2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11246398

Spectral-Temporal Attention for Robust Change Detection

Mayank Thakur, Radhe Shyam Sharma, Vinod K Kurmi, Raj Samant, Badri Narayana Patro

Abstract— **Change detection has long been used for various tasks. With advancements in robotic systems and computer vision, change detection techniques can be further explored for diverse applications. Current state-of-the-art methods primarily use either satellite images or street-level images to detect changes. However, the techniques used for these two types of images differ substantially, though their core objective remains identical.**

We introduce a spectral-temporal attention network capable of detecting changes in both satellite and street-level images across multiple temporal instances. Additionally, we present an indoor environmental dataset featuring significantly more frequent changes. We analyze the impact of temporal and spatial domain shifts on the performance of various methods and demonstrate that performing attention in the spectral domain not only enhances overall performance but also increases robustness against spatial domain shifts.

I. INTRODUCTION

Change detection is a critical task in computer vision and machine learning, focusing on identifying changes in an environment over multiple temporal instances. This capability is essential for various applications, including surveillance [1]–[3], maintaining maps [4], patrolling, land cover monitoring [5], [6], anomaly [7], [8] detection. The ability to detect changes in real-time is crucial in scenarios where immediate action is necessary to maintain safety, security, or operational efficiency. Despite recent advancements, many methods struggle with real-world applicability, often failing outside controlled settings. Traditional techniques like image differencing, optical flow estimation, and background subtraction perform well under ideal conditions but degrade in dynamic lighting, occlusions, or cluttered scenes. These methods are also environment-specific, limiting their adaptability to diverse scenarios. The emergence of deep learning, particularly convolutional neural networks (CNN), has significantly advanced change detection by enabling the extraction of intricate patterns and features from images and videos. Models such as CDNet++ [9] have leveraged CNN architectures like VGG-19 [10] to achieve state-of-the-art results on benchmark datasets. However, these models underperform in densely populated and high-frequency regions where changes are more frequent and large in number. The rise of Unmanned Aerial/Ground Vehicles (UAVs/UGVs) has spurred interest in automated change detection systems for dynamic environments, enhancing autonomous capabilities while reducing human intervention. However, existing solutions still face limitations, particularly in highly cluttered or crowded scenes. Change detection can be classified into multiple categories based on its application and the source of input data. In general, it is divided into street scene change detection and satellite scene change detection. The primary difference between these two lies in the frequency and nature of changes observed over time. In satellite-based change detection, changes typically occur gradually, with significant transformations occurring over extended periods, such as deforestation, urban expansion, or natural disasters. In contrast, street-level change detection involves more frequent and dynamic alterations, such as moving vehicles, pedestrians, and temporary structures, making it significantly more challenging due to the rapid rate of change. Advances in drone technology and onboard computing have expanded change detection applications to disaster relief, and search-and-rescue. Drone-based scene change detection (SCD) merges satellite and street-level paradigms but faces challenges from altitude/angle variability, necessitating adaptive algorithms. Although several state-of-the-art methods and datasets exist for scene change detection, they still struggle with domain adaptation.

The key contributions are as follows:

- We propose a novel Spectral Attention Network (SAN) based approach which is robust to domain shifts, and can be applied to both satellite and street scene change detection.
- We developed a new real-world, indoor lab-based dataset, *A18Robotics*, for indoor scene change detection.
- Furthermore, we demonstrate the experimental validation of our method across different scenarios.

II. RELATED WORK

In this section we study the various change detection techniques used in the past. Several datasets have been developed for street scenes, each with unique characteristics and challenges. The PCD [11] dataset consists of 200 outdoor images, annotated for structural changes in street-view environments. Similarly, the VL-CMU-CD [11] dataset contains 1362 outdoor images, also annotated for structural changes and segmented into nine classes. These datasets facilitate the evaluation of methods under varying environmental conditions such as weather and lighting. The

Mayank Thakur, Radhe Shyam Sharma are with the Centre for Artificial Intelligence and Robotics at Indian Institute of Technology Mandi-175005, India (s23017@students.iitmandi.ac.in, radheshyam@iitmandi.ac.in).

Vinod K Kurmi is with IISER, Bhopal (vinodkk@iiserb.ac.in).

Raj Samant is with NVIDIA, India (rsamant@nvidia.com).

Badri Narayana Patro is with Microsoft, India (badripatro@microsoft.com).

MUDS [12] and DynamicEarthNet [13] are two satellite-based change detection datasets.

