2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

FR-CDNet: Unified Scene Change Detection Model across Viewpoint Variations and Different Temporal Conditions

Yilin Peng[1], Yingchun Fu[2,†], Xiangru Li[3], Zhenhao Li[2], Shuqi Chen[2], Shunping Ji[4]

***Abstract*— Scene Change Detection (SCD) is a critical task for building smart cities, yet its practical application faces dual challenges: existing methods typically rely on temporal conditions present in the training data and the ideal assumption of small viewpoint differences. Consequently, they struggle to handle the common and significant viewpoint variations in real-world scenarios and exhibit strong sensitivity to temporal conditions, leading to drastic performance degradation under unseen temporal settings. To address these challenges, we propose the Fusion-Refinement Change Detection Network (FR-CDNet). By modeling correspondences between objects and preserving spatial prior information from ideally aligned scenes during the disentangled processing of different temporal directions, our network achieves a unified handling of varying degrees of viewpoint variations and different temporal conditions—a capability existing methods lack. Furthermore, FR-CDNet can automatically distinguish the temporal attribution of change entities to better support downstream tasks. To better evaluate performance in real-world settings, we further construct the URSCD dataset, which includes larger viewpoint differences and more diverse change scenarios. Extensive experiments demonstrate the universal scene detection capability of our method: it achieves significant improvement in F1-score on unaligned scenes while maintaining performance comparable to SOTA on aligned scenes. Ablation studies further demonstrate that the proposed framework can be migrated to enhance various mainstream models, effectively eliminating temporal condition dependency while improving overall performance.**

I. Introduction

Scene Change Detection (SCD), as a critical perception capability of autonomous agents, plays an essential role in building smart cities. It can be widely applied in visual surveillance [1], mobile robotics [2], [3], and autonomous vehicles [4], providing technical support for downstream tasks such as anomaly detection [5], infrastructure inspection [6], map updating [7], and natural disaster damage assessment [8]. With large-scale street-view imagery sources (e.g. Google Street View, Baidu Street View) gradually replacing costly field surveys, automated SCD has become an important means for scalable urban monitoring [9].

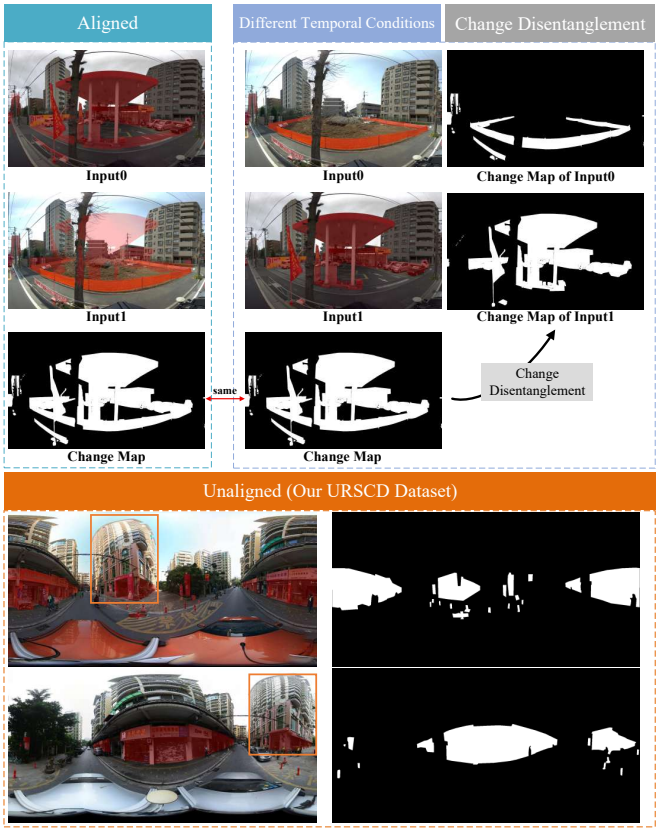

Fig. 1. **Illustration of different scenarios in SCD.** Existing methods experience substantial performance drops in unaligned scenes with significant viewpoint variations and under reversed temporal order.

