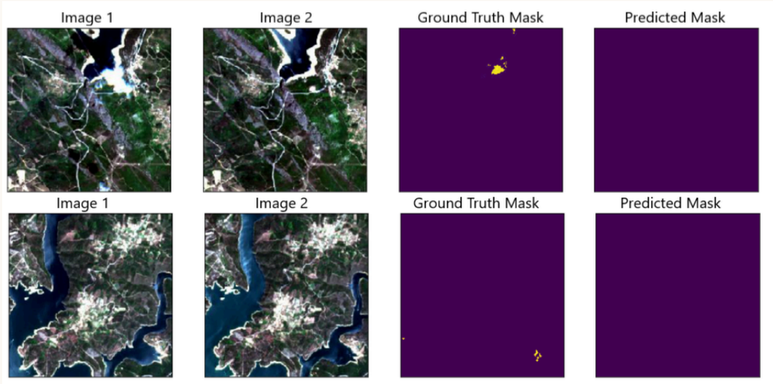
Research Questions

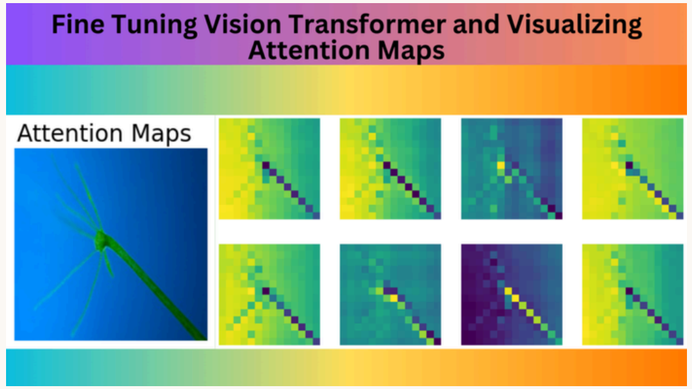
- RQ1 — Can DINOv3 detect subtle changes?
- RQ2 — Can we detect different types of changes?
- RQ3 — Does multi-year comparison improve consistency?
- RQ4 — Which fusion method works best?
- RQ5 — Can the model generalize to different regions?