For indoor environments, the ChangeSim [14] dataset is noteworthy, providing over 130,000 images across 80 scans of 10 scenes in a warehouse setting. It includes annotations for four types of changes: new, missing, rotated, and replaced objects, with additional challenges such as dusty air and low illumination. While PCD [11] and VL-CMU-CD [11] datasets consists of real images, ChangeSim [14] is a synthetic dataset created using the Unreal Engine. ChangeSim [14] consists of paired and unpaired images which are trained and tested on ChangeNet [15] and CSCDNet [14] in different scenarios. Several significant methodological contributions have been made in recent years. Chen et al. [11] introduced a network designed to detect changes in street scenes using dynamic receptive temporal attention. This method leverages attention mechanisms to capture temporal changes in video sequences, improving the accuracy of change detection in complex urban environments. The paper demonstrates the effectiveness of the model on datasets such as PCD [11] and VL-CMU-CD [11]. Rubio et al. [16] explore the use of UAVs for automated change detection. It utilizes image alignment techniques such as ORB (Oriented and Fast Rotated Brief) and sliding windows. This approach is particularly useful for aerial imagery applications, where precise alignment and robust change detection are crucial. Santos et al. [17] presented an unsupervised change detection method for space habitats using Principal Component Analysis (PCA), Gaussian Mixture Models (GMM), and 3D point clouds. The method is specifically designed to compare images to identify anomalies aboard the ISS. Although effective in controlled settings, it requires significant computational resources. One important problem the authors highlight is that the effectiveness of the strategy decreases as environmental changes increase. Because it performs best when identifying up to five changes simultaneously, single image analysis is much more time-consuming and computationally expensive. However, it struggles with more thorough adjustments, particularly when there are more than 25 changes. Prabhakar [9] details an advanced approach using deep neural network feature correlation. The methodology is based on a VGG-19 architecture, employing convolutional blocks for patch-wise feature similarity. This approach has achieved state-of-the-art accuracy in datasets such as VL-CMU-CD, PCD, and GSV. However, it requires prealigned reference and query images, which can limit its applicability in scenarios where exact pixel-wise matching is not guaranteed. ChangeNet [15] is another method which utilizes ResNet encoder and fully connected layer based network to find the change map. SSCDNet [18] is silhouette based technique which makes use ResNet and UNet based network to find changes in the environment. SARAS-Net [19] is another Siamese based network consisting of relation aware, scale aware and cross transformer modules to tackle problem of scene change detection. CSCDNet [18] is another Siamese and correlation based method for detecting changes. C3PO [20] is another method which makes use of a segmentation backbone and merge the various temporal and spatial features to find the changes in the environment. UMAD [21] introduces an anomaly detection dataset which uses change detection techniques for the same. [22] uses Dino-V2 as baseline for detecting changes. [23] uses 3D Gaussian Splatting based features to find changes in the physical object rearrangement.

UTAE [24] is UNet based method to find changes in satellite images. TsVit [25] is a vision transformer based method for satellite change detection. [26] uses normal distribution based voxels to perform spatial change detection. [27] uses non descriptive fields and local object layout comparison to detect changes online. Spectformer [28] is a vision transformer based network which makes use of Spectral Gating Network to find the attention in the spectral domain. This technique not only improves the learning of the model but also increase its performance by reducing the total amount of computation required. Phasor Driven Acceleration [29] is another method which performs convolution in spectral domain. The introduction of spectral domain into a network improves the performance of the network and increases the overall learning of the method. For satellites, there exists a variety of hyper spectral datasets that serve diverse applications. In [30] authors had fused hyper spectral as well as spatial information for change detection.
Inspired by the strengths of various spectral methods in feature learning and domain adaptation, we explore attention mechanisms in the spectral domain to improve performance across diverse scenarios.

III. Method: Spectral Attention Network

We propose to modify the multiUTAE approach [31] by modifying the temporal attention module to calculate attention in the spectral domain. We also propose a new street scene change detection dataset to bridge the gap between the different change detection types. The overall architecture of our method is represented by Fig. 1. Spectral Temporal Attention block in Fig. 1 uses SAN to calculate attention in spectral domain.

A. Proposed Architecture

Our method takes multiple images captured in different temporal instances as input and gives segmented mask for each one of those images. These segmentation mask are compared pixel wise to find the changes among multiple temporal instances. Our method comprises an encoder, a spectral-temporal attention module, and a decoder. The encoder takes time-series input in $\mathbb{R}^{T\times C\times H\times W}$ where T is the number of temporal instances of a spatial location, C is the number of input channels and H and W are the height and width of the images. The encoder block produces a series of feature maps $\mathbf{z}^1$ to $\mathbf{z}^L \in \mathbb{R}^{T\times D}$ where L are number of levels and D is the spatial output dimension. Positional encoding is added to $\mathbf{z}^L$ and passed to spectral temporal attention block. The obtained attention map is up sampled at all levels l where l is

