Region-Aware Token Aggregation for Fine-Grained Change Detection

Bianca-Cerasela-Zelia Blaga

Technical University of Cluj-Napoca
ICCP 2025

Contents

1. Introduction

- Change detection (CD) of remote sensing imagery means identifying meaningful land cover transitions at high spatial resolution [1]
- Critical for applications: agriculture, urban planning, disaster response, environmental monitoring [2]
- Deep learning approaches outperform traditional methods

Problem Statement

- Most CD networks rely on uniform patch-based tokenization
- This leads to:
  - Misalignment with real-world semantic boundaries (fields, parcels)
  - Loss of interpretability in region-level reasoning
  - Errors in boundary precision and small-structure detection
- Key challenge:
  - How to design a CD model that is both semantically coherent (region-level) and spatially precise (pixel-level)?

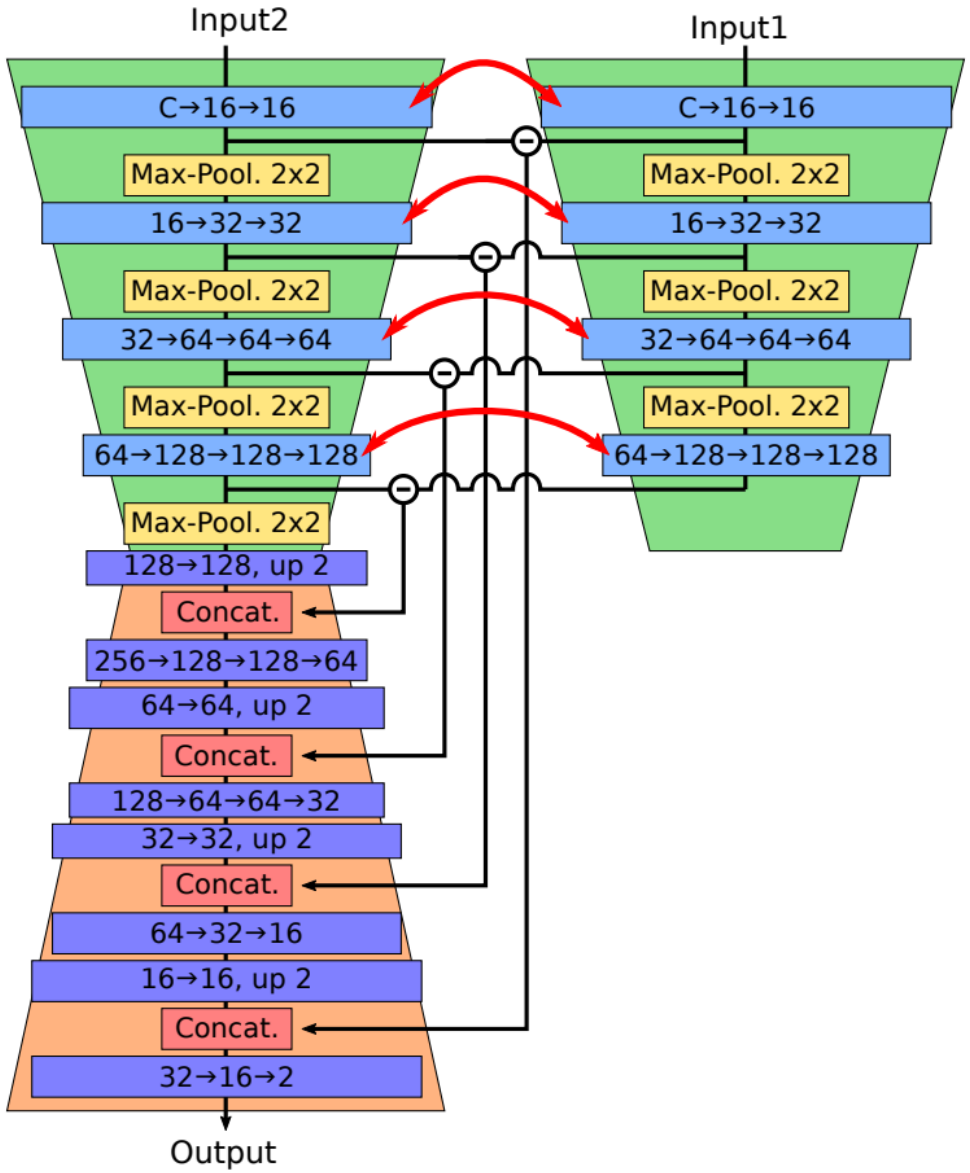
2. Related Work

Network Type	Examples	Strenghts	Limitations
CNN-based	SiamUNet-Diff, RDPNet, A2Net	pixel-level detail	weak long-range reasoning
Transformer-based	BIT, ChangeFormer	global context & token reasoning	uniform patch tokenization
Hybrid CNN-Transformer	ICIFNet, DMINet, SAAN	local–global feature fusion	still patch-based, not region-aware