In real-world deployment, SCD systems must adapt to various complex conditions, including illumination changes, noise interference, temporal condition inconsistency, and unaligned image pairs caused by different camera viewpoints or imperfect matching (Fig. 1). While the first two issues can often be mitigated by employing powerful visual encoders that extract robust deep features[10], [11], [12], existing methods still struggle with viewpoint variation and temporal condition changes. Many current SCD approaches rely on the ideal assumptions of image alignment or small viewpoint variation, and presume consistent temporal conditions between training and testing [13], [14]. As a result, their performance degrades significantly in real-world applications, manifesting as a significant performance drop when viewpoint differences increase or when the input image pair

*This work was supported by the National Natural Science Foundation of China (Grant 42571390, 42071399) and Science and Technology Projects of Xizang Autonomous Region, China (XZ202501ZY0091, XZ202301ZY0021G) and Basic and Applied Basic Research Foundation of Guangdong Province (Grant No.2025A1515011807)

† Corresponding author.
[1] Yilin Peng is with the Beidou Research Institute, South China Normal University, Foshan, China yilinpeng@m.scnu.edu.cn
[2] Yingchun Fu, Zhenhao Li, Shuqi Chen are with the School of Geography, South China Normal University, Guangzhou, China {fuyc, 2024023073, 2024023059}@m.scnu.edu.cn
[3] Xiangru Li is with the School of Computer Science, South China Normal University, Guangzhou, China lixiangru@m.scnu.edu.cn
[4] Shunping Ji is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China jishunping@whu.edu.cn

order is reversed, producing temporal conditions not seen in the training set. Some recent methods claim to extend their applicability to unaligned scenes [15], [16], [17], [18], [14], but they do not sufficiently account for geometric or object information distortion introduced by image or feature transformations, leading to unreliable performance under significant viewpoint differences. Furthermore, these methods typically struggle to maintain detection capability for both aligned and unaligned scenes simultaneously, which remains a critical bottleneck for large-scale SCD applications.

To address these challenges, we propose the **F**usion-**R**efinement **C**hange **D**etection Network (FR-CDNet), which achieves universal scene detection capability for both aligned and unaligned settings without being affected by temporal conditions. Notably, FR-CDNet can distinguish the temporal attribution of change entities (change disentanglement) without requiring additional labels (Fig. 1). This property improves scalability and reduces annotation costs, making it possible to seamlessly support downstream tasks such as semantic or instance-level change detection—where traditional methods typically require extra post-processing models and training data [15].

FR-CDNet introduces a novel perspective. Unlike traditional methods that explicitly distinguish between pre-change and post-change images, we reformulate SCD as a mutual comparison process to identify relative changes, allowing image pairs to be interchangeable as "before" or "after". This perspective introduces an inductive bias for temporal-condition invariance into the network and provides a robust comparison strategy for unaligned scenes. Specifically, we co-design the Spatial Prior-guided Cross-Attention (SPCA) module and the Cross Fusion architecture. The SPCA module leverages cross attention to implicitly establish correspondences between images, performing feature alignment to handle viewpoint differences, while injecting spatial priors to guide attention toward features at the same location, thereby preserving detection capability in aligned scenes. Meanwhile, the Cross Fusion architecture employs bidirectional feature fusion for mutual comparison: When features from image I_0 are transformed and compared in the space of I_1, a reverse transformation-and-comparison process is performed in parallel, enabling complementary verification. This design ensures that change entities from both time points are correctly preserved and represented, thereby preventing object information distortion caused by one-way transformations using SPCA.

To evaluate real-world applicability, we build a new benchmark dataset, URSCD, which includes more significant viewpoint differences and more diverse change scenarios, thereby posing greater challenges for SCD algorithms.

In summary, our contributions are as follows:

- We propose FR-CDNet, which achieves robust detection in both aligned and unaligned scenes without temporal dependency. Moreover, FR-CDNet further enables temporal disentanglement of change entities without requiring additional annotation data, reducing labeling costs and facilitating downstream applications.
- We demonstrate that our framework can be seamlessly transferred to a wide range of mainstream baselines, improving overall performance without introducing additional parameters, thus offering new perspectives and insights for future SCD methods.
- We construct a new challenging benchmark dataset, URSCD, with larger viewpoint differences and diverse scenarios for real-world evaluation.
- We will release the source code of FR-CDNet (including transferred versions of baseline models) along with the dataset to promote large-scale application of SCD techniques : `https://github.com/hhhaaaa/FR-CDNet`.

II. Related Work

Change detection aims to identify changed regions between image pairs or sequences, and has been extensively explored in remote sensing monitoring, video analysis, and natural scene understanding. Remote sensing change detection [19], [20], [21], [22] focuses on identifying land surface changes from satellite or aerial imagery, while video change detection [23], [24], [25] emphasizes foreground-background segmentation of moving objects across consecutive frames. In contrast, our work focuses on Scene Change Detection (SCD) [7], [26], [27], [28], [16], [29], which deals with ground-level imagery such as street views aims to capture distinct changes of local entities like vehicles, pedestrians, and buildings, etc. This type of data is typically collected by mobile platforms such as vehicles and robots, inevitably containing viewpoint differences and background noise, which increases the difficulty of detection.