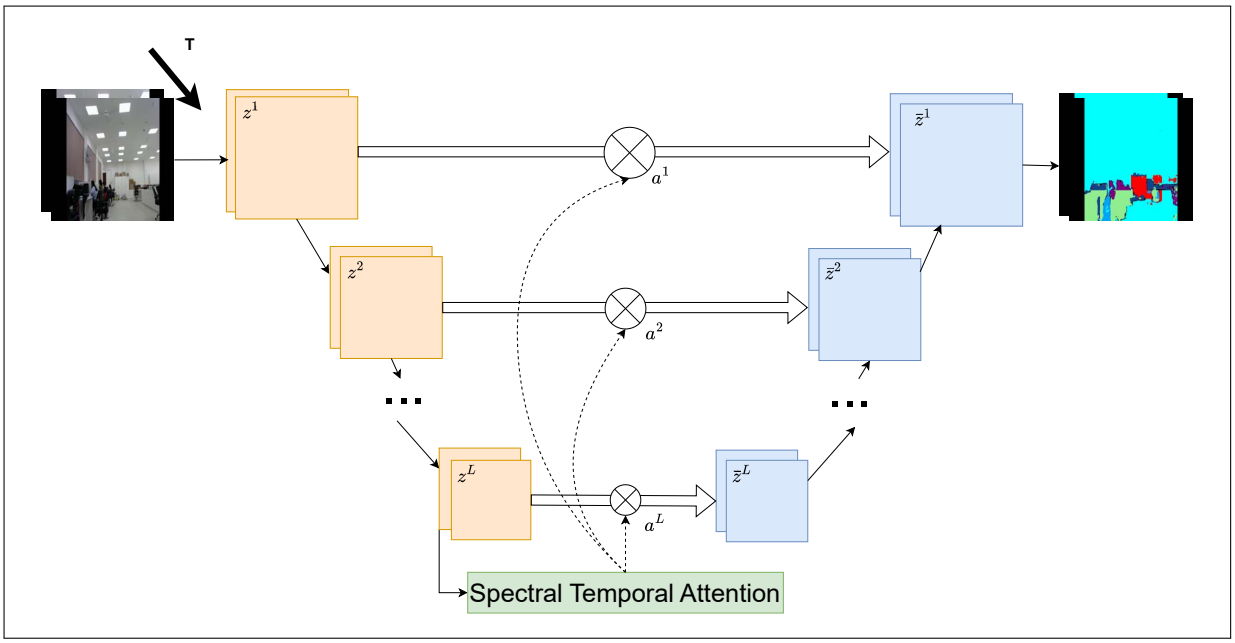

Fig. 1. Proposed model architecture.

between 1,...,L. The combined feature map $\bar{\mathbf{z}}_t^l$ is obtained at different levels for all t in $1, ..., T$. The decoder uses strided transposed convolution to upsample $\bar{\mathbf{z}}^l$ to the next level. The upsampled attention maps $\mathbf{a}^1$,..,$\mathbf{a}^L$ are propagated to upper level with skip connections. The obtained feature maps $\bar{\mathbf{z}}_t^l$ are concatenated with lower-level feature maps $\bar{\mathbf{z}}^{l+1}t$ before up sampling convolution. The model output are segmented maps of $\mathbb{R}^{T\times H\times W\times K}$ where K is the number of semantic classes. The encoder and decoder blocks are strictly aligned with the multiUTAE [31]. The spectral temporal attention block calculates the attention in the spectral domain where feature maps in key and query matrices are converted using Fast Fourier Transforms into spectral domain and attention maps are calculated. The attention maps are then converted back into the spatial domain using Inverse Fast Fourier Transforms. The attention mechanism outputs multi temporal attention maps $\mathbf{a}^L$ in $[0,1]^{T\times T\times H^{'}\times W^{'}}$ where $H^{'}$ is the height and $W^{'}$ is the width of the spatial image at level L. The attention in the spectral domain for a given spatial location (i,j) in $[1,H^{'}]\times[1,W^{'}]$ as:

$$\mathbf{q} = FC^q(FFT(\mathbf{z}_{i,j}^L)) \in \mathbb{R}^{T\times d} \tag{1}$$

$$\mathbf{k} = FC^k(FFT(\mathbf{z}_{i,j}^L)) \in \mathbb{R}^{T\times d} \tag{2}$$

$$\mathbf{a}_{i,j}^L = IFFT(\mathbf{k}\mathbf{q}^\top) \in \mathbb{R}^{T\times T} \tag{3}$$

In equation (1), $\mathbf{q}$ denotes the query matrix, while in equation (2), $\mathbf{k}$ signifies the key matrix. The term $\mathbf{a}_{i,j}^L$ in equation (3) denotes the attention maps that have been derived, while d signifies the dimension associated with keys and queries. The terms FC^q and FC^k in (1) and (2) denote fully connected layers. The FFT in (1) and (2) denotes the Fast Fourier Transforms, while the IFFT in (3) signifies the Inverse Fast Fourier Transforms. The FFT and IFFT are calculated based on equations (4) and (5), respectively.