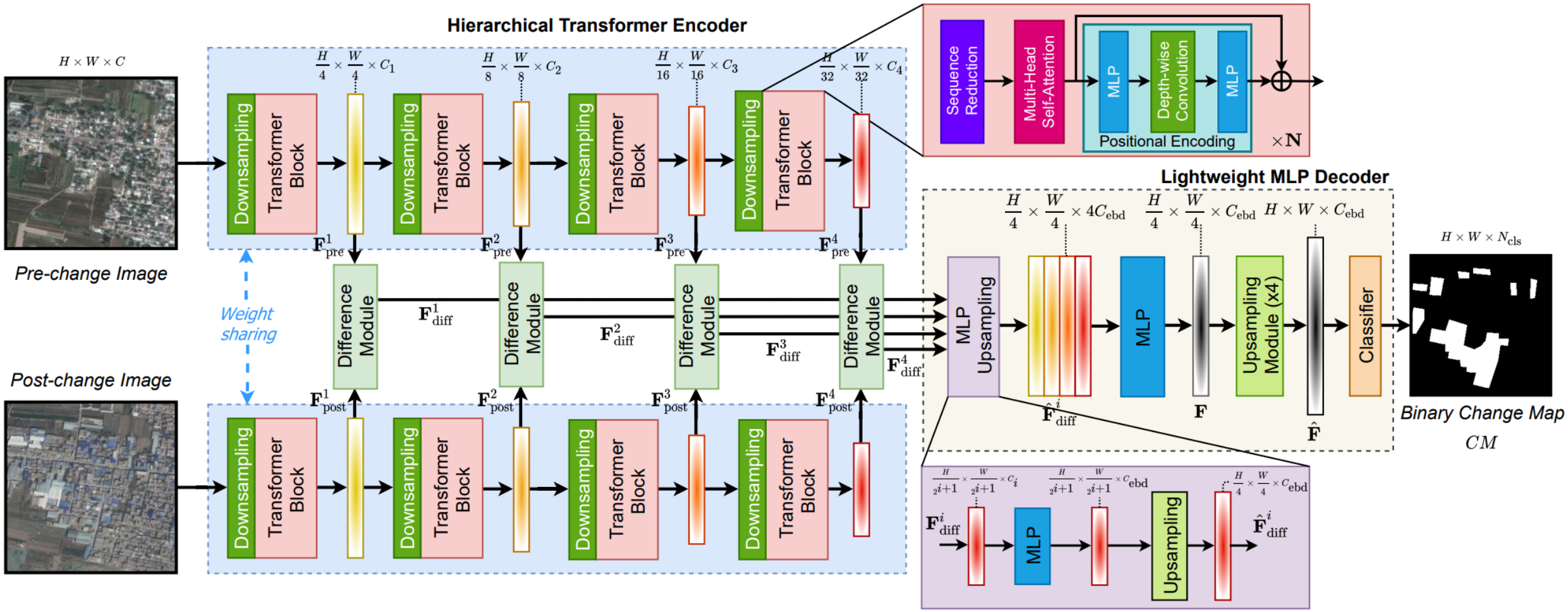
Gap:

- None explicitly leverage domain-specific semantic regions for tokenization + decoding

