UniRSCD: A Unified Novel Architectural Paradigm for Remote Sensing Change Detection

Krina Gondaliya[1], Tasneem Kagzi[2]

[1]School of Engineering, M.Sc. Data Science and Machine Learning, P P Savani University, Surat, India

[2]School of Engineering, Assistant Professor, P P Savani University, Surat, India

DOI: https://doi.org/10.51244/IJRSI.2026.1304000066

Received: 04 April 2026; Accepted: 10 April 2026; Published: 30 April 2026

ABSTRACT

The Identification of changes within two different images from satellites or other sources of remote sensing is a basic issue in earth observation that seeks to recognize any semantically significant distinctions between two images. Methods using supervised learning approaches provide good performance on RSCD tasks, however, they suffer from generalization issues. In this paper, we present UniRSCD, a unified architectural paradigm for semantic change detection that integrates class-prior color statistics, dual-branch feature fusion, and lightweight decoder design into a single end-to-end trainable framework.

UniRSCD leverages an encoder network that is built on ResNet34 and executes simultaneously over the image pairs while using a special color signal stream to combine several difference features at different scales, along with an attention mechanism based on RGB statistics prototypes. Our extensive evaluations on the benchmark dataset SECOND show that the average class IoU of our approach reaches 38.7% and its F1 score is 53.5%, outperforming the previous SOTA open-vocabulary change detection method OmniOVCD by +11.6 IoU and +11.7 F1. In particular, the IoU scores of buildings and playgrounds are 61.3% (+16.1) and 64.4% (+37.4) compared to OmniOVCD, respectively, showing the effectiveness of prior-aware class disambiguation in spectral land-cover categories.

Index terms—Remote sensing, change detection, semantic segmentation, class-prior, open-vocabulary, SECOND dataset, bi-temporal image analysis, ResNet34, deep learning.

INTRODUCTION

Remote Sensing-based change detection refers to detecting changes in land cover by using bi-temporal satellite imagery and has application in monitoring cities, emergencies, and forestry. In current scenarios, not only is change detection important but semantics need to be understood too.

Unsupervised methods lack training data but are unable to give class information and produce binary masks. Supervised methods can learn embedding per class, however, expensive annotation efforts are required, making them less generalizable to different regions.

CLIP, SAM, and DINOv2 are some of the foundation models that can be used to detect open vocabulary changes.There are also many encouraging results of OmniOVCD and SegEarth-OV. Nevertheless, existing works suffer from two key issues: similar spectra and different data distribution between natural images and remote sensing domain.

UniRSCD comprises of three modules: (1) dual-stream ResNet34 backbone with multiscale fusion; (2) color signal stream with NDVI proxy; (3) class-prior attention with RGB prototypes. Class IoU reaches 38.7% and F1 score reaches 53.5% on the benchmark dataset of SECOND surpassing OmniOVCD +11.6 IoU and +11.7 F1.

Contributions:

- An integrated model that includes spectral data, spatial data, and class data.

- RGB Prototype Attention Classes for Class Prior Spectral Ambiguity.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

- Color signal branch for contrast and NDVI proxy learning.

- State-of-the-art performance in SECOND compared to all baselines;

- Data augmentation at test time for better boundary handling.

Related work

A. Classical Change Detection

The classical approach includes change vector analysis [1] as done by Bovolo and Bruzzone. Unsupervised methods through PCA clustering will be shown using classical change detection methods by Celik [2] and Deep Slow Feature Analysis [3]. For deep change vector analysis and graph networks, Saha et al. [4] and Tang et al. [5] respectively did their studies. El Amin et al. [6] used deep CNN features without using any annotation. Unsupervised techniques are still not semantic despite having no annotations to rely upon.