$$\text{FFT}(\mathbf{z}_{i,j}^L) = \hat{\mathbf{Z}}_{u,v}^L = \sum_{i=0}^{T}\sum_{j=0}^{D} \mathbf{z}_{i,j}^L e^{-2\pi i\left(\frac{ui}{T}+\frac{vj}{D}\right)} \tag{4}$$

$$\text{IFFT}(\mathbf{k}\mathbf{q}^\top) = \frac{1}{T^2}\sum_{u=0}^{T-1}\sum_{v=0}^{T-1} \hat{\mathbf{K}}_{u,v} e^{2\pi i\left(\frac{ut}{T}+\frac{vs}{T}\right)} \tag{5}$$

where u and v are the frequency indices in the Fourier domain in equation (4) and (5), t and s are the spatial domain indices in equation (5), and $\hat{\mathbf{K}}_{u,v}$ represents the frequency coefficients in equation (5). The feature maps $\bar{\mathbf{z}}^l$ are calculated as:

$$\bar{\mathbf{z}}_t^l = \sum_{t'=1}^{T} \mathbf{a}_{t,t'}^L \odot \mathbf{z}_{t'}^l \tag{6}$$

where $\odot$ denotes the element-wise multiplication in equation (6). The attention mechanism is visualized in Fig. 2.

B. Dataset Creation

We have created an indoor lab based dataset A18Robotics to test the efficacy of our method in different scenarios. The existing street scene change detection datasets contains very sparse changes, contrastingly, in the real world, the number of changes are dense. We recorded 10 videos of the lab environment in a fixed path with same starting and ending point. We manually segmented and paired frames from 2 videos for training purposes. The different objects in the dataset are divided into seven classes: human, object, chair, table, storage, background and base. The frequency of changes in the environment is very high in human, object and chair classes. The frequency of changes in the storage and

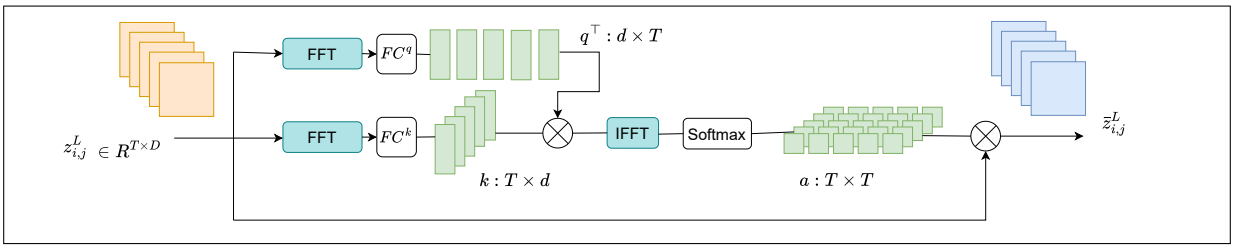

Fig. 2. Spectral Attention Network.

TABLE I

RESULTS OBTAINED IN DIFFERENT SETTINGS FOR DYNAMICEARTHNET DATASET.

	No Domain Shift				Temporal Domain Shift				Spatial Domain Shift			
Method	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU
SITSSCD	24	**30.57**	17.41	52.06	**25.6**	**36.0**	**15.3**	**61.7**	15.8	21.5	10.1	38.5
Ours	**24.36**	29.81	**18.91**	**55**	11.96	22.58	1.33	41.29	**32.22**	**41.28**	**23.16**	**63.14**

For DynamicEarthNet dataset: Results obtained in the No Domain Shift setting use the same training settings and splits for SITSSCD and our method. For the temporal and spatial domain-shift settings, the results for SITSSCD follow those reported in [31].

TABLE II

RESULTS OBTAINED IN DIFFERENT SETTINGS FOR MUDS DATASET.

	No Domain Shift				Temporal Domain Shift				Spatial Domain Shift			
Method	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU
SITSSCD	9.83	18.64	1.02	68.25	11.0	20.1	**1.9**	**71.1**	10.8	**20.8**	0.7	66.2
Ours	**11.24**	**20.91**	**1.56**	**69.70**	**14.32**	**28.03**	0.62	65.92	**11.44**	20.46	**2.42**	**69.10**

For MUDS dataset: Results obtained in the No Domain Shift setting use the same training settings and splits for SITSSCD and our method. For the temporal and spatial domain-shift settings, the results for SITSSCD follow those reported in [31].

table classes is low, and negligible in the background and base classes. The pairing process for the images obtained from the videos was performed manually. The recorded videos vary in length, and while manually pairing images from different temporal instances, the depicted locations may differ slightly.

C. Analysis Techniques

We trained and evaluated our approach on MUDS [12] and DynamicEarthNet [13] dataset in three different settings, along with the effect of noise on A18Robotics, DynamicEarthNet [13] and MUDS [12] datasets.