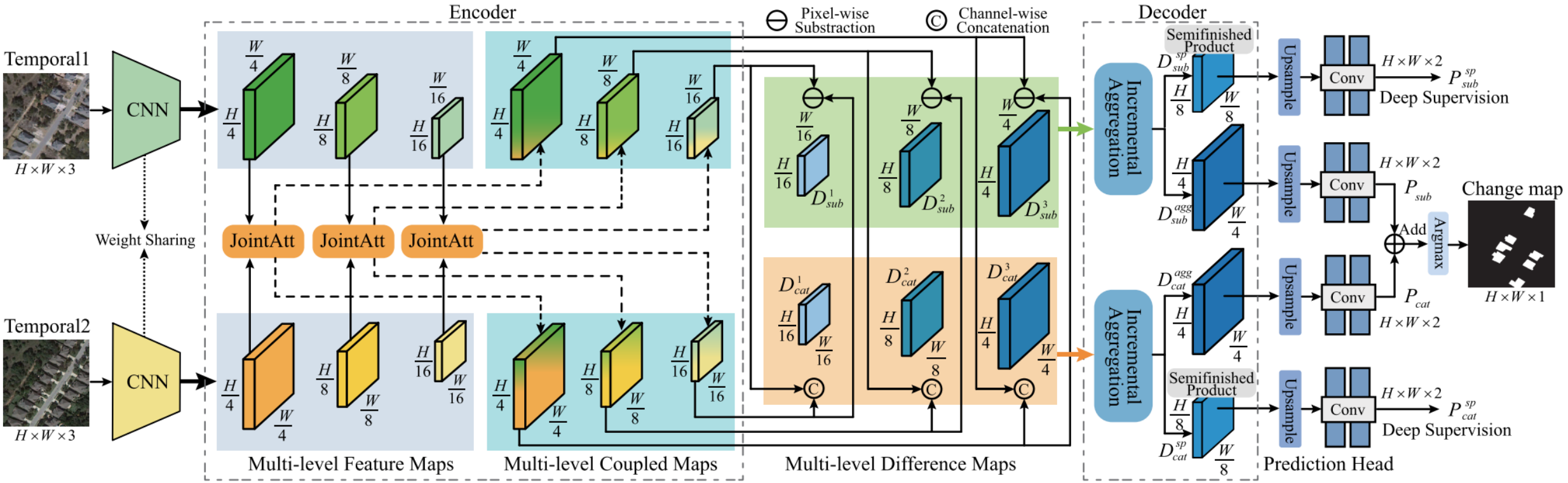
SiamUNet-Diff



ChangeFormer



DMINet

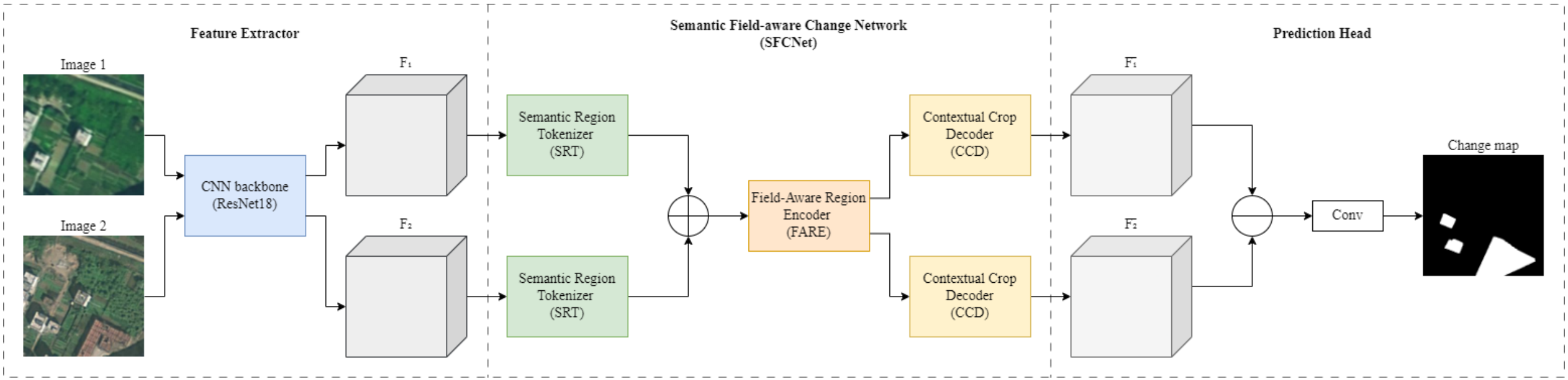

Encoder
Pixel-wise Substraction
Channel-wise Concatenation
Decoder
Temporal1
Temporal2
CNN
CNN
Weight Sharing
JointAtt
JointAtt
JointAtt
Multi-level Feature Maps
Multi-level Coupled Maps
Multi-level Difference Maps
Incremental Aggregation
Incremental Aggregation
Semifinished Product
Semifinished Product
Upsample
Upsample
Upsample
Upsample
Conv
Conv
Conv
Conv
Deep Supervision
Deep Supervision
Add
Argmax
Change map
Prediction Head