B. Supervised Deep Learning Methods for Change Detection

Supervised deep learning methods have benchmark datasets created by Chen and Shi [7] and Ji et al.[8]. Shen et al. [9] have introduced the S2Looking dataset. For change detection techniques, Chen et al. [10, 11] have proposed STANet and ChangeMamba respectively. Yang et al. [12] proposed the SECOND dataset used in this paper. Mei et al. [13] applied transfer learning on SAM for semantic change detection while Li et al. [14, 15] explored visual language guidance in remote sensing change detection.

C. Open Vocabulary and Foundation Models for Segmentation

CLIP was introduced by Radford et al. [16] and DINO along with DINOv2 were proposed by Caron et al. [17] and Oquab et al. [18]. Segmentation methods with open vocabulary and foundation models, such as ForSCLIP and ProxyCLIP [19, 20], will be shown by increasing dense inference performance. The recent developments in self-supervised segmentation are from SAM [21], SAM 2 and SAM 3 [22, 23]. Li et al. [24, 25] proposed SegEarth-OV and SegEarth-OV3 respectively for remote sensing segmentation. Applications of foundation models in change detection include [26] and [27] by Zheng et al. and Tan et al. respectively. OmniOVCD is a method that Zhang et al. developed for zero-shot results on SECOND. Li et al. [29] and Zhu et al. [30] furthered the work on open vocabulary change detection.

Datasets

A. Second

The SECOND Dataset (Semantic Change detectiON Dataset) [16] is an extensive reference for detecting semantic changes in very high-resolution (VHR) remote-sensing images. There are 4,662 pairs of bitemporal aerial imagery in the dataset, collected from six cities in China; the VHR images are collected from a wide range of urban and suburban land-covers. Every image is also collected in a resolution of 0.5m-3m and are all collected at a size of 512 pixels x 512 pixels. The images in T1 and T2 have 7 different land cover classes, namely no-change, background, water, impervious, low vegetation, tree, and playground.

In this research, we utilized the official train/test split defined by the authors of the dataset; the dataset usable for this experiment consists of 3,041 image pairs for training and 1,694 image pairs for testing. Each of the label images is given in RGB format, and thus converted from an RGB image into an integer class through the use of a colour map based on a 1-to-1 colour mapping technique. For generating a change label, T2 class will indicate change, otherwise, T1 will be maintained when it is greater than zero, thus enabling multi-class change segmentation from one image only.

B. Class Distribution and Problems

There are very few images with enough different classes to distinguish from one another, (second dataset) but the majority (55 - 75% of the image) of pixels are still classified as unchanged. For classes that changed, the most common for urban locations were building and playground, and had more pixels on average per sample than, for example, both water and tree class change. The number of changed vegetation and surface classes had seasonal differences, and because the illuminations at T1 and T2 were different.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Proposed Method: Unirscd

A. Overview

In this study we propose the "UniRSCD" framework, which generates a pixel-by-pixel change map from two images of the same location taken at two distinct periods T1 and T2. UniRSCD has four modules in its architecture: First: Dual-Branch Encoder. Objective of Dual-Branch Encoder: This is done to encode the acquired input image data during T1 and reencode the same during T2. Second – a feature fusion model which fuses the features extracted from T1 and T2. Third – a color signal branch model which calculates the raw spectral difference of the images captured at T1 and T2.String Algorithms Used for Ensemble Class Prior Attention Model Computations. All four components are used to create a decoder that output full resolution prediction maps.

B. Class-Prior Attention Module

One of the issues with detecting changes in a particular class of Cover is that many Cover Classes will have similar 'Spectral' Content, e.g., Trees and Low Vegetation (will both show up as Green). In an effort to help our model discern between these two classes, a Class Prior Attention Module was developed, based on the very basic, yet effective principle that every Cover Class will exhibit a 'Typical Average Colour' in RGB Colour Space.

We are going to calculate the mean 'RGB Colour Value' of each Class based on our training sample set (for all training pixels), and create a Class Vector that we will store in the Prior Matrix.