1) No Domain Shift: We divided the MUDS and DynamicEarthNet into 4 subsets and performed 4 fold cross validation on SITSSCD [31] method and our methodology. The subsets used for cross validation are not mentioned in [31], so we trained and evaluated SITSSCD [31] and our method on custom folds.

2) Temporal Domain Shift: For Temporal Domain Shift we divided both the datasets according to the SITSSCD [31]. The training set contains images from year 2018, validation set contains images from first half of year 2019 and test set contains images from second half of 2019 for DynamicEarthNet dataset and randomly assigned images from year 2019 into test and validation splits.

3) Spatial Domain Shift: We have used 5 fold cross validation for spatial domain shift settings and folds which were taken is same as taken in the SITSSCD [31] The cross validation sets used for training in no domain shift and spatial domain shift settings are similar to the train, validation and test sets used in [31].

4) Noise: We added noise to test sets for DynamicEarthNet, MUDS and A18Robotics. The noise is derived from unit Normal Gaussian distribution. For DynamicEarthNet and MUDS dataset the noisy dataset is tested with weights obtained in no domain shift settings. The unit Normal Gaussian Distribution can be described as $X \sim \mathcal{N}(0, 1)$.

D. Training Process

We trained SITSSCD [31] and our method using the focal loss and AdamW optimizer with a learning rate starting from 0 and decreasing it to 10^{-5} after 5000 warm-up steps. We train our method for 120–140 epochs with various configurations, retaining the checkpoint with the highest SCS (Semantic Change Segmentation) score on the validation set. For data augmentation during training, we have followed the techniques described in SITSSCD [31] by cropping the 1024 x 1024 images into 128 x 128 random patches and applying random transformations on patches. The spatial feature size D is kept fixed at 512, d is the dimension of key and query which is kept fixed at 4 and number of levels is kept at 4 as mentioned in [31]. The MUDS and DynamicEarthNet datasets are trained with training length of 12, batch size 8. For Trainings on DynamicEarthNet [13] and MUDS [12] two NVIDIA Quadro 8000 GPU's were used. A18Robotics dataset is trained on SITSSCD [31] and our method, with training length of 2, batch size 2 and GPU used for this is NVIDIA RTX A4000.