3. Methodology

- Propose SFCNet (Semantic Field-aware Change detection Network) with three key components:

1. Semantic Region Tokenizer (SRT)
   - Aggregates CNN features into region-level tokens using field masks
   - Produces compact & interpretable representations
2. Field-Aware Region Encoder (FARE)
   - Uses Slot Attention to capture temporal patterns across regions
   - Enables structured reasoning on crop-level changes
3. Contextual Crop Decoder (CCD)
   - Re-projects region-level features back into pixel space
   - Preserves boundary accuracy and spatial consistency

SFCNet Architecture



Feature Extraction

- Input: bi-temporal satellite images $I_1, I_2 \in \mathbb{R}^{3\times H\times W}$
- Backbone: ResNet-18 encoder (shared weights)
- Output: dense convolutional feature maps

$$F_1 = \mathrm{ResNet}(I_1), \quad F_2 = \mathrm{ResNet}(I_2), \quad F_1, F_2 \in \mathbb{R}^{C\times H\times W}$$

- where C=128
- These features encode rich semantic & spatial information

Semantic Region Rokenizer (SRT)

- Goal: represent features at the field level (not just pixels/patches)
- Input: feature maps $F_1, F_2 \in \mathbb{R}^{C \times H \times W}$ and field segmentation mask $\mathcal{M} \in \{0,1\}^{N \times H \times W}$
- For each region $i \in \{1, \ldots, N\}$ and time $t \in \{1, 2\}$
  - Average features across all pixels in region

$$z_t^{(i)} = \frac{1}{|\Omega_i|} \sum_{(h,w) \in \Omega_i} F_t[:, h, w] \cdot \mathcal{M}_i[h, w]$$

  - Create a region-level token

$$Z_t = \{z_t^{(1)}, \ldots, z_t^{(N)}\} \in \mathbb{R}^{N \times C}$$

  - Concatenate the two time points

$$Z_{\text{concat}} = [Z_1 \,\|\, Z_2] \in \mathbb{R}^{N \times 2C}$$

Field-Aware Region Encoder (FARE)

- Based on Slot Attention mechanism
- Input: concatenated region token

$$Q = W_q S, \quad K = W_k Z_{\text{concat}}, \quad V = W_v Z_{\text{concat}},$$
$$\alpha = \text{softmax}\left(\frac{QK^\top}{\sqrt{C}}\right), \quad U = \alpha V,$$
$$S \leftarrow \text{GRU}(U, S) + \text{MLP}(\text{LayerNorm}(S)),$$

- Output:
  - Context-aware slot embeddings representing temporal interactions across fields
- Advantage:
  - Groups regions by change dynamics
  - Cluster similar regions
  - Builds high-level reasoning beyond pixel differences

Contextual Crop Decoder (CCD)

- Goal: restore pixel-level change maps while preserving field boundaries
- Input: enriched slot embeddings from FARE

1. Split slot embeddings into time-specific features

$$S = [S_1 \,\|\, S_2], \quad S_1, S_2 \in \mathbb{R}^{K \times C}$$

2. Broadcast features back to spatial domain using original field masks and reconstruct semantically enriched feature maps

$$\bar{F}_t = \sum_{k=1}^{K} s_t^{(k)} \otimes \mathcal{M}_k$$

- Output: pixel-level feature maps aligned with region geometry
- Advantage:
  - Preserves boundaries & spatial consistency
  - Bridges gap between region-level reasoning and pixel-level prediction

Change Prediction Head

- Input: reconstructed feature maps from CCD

1. Compute pixel-wise absolute difference

$$D = |\bar{F}_2 - \bar{F}_1|, \quad D \in \mathbb{R}^{C \times H \times W}$$

2. Pass through a shallow convolutional decoder

$$\hat{Y} = \mathrm{Conv}(D), \quad \hat{Y} \in [0,1]^{H \times W}$$

- Output: pixel-wise probability map of change

Loss Function

- Training: Binary Cross-Entropy (BCE) loss

$$\mathcal{L}_{\text{BCE}} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\left[Y_{h,w}\log\hat{Y}_{h,w} + (1-Y_{h,w})\log(1-\hat{Y}_{h,w})\right]$$

- Advantage:
  - Highlights fine spatial differences
  - Balances local detail and global reasoning
  - Penalizes false positives and false negatives