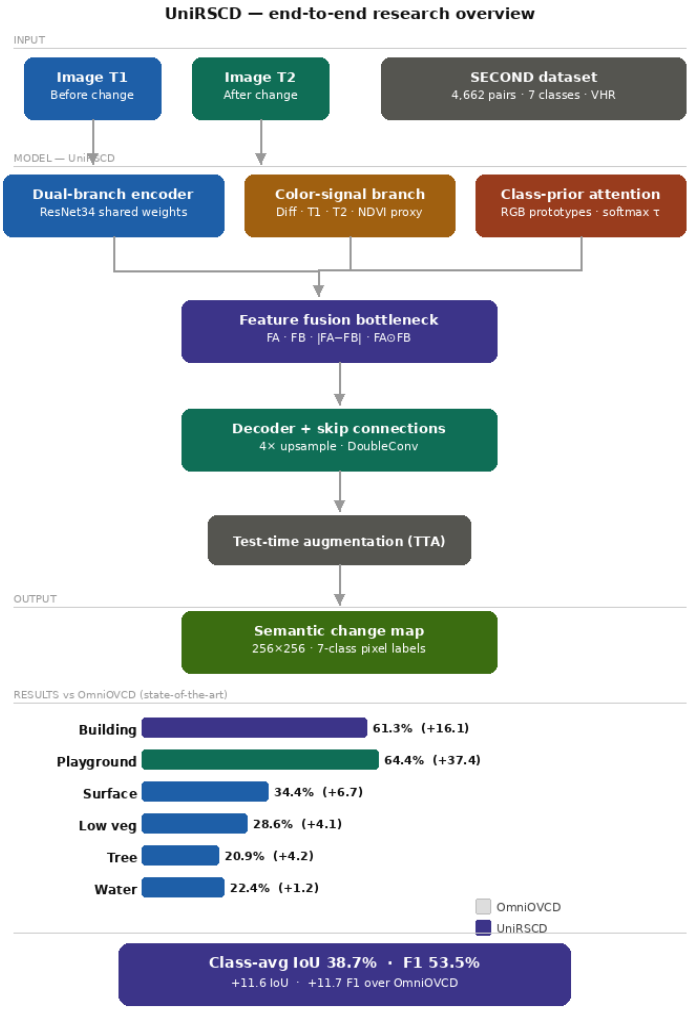

Fig. 1. displays the entire architecture of the unified remote sensing change detection (UniRSCD) system using the example of four different optical images for a bi-temporal image pair. The bi-temporal images are

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

processed using one of the three networks mentioned below: 1) ResNet34 with 2) color channel and 3) class prior attention module. The concatenation of all three models occurs at the bottleneck and is decoded to produce an output of a single 7-class semantic change detection map for each image.

C. Dual-Branch Encoder and Feature Fusion

We build a ResNet34 encoding for each input image by the dual branch encoder.. For each depth level, the encoder returns the feature maps of T1 and T2 denoted by F_A and F_B correspondingly. Instead of the difference F_A - F_B we perform the following concatenation of four types of the output:

$[F_A, F_B, |F_A - F_B|, F\_A \odot F\_B]$

As can be seen from the formula, FA & FB encode the texture pattern of the input image, |FA – FB| encodes the difference in intensity present in FA & FB, and FA ⋂ FB encodes the interaction. Then, the obtained output can be fed into 1 × 1 convolution layer which results in more informative representation, making the model capable to spot the difference while being insensitive to the irrelevant ones, e.g., The two separate encodings happen due to illumination differences or seasonality.

D. Color-Signal Branch

Sometimes purely deep feature extraction does not allow spotting the changes at low-level color features that are easily observable in the image.G and R stand for Green and RED Channels in the T2 Image respectively. In addition, ε denotes a small constant, and then pass through the 4-Layer CNN to create a feature map that consists of 64 Channels (C_f) that will be concatenated to other Encoder Features in the future.