Traditional SCD methods, including pixel-based [30], [31] and object-based [32], [33], [34] approaches, often rely on image differencing and segmentation algorithms. However, they are highly sensitive to noise, registration errors, and pseudo-changes. With the development of deep learning, traditional methods have been gradually superseded. Powerful image encoders and their variants are now widely used to extract discriminative features, combined with various feature comparison algorithms to accomplish SCD tasks [35], [36], [37], [28], [38], [14], demonstrating promising results. Nevertheless, these methods typically assume that image pairs are captured from the identical viewpoints and positions, leading to significant performance degradation when viewpoint differences exist. Some studies have attempted to introduce registration mechanisms in feature space by leveraging feature correlation [15], [16], [14] or additional auxiliary optical flow data [26], [17], [18]. Although these approaches improve adaptability to unaligned scenes to some extent, they often overlook the object distortion caused by feature transformation. Consequently, they still struggle to maintain robust performance for both aligned and unaligned scenes and exhibit limited stability under significant viewpoint differences [14]. Moreover, existing methods usually depend on the temporal conditions of the training data, with explicitly defined change directions (e.g., pre-/post-change).

When tested under unseen temporal conditions like reversed temporal order, their performance degrades notably [13].

Some works have further explored Semantic Scene Change Detection (SSCD) [39], [15], [40], [13], which not only identifies changes but also predicts semantic categories. SSCD often relies on change disentanglement before semantic classification. [15] proposed a weakly supervised SSCD framework to reduce annotation cost by leveraging post-processing models and synthetic datasets for change disentanglement. In contrast, our method achieves end-to-end change disentanglement jointly with change detection, eliminating the need for additional steps or training data. This further reduces research costs, improves scalability, and provides more efficient support for SSCD.

III. Method

We introduce FR-CDNet, a dual-branch disentangled change detection architecture. It takes a pair of images $\mathbf{I}_0, \mathbf{I}_1 \in \mathbb{R}^{3\times H\times W}$ as input and outputs binary change masks $\mathbf{M}_0, \mathbf{M}_1, \mathbf{M} \in \{0,1\}^{H\times W}$, which indicate the change entities in $\mathbf{I}_0, \mathbf{I}_1$, and the union of these changed regions, respectively, regardless of viewpoint differences or temporal order. As illustrated in Fig. 2, FR-CDNet primarily consists of two components: Cross Fusion and SPCA, which will be detailed in Sec. III-A and III-C, respectively.

A. Cross Fusion

Cross Fusion adopts a dual-branch structure to perform mutual comparisons via forward and backward paths. The forward path compares $\mathbf{I}_0$ against $\mathbf{I}_1$ to detect change entities in $\mathbf{I}_0$, while the backward path performs the opposite comparison to detect change entities in $\mathbf{I}_1$. Given feature representations $\mathbf{F}_0, \mathbf{F}_1 \in \mathbb{R}^{c\times h\times w}$ of the input image pair extracted by a visual encoder, a feature remapping function ξ generates new feature representations to handle viewpoint variation, and a feature fusion operator fuse computes the differences between features. The two paths define different orders of remapping and fusion. Specifically, the forward path detection branch is formulated as:

$$\mathbf{m}_0 = \text{fuse}(\mathbf{F}_0, \mathbf{F}_1^{warped}) \tag{1}$$

$$\mathbf{F}_1^{warped} = \xi_{\mapsto \mathbf{F}_0}(\mathbf{F}_1) \tag{2}$$

Here, $\xi_{\mapsto \mathbf{F}_0}(\mathbf{F}_1)$ remaps $\mathbf{F}_1$ into the feature space of $\mathbf{F}_0$, yielding $\mathbf{F}_1^{warped}$ aligned to $\mathbf{F}_0$. As illustrated in Fig. 2, the forward path preserves object location and shape information in $\mathbf{F}_0$ while warping the objects in $\mathbf{F}_1$. Similarly, the backward path follows the process in reverse order:

$$\mathbf{m}_1 = \text{fuse}(\mathbf{F}_1, \mathbf{F}_0^{warped}) \tag{3}$$

$$\mathbf{F}_0^{warped} = \xi_{\mapsto \mathbf{F}_1}(\mathbf{F}_0) \tag{4}$$

To correctly distinguish the direction of feature comparison, we explicitly constrain the fuse operator to be non-commutative ($\exists \mathbf{X}, \mathbf{Y} \in \mathbb{R}^{c\times h\times w}$ such that $\text{fuse}(\mathbf{X}, \mathbf{Y}) \neq \text{fuse}(\mathbf{Y}, \mathbf{X})$), preventing homogeneous outputs from the two paths.