3. Experimental Results

- CLCD (Cropland Change Detection) [3] - benchmark for fine-grained agricultural CD
- Content:
  - 600 bi-temporal image pairs (size 512 × 512)
  - Collected from Gaofen-2 satellite imagery (0.5–2 m resolution)
  - Covers cropland areas in Guangdong Province, China (2017–2019)
- Annotations:
  - Binary masks for cropland changes
  - Change types: conversion to buildings, roads, lakes, bare soil, land-use changes
- Split: 320 training, 120 validation, 120 test pairs

Implementation Details

- Hardware: NVIDIA RTX 3090 GPU
- Unified Change Detection Framework (for standard comparisons)
- Training setup:
  - Optimizer: Adam
  - Learning rate: 0.0001
  - Batch size: 16
  - Epochs: 100
- Backbone: ResNet-18 (shared across time steps)
- Loss function: Binary Cross-Entropy (BCE)

Evaluation metrics

- IoU, Accuracy, F1 Score, Precision, Recall

$$\text{IoU} = \frac{TP}{TP + FP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Networks for Comparison

Category	Model	Key Idea
CNN-based	SiamUNet-Diff [4]	Siamese U-Net, feature subtraction
	A2Net [5]	Lightweight, MobileNet backbone + attention
	RDPNet [6]	Boundary-aware loss for fine detail
Transformer	BIT [7]	Tokenized bi-temporal features, transformer reasoning
	ChangeFormer [8]	Fully transformer-based Siamese architecture
Hybrid	ICIFNet [9]	Cross-scale feature interaction + attention
	DMINet [10]	Dual-branch, subtraction + concatenation + joint attention

Quantitative Results

Model	IoU	Accuracy	F1	Precision	Recall
A2Net	23.42	92.04	37.95	45.18	32.71
SiamUNet-Diff	32.57	93.58	49.13	59.83	41.68
RDPNet	37.28	92.88	54.31	51.95	56.90
ICIFNet	39.17	94.16	56.29	63.54	50.52
DMINet	40.28	93.91	57.42	59.86	55.18
ChangeFormer	45.07	94.31	62.14	61.50	62.79
BIT	49.16	94.88	65.91	65.26	66.58
SFCNet (Ours)	**52.47**	**95.20**	**67.23**	**67.85**	**68.37**

Quantitative Results

- SFCNet outperforms all baselines across IoU, F1, Precision, Recall
- Achieves 52.47% IoU, which is +3.31% higher than BIT (the previous best)
- Delivers 67.23% F1 score, showing a strong balance between precision (67.85%) and recall (68.37%)
- CNN-based models struggle with complex spatial-temporal dependencies
- Transformer-based models improve global reasoning but still miss fine boundary accuracy
- Region-aware modeling ensures both semantic consistency and boundary precision