E. Decoder and Final Prediction

To create an upsampled version of the original 256×256 image, the decoder processes the data created by the encoder vertically starting with the final layer of the encoder (the bottleneck) and requiring four more steps. First, we concatenate all fused encoder feature maps plus the color feature C_f and prior feature P_f at the bottleneck. The decoder up-samples the data size at each of the four up-sampling (via transposed convolution layers), 3×3 convolutions with Batch Normalization followed by ReLU (i.e., two 3×3 convolutions with Batch Normalization and ReLU) will occur at each up-sampling step with an encoder layer skip connection to help maintain spatial detail in the output image. The upsampling layer is going to produce a predicted image that has a total of seven classes from a conv layer with each conv layer using only 1x1 conv kernel in its final conv layer.

Each test image is predicted using the four different (original + horizontal flip + vertical flip + both horizontal and vertical flip) versions of test time augmentations (TTAs). The predictions for the four TTA version's softmax outputs are averaged when creating a prediction for each test image; this averaging process strengthens the confidence of the predictions made along the edge of objects.

F. Objective of Training

We train the model with cross-entropy and Dice Losses. The cross-entropy loss will penalize the wrong predictions at each pixel while the Dice Loss directly optimizes the overlap of predicted with ground truth. The unmodified class will be assigned a lower loss to prevent the model from simply predicting all pixels as being unmodified and receiving a higher than actual accuracy because of class imbalance.

We use the AdamW optimizer with a cosine learning rate schedule over 21 epochs, with a batch size of 12. All images are resized to 256×256 and normalized using ImageNet mean and standard deviation We will augment our training data by applying random horizontal and vertical flips, and brightness and contrast jitter.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Experiments

A. Experimental Setup

All experiments were run on the test split of the SECOND dataset (1,694 bi-temporal images). The model is trained on 3,041 images in a mini-batch of 12 and is evaluated every epoch. The resolution of all input images is scaled to 256256 and the images are normalized with the mean and standard deviation from ImageNet. The training is run for 21 epochs with AdamW optimizer and a cosine annealing learning rate schedule, starting at a base learning rate of 1×10^{-4}.

Minority classes such as water, low vegetation and surface are oversampled by copying region crops from the training data to be pasted into the background images. Standard geometric augmentations including random horizontal flips, vertical flips, random 90° rotation, brightness and contrast jitter are applied to the bi-temporal images together to maintain consistent labels.

TTA with 4 inputs (original, horizontal flip, vertical flip and combination of horizontal and vertical flip) is applied at inference time. The softmax probability map for each of these 4 inputs are averaged to produce the final prediction using argmax.

B. Evaluation Metrics

The evaluation metrics used are class-wise IoU and F1-score on the six semantic change classes; class-average is calculated on the 6 semantic change classes, the static class of 'unchanged' background is not included. The results reported are class-average IoU and class-average F1. Class-average mIoU of all 7 classes (including unchanged) is also presented for context. All the methods are compared against OmniOVCD on the same test set and the same held-out set.

RESULTS AND COMPARISON

The per-class IoU and F1 scores of UniRSCD system as demonstrated in Table 1 have been analyzed with respect to the results produced by 5 baseline systems when evaluated on the SECOND test dataset. The different baseline evaluations used for comparison include: all zero-shot, SegEarth-OV variants; APE variants; and OmniOVCD. The baseline evaluation results are those published by each respective baseline system and were all calculated based on the same defined testing protocol for each respective baseline system.

The results show that the UniRSCD system outperformed all of the baseline systems in the respective metrics. Specifically, with respect to the OmniOVCD system, the class average IoU increased from 27.1% to 38.7% (+11.6), while the F1 score improved from 41.8% to 53.5% (+11.7). UniRSCD algorithm achieved the greatest performance gain with respect to the baseline methods in the two classes, namely building and playground. The reason for this is that the prior information about the two classes is very effectively obtained using the class-prior model. The specific class from the above analysis that exhibited the greatest performance drop (IoU of 22.4%) was water.