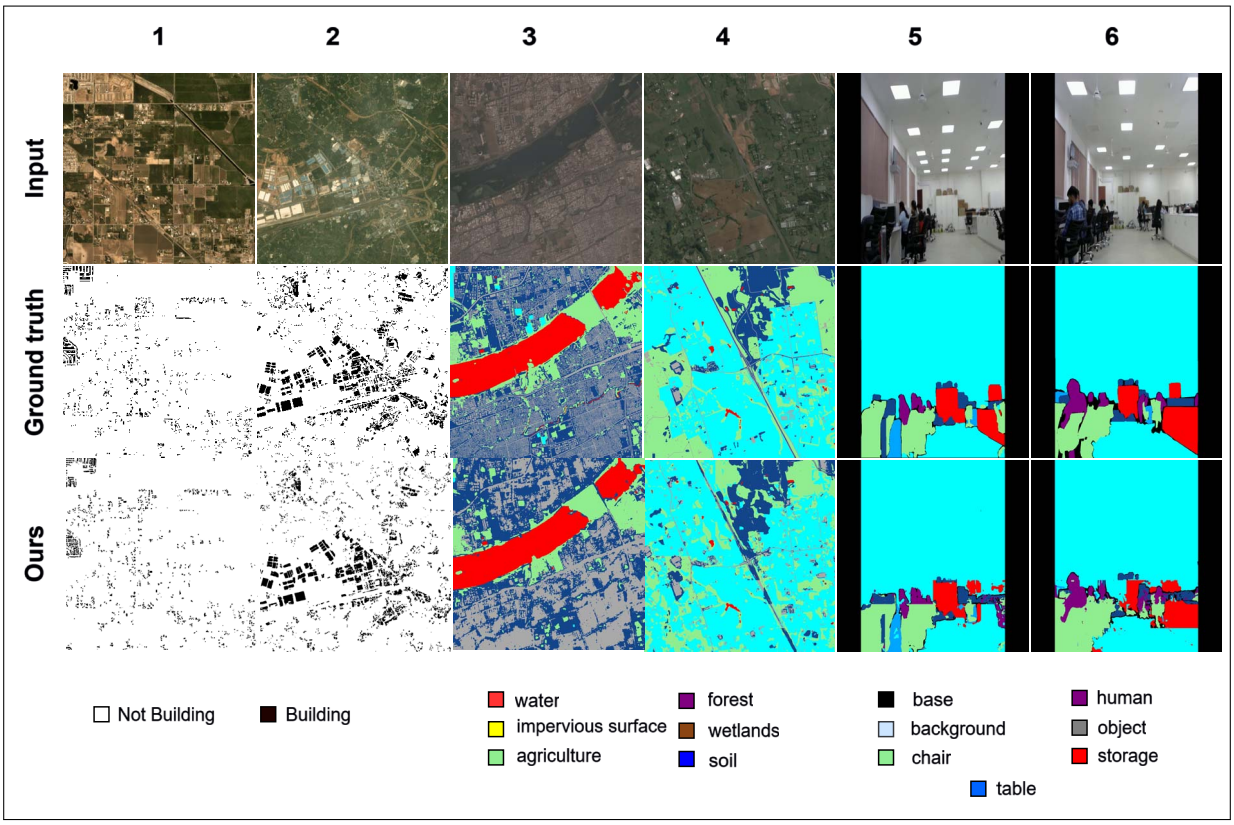

Fig. 3. Segmentation output. Images 1 and 2 are taken from the MUDS dataset, images 3 and 4 are taken from the DynamicEarthNet dataset and images 5 and 6 are taken from the A18Robotics dataset.

IV. Experiments and Results

A. Datasets and Metrics Used

We have used two satellite-based change detection datasets and our custom A18Robotics dataset for evaluation of our method and comparison with SITSSCD [31].

- **MUDS [12]:** It is a two-class dataset in which RGB images of 60 locations are collected over a period of 2 years (24 months) at the start of each month from January 2018 to December 2019. The 2 classes are building and not building. The dataset used for our work is taken from [31].
- **DynamicEarthNet [13]:** A six-class dataset with RGB and depth images collected over a period of 2 years with multiple images in each month and over 75 locations (of which 55 are used for spatial domain shift settings). The classes of this dataset are impervious surface, agriculture, forest, wetlands, soil and water. The 'snow' class is merged with water as it is present in very few images [31].
- **A18Robotics:** A seven-class dataset with RGB videos captured over different time intervals with each video is of varying length, but the start and end points of each video sequence are the same. This dataset includes seven classes (base, background, chair, human, object, storage, table), with each video lasting 45–60 seconds.

We have used the following metrics used in [13] and [31] to evaluate the performance of various methods, on the above stated datasets.

TABLE III
Results obtained for A18Robotics Dataset.

Method	SCS	SC	BC	mIoU
SITSSCD	36.85	14.16	**59.54**	29.54
Ours	**43.23**	**27.85**	58.62	**44.25**

- **Mean Intersection over Union (mIoU)**: Indicates the ability of the method to predict the correct semantic segmentation, irrespective of change.
- **Binary Change (BC)**: Depicts how well a method predicts semantic change.
- **Semantic Change (SC)**: Focuses on semantic prediction for pixels where a change actually occurs.
- **Semantic Change Segmentation (SCS)**: The average of BC and SC.

A high mIoU indicates that the method predicts segmentation maps closer to the ground truth. A high SC indicates that pixels where change occurs truly are predicted correctly. A high BC indicates that more correct changes are detected by the method. A high SCS indicates the overall capability of the model to predict semantic labels and changes correctly. We use SCS as our parameter for comparison of models during training as well as in testing results.

TABLE IV

RESULTS OBTAINED WHEN NOISE FROM $X \sim \mathcal{N}(0,1)$ IS ADDED IN THE TEST SET

Dataset	DynamicEarthNet				MUDS				A18Robotics			
Method	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU	SCS	SC	BC	mIoU
SITSSCD	16.02	21.71	10.32	39.92	5.78	11.42	0.14	59.07	14.52	6.99	26.48	20.12
Ours	**18.54**	**26.17**	**10.91**	**42.72**	**13.84**	**25.41**	**2.26**	**62.64**	**20.40**	**9.40**	**31.34**	**23.54**

The DynamicEarthNet and MUDS datasets are tested using weights obtained from the no-domain-shift setting.

B. Results and Discussions

Results of our proposed approach are compared with SITSSCD [31] in Table I, II, III, IV and V. Fig. 3 depicts the output of our method and ground truth with semantic labels for different images taken from MUDS, DynamicEarthNet and A18Robotics dataset. In Fig. 3 images 1 and 2 are taken from MUDS dataset, images 3 and 4 are taken from DynamicEarthNet dataset and images 5 and 6 are taken from A18Robotics dataset. The class labels for each datasets are represented below the images of respective dataset.

TABLE V

ACCURACY ACHIEVED ON DIFFERENT DATASETS ON DIFFERENT TRAINING METHODS.

Dataset	DynamicEarthNet	MUDS	A18Robotics
Method	Accuracy		
SITSSCD	63.59	74.83	41.67
Ours	**65.27**	**78.92**	**49.78**

For the DynamicEarthNet and MUDS datasets, accuracies are compared under the no-domain-shift setting.