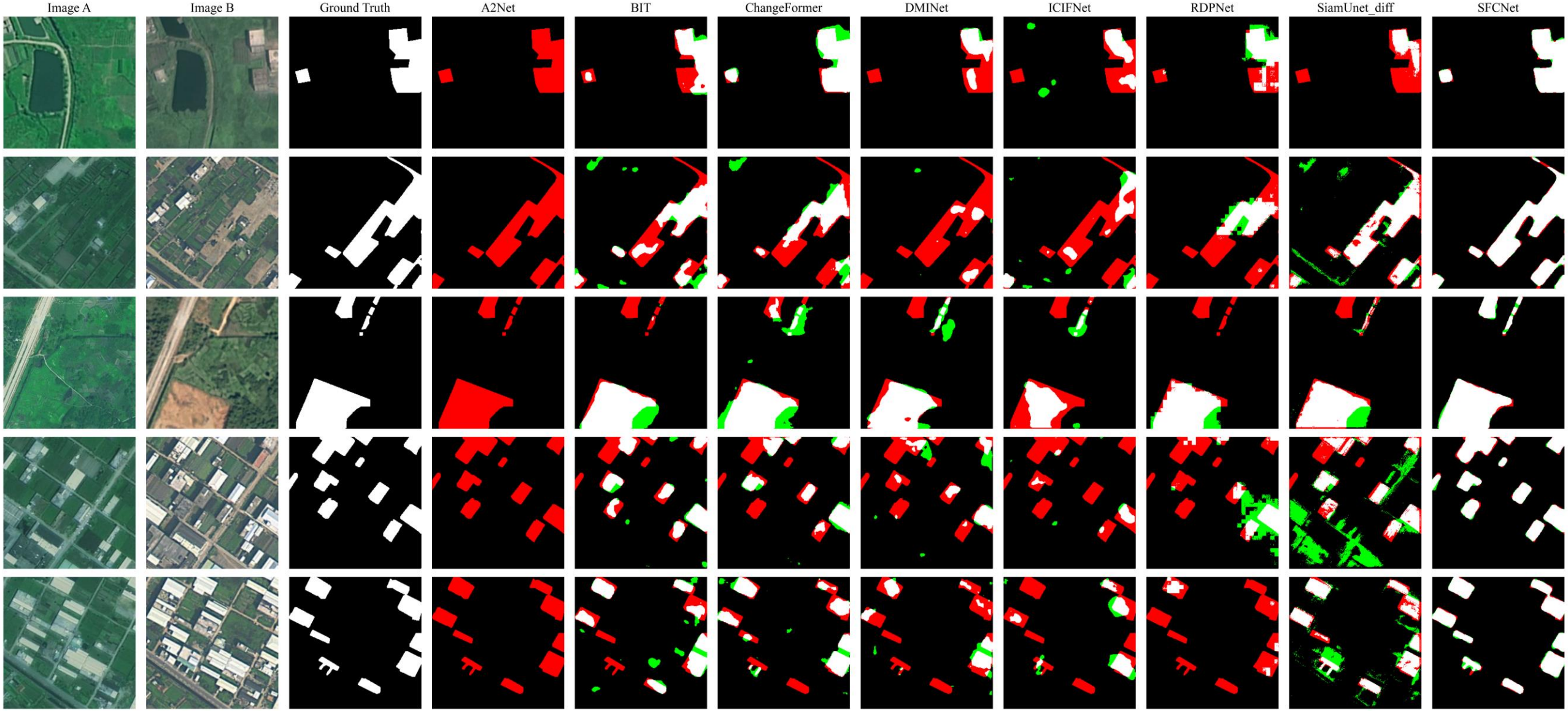
Performance Results

Model	IoU	Accuracy	Parameters (M)	FLOPs (G)	Training time (h)	Inference time (ms)
A2Net	23.42	92.04	3.60	2.86	0.43	1.54
BIT	32.57	93.58	11.39	8.28	0.98	2.01
ChangeFormer	37.28	92.88	39.13	129.70	26.13	7.43
DMINet	39.17	94.16	6.44	16.16	0.65	3.42
ICIFNet	40.28	93.91	24.64	22.97	2.02	5.38
RDPNet	45.07	94.31	1.62	1.63	0.62	6.83
SiamUNet-Diff	49.16	94.88	1.29	3.99	0.37	1.74
SFCNet (Ours)	**52.47**	**95.20**	**14.51**	**6.08**	**1.26**	**1.89**