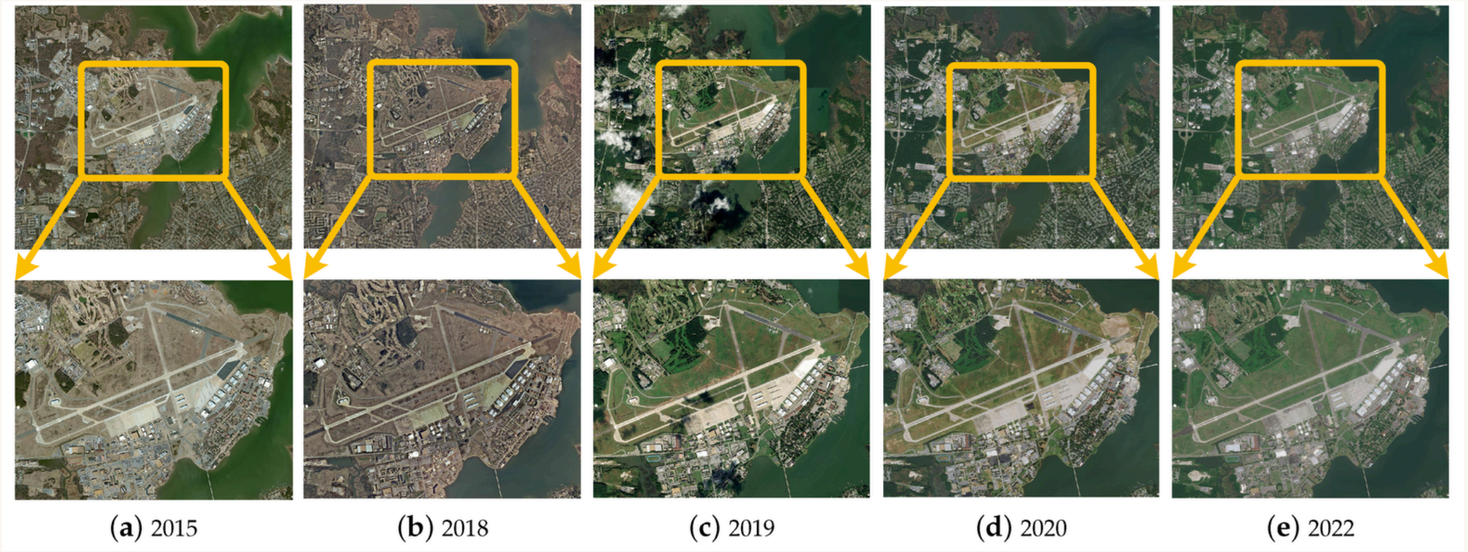


Hypotheses

- H1: DINOv3 features improve sensitivity to fine-grained structural and land-cover changes.
- H2: A dual-branch or multi-temporal architecture outperforms single-image models for identifying long-term transformations.
- H3: Spatial augmentations and patch-based processing increase robustness to noise, lighting differences, and seasonal effects.







Methodology & Workflow

Methodology & Workflow

1.Data Preparation:
Utilizing the OSCD dataset to obtain multi-temporal satellite imagery

2. Preprocessing:
High-resolution satellite imagery is spatially aligned, normalized for lighting, and processed using 256x256 pixel patches with a 128-pixel overlapping stride to ensure precise feature extraction

3.Feature Extraction:
Implementing a dual-branch hierarchical transformer encoder based on DINOv3 to extract deep spatial features from pre-change and post-change images.

4.Feature Fusion:
Using a difference module to compare features from different years and highlight structural transformations.

5.Classification:
The fused multi-depth features are decoded through a convolutional segmentation head, producing a dense per-pixel change probability map via a 1×1 convolution and sigmoid activation.

OSCD Dataset samples used for training and testing (Ref. Brasilia city)

Pre-image

Post-image

Ground Truth

Development Choices & Techniques

Development Choices & Techniques

Initial Design Choices

- Starting with a Transformer-based encoder to capture global spatial context.
- A Mask2Former-style decoder was initially used for segmentation.
- Training used Binary Cross-Entropy loss and full backbone fine-tuning.

Issues Observed in Early Experiments

- Training was unstable when fully fine-tuning a large Vision Transformer.
- Binary Cross-Entropy loss was dominated by background pixels.
- Overlap-based metrics (IoU, F1) remained low despite high overall accuracy.

Backbone: Vision Transformer with DINOv3

- We use a large Vision Transformer pretrained using self-supervised DINOv3.
- Exact model:
- vit_large_patch16_dinov3.sat493m
- Self-supervised pretraining provides strong representations without dense labels.

Siamese Encoder Design

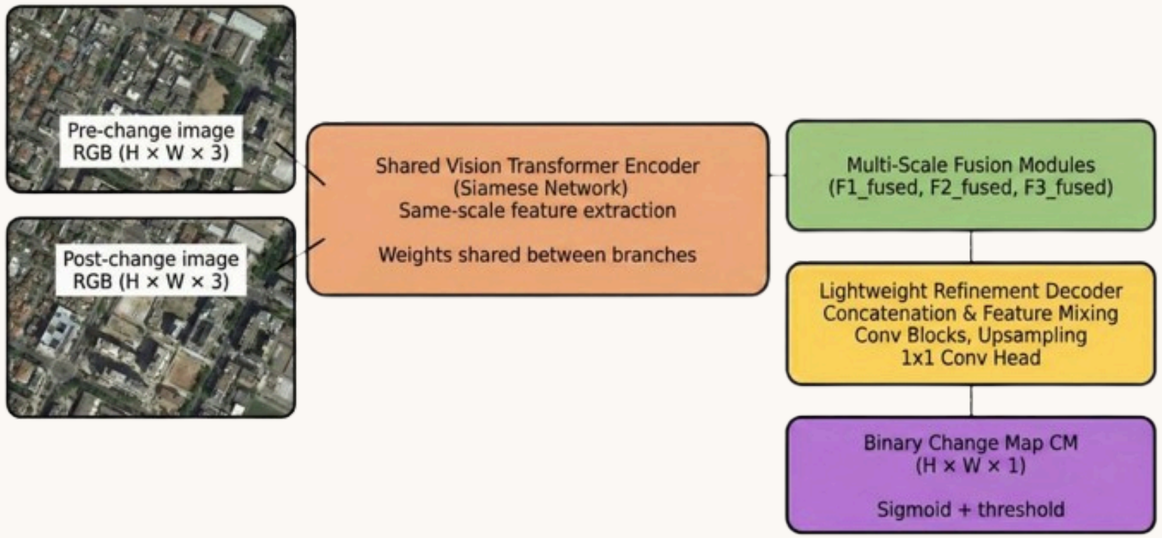
- Pre-change and post-change images are processed by the same backbone.
- Weights are shared to enforce temporal consistency.
- The backbone outputs multi-depth feature maps.