1) Comparison with SOTA: Table I shows the comparison of our approach with SITSSCD [31] for DynamicEarthNet Dataset. Table II shows the comparison of our approach with SITSSCD [31] for MUDS Dataset. Table III shows the comparison of our approach with SITSSCD [31] for A18Robotics Dataset. Table IV represents the comparison of our method with when SITSSCD [31] when noise from Unit Normal Gaussian ($X \sim \mathcal{N}(0,1)$) is added in the images. Table V shows the accuracies achieved on our method and SITSSCD [31] for different datasets. Our method outperforms SITSSCD [13] in the spatial domain shift setting for DynamicEarthNet, as shown in Table I. All the evaluation metrics are doubled in this case. This improvement is attributed to the use of spectral attention. Our method also performs comparably in no domain shift settings. For the MUDS dataset in Table II, we get a high score for BC and mIoU in spatial domain shift settings which depicts our methods capability find change and semantically segment the image in a better way. Our method performs similarly to SITSSCD [31] in the no domain shift and temporal domain shift settings. Table III depicts that our method performs better than SITSSCD [31] in SCS (Semantic Change Segmentation), SC (Semantic Change) and mIoU (Mean Intersection over Union) while for BC (Binary Change) the results are similar. Table IV depicts that our method is robust to various types of changes which can add noise in the images. For DynamicEarthNet and MUDS dataset the performance of our method do not degrade much as compared to SITSSCD [31] when noisy images are tested on weights of no domain shift training method. For A18Robotics dataset our methods performs better than SITSSCD [31]. We compare the accuracies of different methods in Table V, our method gives accurate labels than SITSSCD [31] for DynamicEarthNet, MUDS and A18Robotics dataset. The accuracies for DynamicEarthNet and MUDS dataset are obtained in case of no domain shift settings.

2) Advantages: From Table I and II we can see that our method performs pretty impressive on spatial domain shift settings. This depicts that our method is robust to changes in environment during real world applications. In [31] authors have mentioned that the value of binary change scores are very low in almost all their training cases because there are very few changes present in DynamicEarthNet and MUDS dataset. This is observed in Table I and II in our case also. But for A18Robotics in Table III we observe a high binary change compared to semantic change. This shows our methods ability to adapt to real world street level scenarios where the changes present are large in number. Table IV depicts our methods' ability to sustain performance when there is a noise in images or videos which is generally the case in real world applications due to weather, seasons and time of the day when images or videos are captured. We trained our method on satellite and indoor lab images. Our method yields strong results in both the cases, which can be further adapted for drone based scene change detection in future.

V. CONCLUSION

A spectral attention network has been designed for semantic change detection across both satellite and street environments. Our approach significantly enhances the robustness of existing methods, particularly in the presence of domain shifts. Experimental evaluations show that our method outperforms existing techniques in scenarios with no domain shift and spatial domain shift while achieving comparable performance in temporal domain shift settings. Moreover we introduce a novel benchmarking dataset for indoor environmental scene change detection, characterized by densely occurring changes, making it highly relevant to real-world applications. Furthermore, the versatility of our method in both satellite and street-level change detection opens avenues for adaptive, efficient, drone-based monitoring systems.

ACKNOWLEDGMENT

The authors would like to thank Niharika, Vaishalee, Dharmendra Sharma and John Rebeiro for providing their help in image annotations to create dataset and for technical discussion.

REFERENCES

[1] K. Svendsen, T. Kristiansen, J. Martin, A. Askø, J. Bjørlo, R. Khosravanian, C. Holt, and F. Ruel, "Automated computer vision system for real-time detection of drilled cuttings and cavings," in *SPE/IADC Drilling Conference and Exhibition*. SPE, 2025, p. D031S024R002.

[2] H. Bhanushali, H. Maheshwari, S. Gohil, G. K. Gautam, and N. P. Koshti, "Enhancing human detection and counting efficiency through deep learning: A comprehensive approach," in *AIP Conference Proceedings*, vol. 3255, no. 1. AIP Publishing, 2025.

[3] J. P. Underwood, D. Gillsjö, T. Bailey, and V. Vlaskine, "Explicit 3d change detection using ray-tracing in spherical coordinates," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 4735–4741.

[4] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Transactions on Image Processing*, vol. 14, no. 3, pp. 294–307, 2005.

[5] I. Atheaux, T. Hughes, J. Peng, C. Johnson, C. Yarman, A. Medvedev, and S. Makarychev-Mikhailov, "Xrf-ftir data fusion and machine learning models for mineralogical analysis of solid mineral mixtures and solid dispersions in drilling fluids," in *SPE/IADC Drilling Conference and Exhibition*. SPE, 2025, p. D021S014R001.