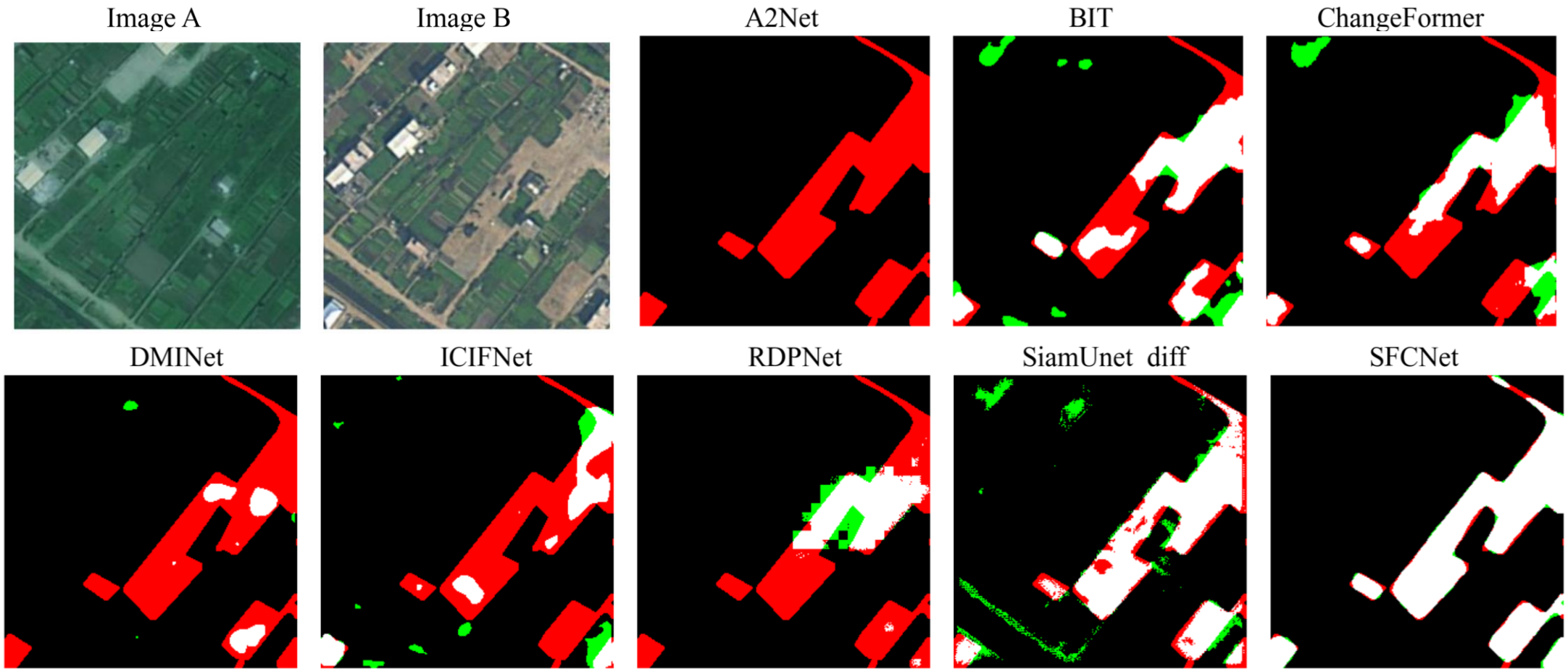
Performance Results

- SFCNet offers the best trade-off between accuracy and efficiency
- Despite having 14.5M parameters, it remains lightweight compared to ChangeFormer (39M)
- Requires only 6.08 GFLOPs, significantly less than ChangeFormer (129 GFLOPs)
- Achieves fast inference (1.89ms), faster than BIT, ICIFNet and ChangeFormer
- Key takeaway:
  - CNN models → efficient but less accurate
  - Transformer models → accurate but heavy
  - SFCNet → combines the best of both worlds: accurate, efficient, practical for real-time cropland monitoring