The final change maps $\mathbf{m}_0, \mathbf{m}_1 \in \mathbb{R}^{2\times H\times W}$ from the two branches contain two channels: unchanged (channel 0) and changed (channel 1). Their binarized masks $\mathbf{M}_0, \mathbf{M}_1 \in \{0,1\}^{H\times W}$ indicate the change entities in $\mathbf{I}_0$ and $\mathbf{I}_1$, respectively. To achieve weakly-supervised change disentanglement, we design a Combine Operator to generate the union change map $\mathbf{m}$:

$$\mathbf{m} = \left[\min(\mathbf{m}_0^0, \mathbf{m}_1^0), \max(\mathbf{m}_0^1, \mathbf{m}_1^1)\right] \tag{5}$$

where $[\cdot]$ denotes channel-wise concatenation and the superscript denotes the channel index.The Combine Operator maintains consistency by ensuring that the union map covers all regions predicted as changed in $\mathbf{m}_0$ and $\mathbf{m}_1$, ensuring correct gradient flow. This design enables weakly supervised learning by leveraging the union mask as the only supervision signal. Cross Fusion defines a general SCD architecture where both branches preserve the object information of their respective target image, avoiding distortions caused by one-way remapping. Moreover, the bidirectional comparison removes the distinction based on temporal conditions. It can be proven that this architecture yields consistent detection results regardless of the input temporal order of the image pair.

B. Fusion Operator

For the fusion operator, we slightly modify the Merge Temporal Features (MTF) module proposed by C3PO [38]. The original MTF is commutative, which conflicts the design constraint of Cross Fusion. Therefore, we introduce the Non-commutative MTF (Nc-MTF), defined as:

$$\text{Nc-MTF}(\mathbf{X}, \mathbf{Y}) = \text{Conv}(\mathbf{X}) + \text{Conv}(\text{ReLU}(\mathbf{X} - \mathbf{Y})) \tag{6}$$

Nc-MTF produces different outputs for different input orders, thereby explicitly encoding the directionality of feature comparison.

C. Spatial Prior-guided Cross Attention (SPCA)

We propose the SPCA module as the feature remapping function in our network. As shown in Fig. 2, unlike standard Cross Attention[14], SPCA explicitly incorporates spatial priors into correlation modeling. This allows it to maintain attention on features at the same location when searching for cross-image correspondences, thereby adapting to both aligned and unaligned scenes. Specifically, for a given feature location (i, j), taking the generation of $\mathbf{F}_1^{warped}(i, j)$ as an example, SPCA first computes the global correlation between the query $\mathbf{F}_0(i, j) \in \mathbb{R}^{c\times 1}$ and all locations in key $\mathbf{F}_1$:

$$\mathbf{Corr}(k, \ell) = \mathbf{F}_0(i, j)^\top \mathbf{F}_1(k, \ell) \tag{7}$$

Each location (k, ℓ) in $\mathbf{Corr} \in \mathbb{R}^{h\times w}$ represents the correlation score between $\mathbf{F}_0(i, j)$ and $\mathbf{F}_1(k, \ell)$. In an ideally aligned scenario, we expect $\mathbf{F}_0(i, j)$ to correspond to $\mathbf{F}_1(i, j)$, i.e., $\mathbf{Corr}(i, j)$ should be significantly higher than other locations. This property is referred to as the spatial prior. To handle both aligned and unaligned scenes, we introduce a controllable spatial prior:

$$\mathbf{A} = \text{Softmax}(\text{keep}_{(i,j)}(\text{top-}k(\mathbf{Corr}))) \tag{8}$$