[6] A. Kutakova, V. Kokhan, and D. Bocharov, "Dataset and feature point-based matching algorithm for satellite and aerial remotely sensed images," in *International Conference on Machine Vision*, vol. 13517. SPIE, 2025, pp. 36–43.

[7] A. Aboah, "A vision-based system for traffic anomaly detection using deep learning and decision trees," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4207–4212.

[8] Y. Jiang, W. Wang, and C. Zhao, "A machine vision-based realtime anomaly detection method for industrial products using deep learning," in *Chinese Automation Congress (CAC)*. IEEE, 2019, pp. 4842–4847.

[9] K. R. Prabhakar, A. Ramaswamy, S. Bhambri, J. Gubbi, R. V. Babu, and B. Purushothaman, "Cdnet++: Improved change detection with deep neural network feature correlation," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[10] E. D. Çatmabacak and İ. Çetinkaya, "Deep learning algorithms for detecting fractured instruments in root canals," *BMC Oral Health*, vol. 25, p. 293, 2025.

[11] S. Chen, K. Yang, and R. Stiefelhagen, "Dr-tanet: Dynamic receptive temporal attention network for street scene change detection," in *IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 502–509.

[12] A. Van Etten, D. Hogan, J. M. Manso, J. Shermeyer, N. Weir, and R. Lewis, "The multi-temporal urban development spacenet dataset," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 6398–6407.

[13] A. Toker, L. Kondmann, M. Weber, M. Eisenberger, A. Camero, J. Hu, A. P. Hoderlein, Çaglar Senaras, T. Davis, D. Cremers, G. B. Marchisio, X. X. Zhu, and L. Leal-Taix'e, "Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21 126–21 135, 2022.

[14] J.-M. Park, J.-H. Jang, S.-M. Yoo, S.-K. Lee, U.-H. Kim, and J.-H. Kim, "Changesim: Towards end-to-end online scene change detection in industrial indoor environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8578–8585.

[15] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[16] V. García Rubio, J. A. Rodrigo Ferrán, J. M. Menéndez García, N. Sánchez Almodóvar, J. M. Lalueza Mayordomo, and F. Álvarez, "Automatic change detection system over unmanned aerial vehicle video sequences based on convolutional neural networks," *Sensors*, vol. 19, no. 20, 2019.

[17] J. Santos, H. Dinkel, J. Di, M. Moreira, B. Coltin, P. Borges, T. Smith, and O. Alexandrov, "Unsupervised change detection for space habitats using 3d point clouds," in *AIAA SCITECH 2024 Forum*. American Institute of Aeronautics and Astronautics, Jan. 2024.

[18] K. Sakurada, M. Shibuya, and W. Wang, "Weakly supervised silhouette-based semantic scene change detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6861–6867.

[19] C.-P. Chen, J.-W. Hsieh, P.-Y. Chen, Y.-K. Hsieh, and B.-S. Wang, "Saras-net: Scale and relation aware siamese network for change detection," in *AAAI Conference on Artificial Intelligence*, vol. 37, Jun. 2023, pp. 14 187–14 195.

[20] G.-H. Wang, B.-B. Gao, and C. Wang, "How to reduce change detection to semantic segmentation," *Pattern Recognition*, vol. 138, p. 109384, 2023.

[21] D. Li, L. Chen, C.-Z. Xu, and H. Kong, "Umad: University of macau anomaly detection benchmark dataset," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 5836–5843.

[22] C.-J. Lin, S. Garg, T.-J. Chin, and F. Dayoub, "Robust scene change detection using visual foundation models and cross-attention mechanisms," 2025.

[23] Z. Lu, J. Ye, and J. Leonard, "3dgs-cd: 3d gaussian splatting-based change detection for physical object rearrangement," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2662–2669, 2025.

[24] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4852–4861, 2021.

[25] M. Tarasiou, E. Chavez, and S. Zafeiriou, "Vits for sits: Vision transformers for satellite image time series," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 418–10 428, 2023.

[26] U. Katsura, K. Matsumoto, A. Kawamura, T. Ishigami, T. Okada, and R. Kurazume, "Spatial change detection using voxel classification by normal distributions transform," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2953–2959.

[27] J. Fu, Y. Du, K. Singh, J. B. Tenenbaum, and J. J. Leonard, "Robust change detection based on neural descriptor fields," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 2817–2824.

[28] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "Spectformer: Frequency and attention is what you need in a vision transformer," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 9543–9554.

[29] E. Reis, T. Akilan, and M. Khalid, "Phasor-driven acceleration for fft-based cnns," 2024.

[30] X. Qin, Y. Zhang, and Y. Dong, "Domain alignment dynamic spectral and spatial feature fusion for hyperspectral change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 557–568, 2025.

[31] E. Vincent, J. Ponce, and M. Aubry, "Satellite image time series semantic change detection: Novel architecture and analysis of domain shift," 2024.