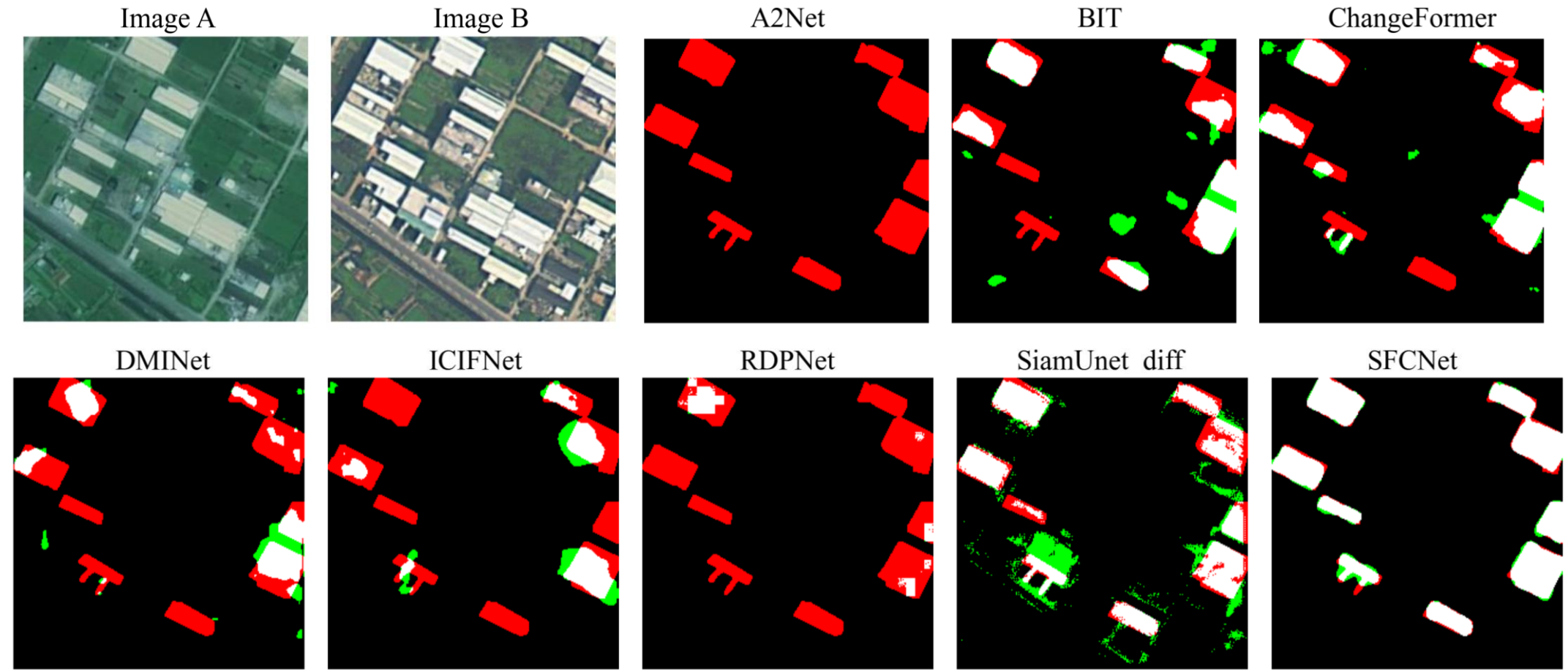
Qualitative Results



Qualitative Results



Qualitative Results



Qualitative Results

- SFCNet produces cleaner boundaries and fewer false positives/negatives
- Competing models often:
  - Overestimate change regions (BIT, ChangeFormer)
  - Miss narrow or fine structures (DMINet, SiamUNet-Diff)
- Our model successfully delineates fine structural changes while suppressing background noise
- Demonstrates strong region coherence, predictions align with field boundaries
- Particularly effective in heterogeneous and subtle cropland transitions
- Key takeaway: SFCNet achieves both precision and consistency — outperforming SOTA not just in numbers, but also in visual interpretability

Conclusions

- Proposed SFCNet: a region-aware change detection framework for remote sensing images
- Key innovations:
  - Semantic Region Tokenizer (SRT): aligns features with real-world field boundaries
  - Field-Aware Region Encoder (FARE): slot attention for structured temporal reasoning
  - Contextual Crop Decoder (CCD): precise boundary-preserving reconstruction
- Outperforms state-of-the-art models by +3.31% IoU on CLCD
- Achieves a strong balance of accuracy, efficiency, and interpretability

Future Work

- Explore weakly supervised training without ground-truth field masks
- Extend SFCNet to heterogeneous scenes (urban, building change detection, post-disaster environments)
- Investigate scalability for large-scale agricultural monitoring
- Integrate with foundation models for open-vocabulary semantic change detection

References

[1] L. Khelifi and M. Mignotte, “Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis,” IEEE Access, vol. 8, pp. 126 385–126 400, 2020.

[2] T. Bai, L. Wang, D. Yin, K. Sun, Y. Chen, W. Li, and D. Li, “Deep learning for change detection in remote sensing: a review,” Geo-spatial Information Science, vol. 26, no. 3, pp. 262–288, 2023.

[3] M. Liu, Z. Chai, H. Deng, and R. Liu, “A CNN-Transformer Network With Multiscale Context Aggregation for Fine-Grained Cropland Change Detection,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 4297–4306, 2022.

[4] R. Caye Daudt, B. Le Saux, and A. Boulch, “Fully Convolutional Siamese Networks for Change Detection,” in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 4063–4067.

[5] Z. Li, C. Tang, X. Liu, W. Zhang, J. Dou, L. Wang, and A. Y. Zomaya, “Lightweight Remote Sensing Change Detection With Progressive Feature Aggregation and Supervised Attention,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–12, 2023.

[6] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, “RDP-Net: Region Detail Preserving Network for Change Detection,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–10, 2022.

[7] H. Chen, Z. Qi, and Z. Shi, “Remote Sensing Image Change Detection With Transformers,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.

[8] W. G. C. Bandara and V. M. Patel, “A Transformer-Based Siamese Network for Change Detection,” in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 207–210.

[9] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, “ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.

[10] Y. Feng, J. Jiang, H. Xu, and J. Zheng, “Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1– 15, 2023.

Region-Aware Token Aggregation for Fine-Grained Change Detection

Bianca-Cerasela-Zelia Blaga

Technical University of Cluj-Napoca
ICCP 2025