Development Choices & Techniques

• The refinement decoder aggregates depth-wise semantic change features into a single representation and progressively refines them to produce a clean binary change logit map.
• Feature maps from multiple backbone stages are gradually integrated across resolutions.
• High-level semantic information guides the refinement of finer spatial features.

• This enables precise pixel-level localization of subtle changes.
• Urban changes are often small and spatially sparse.
• Single-scale decoding leads to coarse or fragmented predictions.

• Aggregates depth-wise change-aware ViT features into a unified latent representation.
• Uses 1×1 channel mixing followed by lightweight 3×3 refinement blocks to improve spatial consistency. 3×3 refinement blocks to improve spatial consistency.
• Produces a binary change logit map via a 1×1 prediction head with late upsampling.



Evaluation and Training Strategy & Expected Outcomes

Loss Functions: BCE vs Focal Loss

- Binary Cross-Entropy treats all pixels equally.
- In change detection, background pixels dominate the loss.
- Focal loss down-weights easy negatives and focuses learning on hard change pixels.

Training Strategy Evolution

- Stage 1: freeze the backbone to stabilize early training.
- Stage 2: unfreeze only the last two Transformer blocks.
- This prevents catastrophic forgetting while allowing task-specific adaptation.

Threshold Tuning

- The model outputs a probability map via a sigmoid activation.
- A decision threshold is tuned on the validation set.

We select the threshold that maximizes the F1-score to balance precision and recall.

Expected Outcomes

- More accurate identification of long-term changes (urban expansion, deforestation, infrastructure growth)
- Cleaner and more interpretable change maps compared to classical methods
- Higher sensitivity to subtle transformations thanks to DINOv3 features
- Robust performance across years with varying lighting and seasonal artifacts
- Practical usefulness for analysts, planners, and environmental teams

Experiments & Results

Experiments & Results

Experimental Setup & Configuration

Parameter	Value	Purpose / Details
Backbone	vit_large_patch16_dinov3.sat493m.	Self-supervised pre-trained model for robust feature extraction.
Patch Size	256x256	Optimized for Vision Transformer spatial resolution.
Stride	128	Provides 50% overlap to ensure spatial consistency.
Optimization	AdamW	Standard for Transformers; used with 100 epochs.
Learning Rates	Decoder: 3x10^-5; Backbone: 5 x10^-7	Differential rates to protect pre-trained backbone features.
Loss Function	Focal Loss	Specifically used to mitigate severe class imbalance.
Strategy	Two-stage Fine-tuning	20 epochs frozen, then unfreeze last 2 ViT blocks.
Augmentation	Flips, Rotations, Rescaling	Paired transforms to reduce seasonal and illumination bias.

Comparative Results

Method	Overall Accuracy (OA)	Change Accuracy
OSCD Siamese CNN (Baseline)	84.13%	78.57%
OSCD Early Fusion CNN	83.63%	82.14%
Siamese DINOv3 + Refinement Decoder(Ours)	**94.84%**	**53.44%**