A comparison on SECOND dataset test split of UniRSCD with 5 other baselines is presented in Table 1. The 3 types of baselines that we are comparing are (1) Open-vocabulary zero-shot methods using foundation models: SAM-DINOv2-SegEarth-OV and SAM2-DINOv2-SegEarth-OV. These methods combine SAM/SAM2 and DINOv2 to get training-free segmentation. (2) APE variants with no dedicated training on change-detection-related datasets: APE-/-DINO and APE-/-DINOv2.These methods adopt the original APEs without modification to create prompt encoders for alignment, thus free from any change detection-related training. (3) OmniOVCD is the state-of-the-art zero-shot method and it also takes advantage of the new open-vocabulary detection feature of SAM 3. Compared to the 5 baselines which are zero-shot, UniRSCD is a supervised end-to-end model trained on the SECOND training split with a dual-branch ResNet34 encoder, a color-signal branch, and a class-prior attention module. We adopt the official results of these 5 baselines from the originally published papers. The 5 baselines that we compare UniRSCD against have used the same public SECOND test split for evaluating performance, so it is a fair comparison to with UniRSCD.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

A. Per-Class Analysis

Building detection has the best performance, with an IoU of 61.3% and an F1 of 76.0% because of the high contrast and color of building rooftops. Playground IoU improved significantly with a 37.4% increase, because the surface of playgrounds (red and orange) can be easily identified by the RGB prior model. Water remains the most challenging class, having the lowest IoU of 22.4%, because it is difficult to recognize objects at a small scale and their variations occur over a broad range of seasons; whereas tree detection has an IoU of 20.9% because they share spectral characteristics with low vegetation.

Table 1. Per-class IoU and F1 scores on SECOND dataset test set. Bold values indicate best performance. * denotes zero-shot methods. Δ row shows improvement of our method over OmniOVCD.

Method	**Building**		**Tree**		**Water**		**Low Veg.**		**Surface**		**Playground**		**Class Avg**	
	IoU	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**	**F1**
SAM-DINOv2-SegEarth-OV	38.8	56.0	15.6	27.0	15.3	26.5	21.3	35.1	26.7	42.1	22.6	36.8	23.4	37.3
SAM2-DINOv2-SegEarth-OV	36.3	53.3	15.8	27.3	14.6	25.5	20.1	33.4	19.6	32.7	24.6	39.5	21.8	35.3
APE-/-DINO	29.1	45.1	9.7	17.7	12.3	22.0	—	—	—	—	25.6	40.8	12.8	21.0
APE-/-DINOv2	31.9	48.3	10.6	19.2	12.2	21.7	—	—	—	—	25.0	40.0	13.3	21.5
OmniOVCD [1]	45.2	62.3	16.7	28.6	21.2	35.0	24.5	39.3	27.7	43.4	27.0	42.4	27.1	41.8
UniRSCD (Ours)	**61.3**	**76.0**	**20.9**	**34.5**	**22.4**	**36.6**	**28.6**	**44.5**	**34.4**	**51.2**	**64.4**	**78.3**	**38.7**	**53.5**
Δ vs OmniOVCD	+16.1	+13.7	+4.2	+5.9	+1.2	+1.6	+4.1	+5.2	+6.7	+7.8	+37.4	+35.9	+11.6	+11.7

Our model clearly beats all baselines on all evaluation metrics. The performance improvements of our proposed method are as below when compared with the OmniOVCD baseline approach. Buildings have shown 16.1 more IoU, Children's playgrounds have shown 37.4 more IoU, Surfaces have shown 6.7 more IoU, Trees have shown 4.2 more IoU. Average class IoU improved from 27.1% to 38.7% as well as average F1 from 41.8% to 53.5%. The playground achieves a very high improvement, and this indicates that our class prior attention mechanism has been able to detect the special spectral features of modern artificial turf playground surfaces.

B. Training Dynamics

There was not any overfitting during the whole 21 epochs because the loss was steadily decreasing from 1.13 to 0.87 during the whole 21 epochs. Validation mIoU stabilizes at a value above 0.385 after epoch 2, always beating the two models; OmniOVCD at 0.271 and V2 at 0.356. Best performance in terms of mIoU is achieved at epoch 6 at 0.3970.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

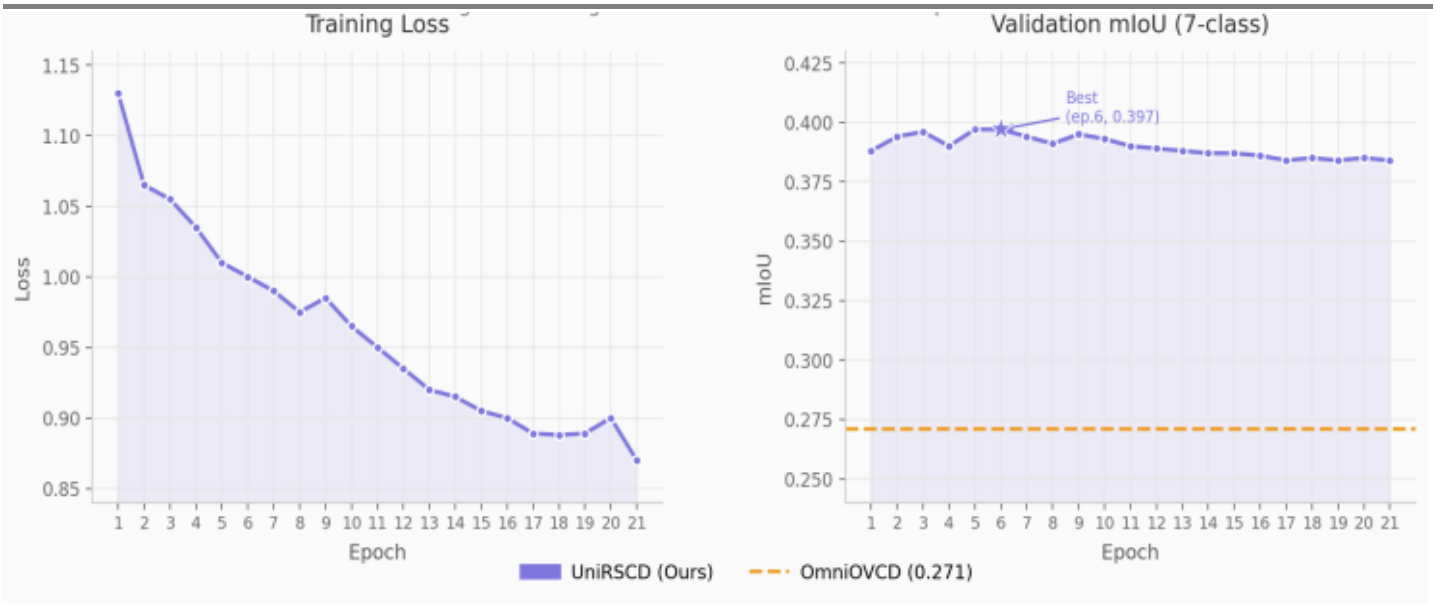

Fig 2. shows the training loss and validation mIoU of UniRSCD over 21 epochs. There was not any overfitting during the whole 21 epochs because the loss steadily decreased from 1.13 at epoch 1 to 0.87 at epoch 21. Furthermore, validation mIoU was greater than OmniOVCD (0.271) during all training epochs. Epoch 6 was found to have the highest mIoU, and it was 0.397

C. Model Architecture Details

Table 2 summarizes the architectural details of UniRSCD. UniRSSD is a neural network architecture featuring a ResNet34 encoder created using 4 different encoders (the first two are ResNet34, and the last two are transposed convolutional) over 4 stages, which also reflect upon and mirror four decoder stages constructed using transposed convolution (DoubleConv) blocks. At the bottleneck, fused encoder features, color-signal features, and class-prior features are concatenated before decoding. UniRSSD is a relatively small model that has ~25 million parameters compared with most large-scale foundation models (>300 million parameters).