Experiments & Results

№	version	description	model name	num of epochs	freezed epochs	batch_size	initial_eval_thres	auto_eval_thres	Overall Accuracy	IoU score	F1 score	Precision	ChangeAcc	NoChange Acc	lr_decoder	lr_backbone	loss
1	m1	use Mask2Former decoder with	vit_large_patch16_dinov3.sat493m	40	10	2			0,9352	0,2704	0,4256				0,0001	0,0001	dice+bce
2	m2	use Mask2Former decoder with		40	0	2			0,9067	0,2023	0,3365				0,0001	0,0001	dice+bce
3	m3	use Mask2Former decoder with		80	20	2			0,9272	0,2838	0,4421				0,0001	0,0001	dice+bce
4	m4	use Mask2Former decoder with		100	20	4			0,9403	0,2713	0,4268				0,0001	0,0001	dice+bce
5	m5	use Mask2Former decoder with ordinary fusion method		80	15	4			0,947	0,2351	0,3808				0,0001	0,0001	dice+bce
6	m6	uses Mask2Former decoder with ordinary fusion method and utilizes new metrics.		40	10	4			0,9405	0,2039	0,3388	0,4282	0,4282		0,0001	0,00001	bce
7	m7	uses Mask2Former decoder with normalized fusion method.		100	20	4	0,35	0,26	0,9407	0,285	0,4436	0,4533	0,4533		0,0001	0,0000001	focal
8	m8	uses multi-depth backbone with FPN decoder and normalized multi-depth fusion		100	20	4	0,35	0,34	0,9519	0,3594	0,5287	0,5662	0,5662		0,00001	0,0000001	focal
9	m9	is the same as M8 but also utilizes new metrics.		100	20	4	0,35	0,36	0,9484	0,3602	0,5297	0,525	0,5344	0,9722	0,00001	0,0000001	focal
10	m10	uses multi-depth backbone with refinement decoder and normalized multi-depth fusion method.		100	20	4	0,35	0,34	0,9439	0,3536	0,5225	0,4863	0,5644	0,9657	0,00001	0,0000001	focal

Experiments & Results

- Compared to Mask2Former, the refinement decoder shows more stable optimization and higher sensitivity to sparse urban changes in our setting.
- While Mask2Former’s query-based masked attention is effective for large-scale segmentation, it exhibits a strong background bias under severe class imbalance, leading to lower overlap-based metrics.
- In contrast, the refinement decoder directly regularizes fixed-resolution, change-aware ViT features, providing a simpler and more robust solution for binary change detection.

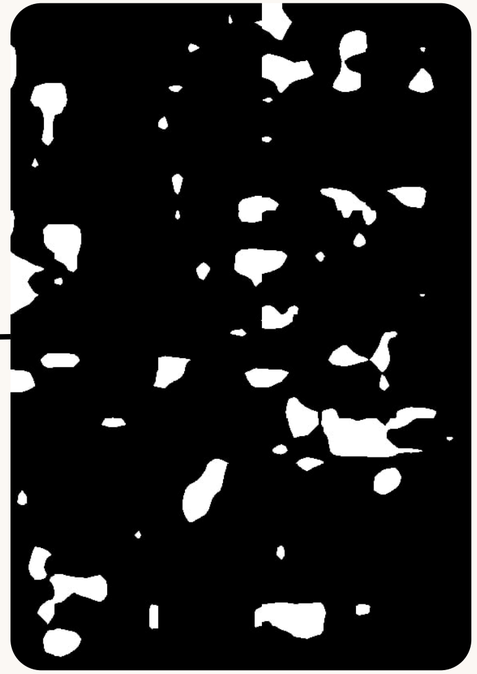
Examples of the results
(Ref. Las-Vegas city)

Pre-image

Post-image

Ground Truth

Prediction



Conclusions

- ***Model Effectiveness***:
  - Our Siamese DINOv3 + Refinement decoder architecture achieved a superior Overall Accuracy of 94.84%, proving highly effective for urban change detection.
- ***Technical Insight***:
  - The use of Focal Loss and a two-stage fine-tuning strategy (partial unfreezing) significantly improved training stability and handled class imbalance.
- ***Practical Impact***:
  - The system provides a reliable, automated pipeline for monitoring infrastructure development with minimal manual oversight.

Future Work

- ***Per-city Error Analysis:***
  - Conduct a deeper investigation into how different urban landscapes (e.g., Las Vegas vs. Paris) affect model precision.
- ***Multi-class Detection:***
  - Extend the model to distinguish between types of changes (e.g., new buildings vs. road construction).
- ***Alternative Fusion:***
  - Test different feature fusion operators to further improve the sensitivity to sparse changes.

Thank you for attention!