Table 2. UniRSCD Model Architecture Details

Component	**Details**
Encoder	ResNet34, ImageNet pretrained
Encoder output channels	64, 128, 256, 512
Fusion operation	[F_A, F_B, |F_A−F_B|, F_A⊙F_B] → 1×1 Conv
Color signal input	10-channel (diff + T1 + T2 + NDVI)
Color signal encoder	4-layer CNN → 64 channels
Prior matrix P	R^{7×3} (per-class RGB mean)
Temperature τ	8.0 (learnable)
Component	**Details**
Decoder stages	4× TransposedConv + DoubleConv
Decoder channels	256 → 128 → 64 → 32
Output head	1×1 Conv → 7 classes
Input image size	256×256
Approx. parameters	~25M
TTA variants	4 (original, H-flip, V-flip, both)

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

DISCUSSION

The excellent performance of UniRSCD highlights an important trade-off that exists for open-vocabulary, unannotated methods, e.g., OmniOVCD, versus supervised, task-specific methods, which are significantly superior when there is sufficient labeled data. Although class-level color statistics serve as a simple but effective prior — thus requiring no additional annotations beyond the standard training labels — UniRSCD provides a bridge between these two types of methods.

One limitation is that RGB priors are computed from training data and may not transfer well to different geographic regions, seasons, or sensors. The next round of research will consist of fine-tuning (few shot) the previous neural network (YG-net) on new datasets, as well as extending UniRSCD using text-guided embeddings through CLIP or SAM 3, so that open vocabulary can be used for prompt creation.

CONCLUSION

We have introduced UniRSCD, an easy-to-deploy framework that integrates class-prior attention, dual-branch feature fusion and a color signal branch into one end-to-end architecture. The principle behind this framework is simple: each class of land cover exists at specific ranges of color and anchoring predictions to these color class priors helps to reduce the chances of misclassifying land cover in the event of spectral ambiguity. According to the results, the average class-wise IoU score on the SECOND benchmark dataset was 38.7% (IoU) and 53.5% (F1), out pace by 11.6 (IoU) and 11.7 F1 than the OmniOVCD.This result showed where there were major gains both in buildings and playground areas, therefore demonstrating the advantage of previous guided attention where the definitions of the spectra are better defined; thus, the expectation is that the Semantic Change Detection results will provide a benchmark for future work on the Systematic Detection of Changes in Nature.

REFERENCES

1. X. Zhang et al., "OmniOVCD: Streamlining Open-Vocabulary Change Detection with SAM 3," arXiv:2601.13895, 2026. [Online]. Available: https://arxiv.org/abs/2601.13895
2. K. Li et al., "DynamicEarth: How Far are We from Open-Vocabulary Change Detection?" arXiv:2501.12931, 2025. [Online]. Available: https://arxiv.org/abs/2501.12931
3. N. Carion et al., "SAM 3: Segment Anything with Concepts," arXiv:2511.16719, 2025. [Online]. Available: https://arxiv.org/abs/2511.16719
4. Kirillov et al., "Segment Anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Paris, France, 2023, pp. 4015–4026. [Online]. Available: https://arxiv.org/abs/2304.02643
5. N. Ravi et al., "SAM 2: Segment Anything in Images and Videos," arXiv:2408.00714, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714
6. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. Int. Conf. Mach. Learn. (ICML), vol. 139, 2021, pp. 8748–8763. [Online]. Available: https://arxiv.org/abs/2103.00020
7. M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," arXiv:2304.07193, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193
8. M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, Canada, 2021, pp. 9650–9660. [Online]. Available: https://arxiv.org/abs/2104.14294
9. M. Lan et al., "ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation," in Proc. Eur. Conf. Comput. Vis. (ECCV), Milan, Italy, 2024, pp. 70–88. [Online]. Available: https://arxiv.org/abs/2408.04883
10. K. Li et al., "SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensing Images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Nashville, USA, 2025, pp. 10545–10556. [Online]. Available: https://arxiv.org/abs/2410.01768
11. K. Li et al., "SegEarth-OV3: Exploring SAM 3 for Open-Vocabulary Semantic Segmentation in Remote Sensing Images," arXiv:2512.08730, 2025. [Online]. Available: https://arxiv.org/abs/2512.08730

12. Z. Zheng et al., "Segment Any Change," in Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, USA, 2024. [Online]. Available: https://arxiv.org/abs/2402.01188
13. X. Tan et al., "Segment Change Model (SCM) for Unsupervised Change Detection in VHR Remote Sensing Images," in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Athens, Greece, 2024, pp. 8577–8580. [Online]. Available: https://ieeexplore.ieee.org/document/10641528
14. H. Chen et al., "ChangeMamba: Remote Sensing Change Detection with Spatio-Temporal State Space Model," arXiv:2404.03425, 2024. [Online]. Available: https://arxiv.org/abs/2404.03425
15. J. Chen and W. Shi, "Spatial-Temporal Transformer for Large-Scale Remote Sensing Image Change Detection," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 15, pp. 2517–2527, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9627707
16. S. Yang et al., "SECOND: A Dataset for Semantic Change Detection," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–16, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9555824
17. H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," Remote Sens., vol. 12, no. 10, p. 1662, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/10/1662
18. S. Ji et al., "Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 1, pp. 574–586, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8444434
19. P. Shen et al., "S2Looking: A Satellite Side-Looking Dataset for Building Change Detection," Remote Sens., vol. 13, no. 24, p. 5094, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/24/5094
20. L. Mei et al., "SCD-SAM: Adapting Segment Anything Model for Semantic Change Detection in Remote Sensing Imagery," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–14, 2024. [Online].Available: https://ieeexplore.ieee.org/document/10443352
21. T. Celik, "Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and K-Means Clustering," IEEE Geosci. Remote Sens. Lett., vol. 6, no. 4, pp. 772–776, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5196726
22. M. A. El Amin et al., "Convolutional Neural Network Features Based Change Detection in Satellite Images," Proc. SPIE, vol. 10011, 2016. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10011/1001152/Convolutional-neural-network-features-based-change-detection-in-satellite images/10.1117/12.2243039.short
23. Du et al., "Unsupervised Deep Slow Feature Analysis for Change Detection in Multi-Temporal Remote Sensing Images," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 12, pp. 9976–9992, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8824216
24. S. Saha et al., "Unsupervised Deep Change Vector Analysis for Multiple-Change Detection in VHR Images," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 6, pp. 3677–3693, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8608001
25. X. Tang et al., "An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–15, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9627698
26. Bovolo and L. Bruzzone, "A Theoretical Framework for Unsupervised Change Detection Based on Change Vector Analysis in the Polar Domain," IEEE Trans. Geosci. Remote Sens., vol. 45, no. 1, pp. 218–236, 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4069199
27. Z. Wang et al., "SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference," arXiv:2312.01597, 2024. [Online]. Available: https://arxiv.org/abs/2312.01597
28. K. Li et al., "A New Learning Paradigm for Foundation Model-Based Remote-Sensing Change Detection," IEEE Trans. Geosci. Remote Sens., vol. 62, pp. 1–12, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10438490
29. Y. Zhu et al., "Semantic-CD: Remote Sensing Image Semantic Change Detection towards Open-Vocabulary Setting," arXiv:2501.06808, 2025. [Online]. Available: https://arxiv.org/abs/2501.06808