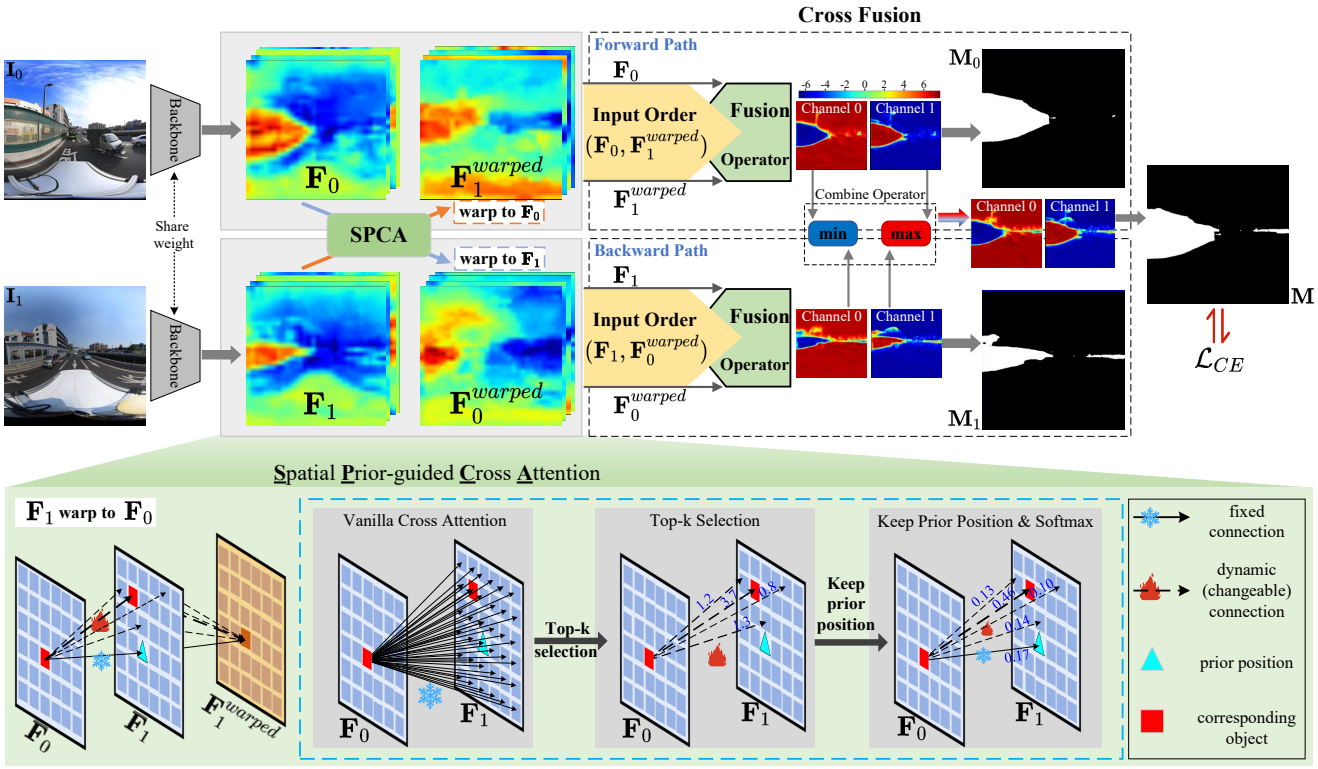

Fig. 2. **Architecture Overview.** Our model mainly consists of two components: (1) The SPCA module warps the extracted image features. (2) The Cross Fusion architecture with two paths. The forward path preserves the object information of $\mathbf{I}_0$ and fuses it with the warped features of $\mathbf{I}_1$ to indicate the change entities in $\mathbf{I}_0$, while the backward path performs the reverse process. Finally, the Combine Operator generates a union mask covering all changed regions to enable weakly supervised change disentanglement.

The top-k operation retains the k largest values in $\mathbf{Corr}$ and masks out the rest, while $\text{keep}_{(i,j)}$ always preserves the score at location (i,j) to ensure that the prior position is not masked. Subsequently, Softmax normalizes the scores at these retained locations, producing a sparse attention weight matrix $\mathbf{A}$. Finally, $\mathbf{F}_1^{warped}(i,j)$ is obtained by the weighted sum of original $\mathbf{F}_1$:

$$\mathbf{F}_1^{warped}(i,j) = \sum_{(k,\ell)} \mathbf{A}(k,\ell) \cdot \mathbf{F}_1(k,\ell) \tag{9}$$

In Eq. 8, the top-k operation allows us to control the strength of the spatial prior. A larger k weakens the prior as attention weights are distributed over more locations. When $k = h \times w$, SPCA reduces to standard Cross Attention. Conversely, a smaller k strengthens the prior, and when $k = 0$, SPCA reduces to an identity mapping.

D. Loss Function

We use the softmax cross-entropy loss as the objective function. Although our network outputs disentangled change results, we compute the loss only on the union change map $\mathbf{m}$, thus requiring no additional labels for change disentanglement. Moreover, since change and unchanged regions are highly class imbalanced, following prior work [38], we use weighted cross-entropy loss, where weights are determined by the pixel ratios of changed and unchanged classes in each dataset.

IV. Experiments

A. Datasets and Preprocessing

We evaluate our method on three datasets: VL-CMU-CD [7], PSCD [15], and the URSCD dataset (Urban Street-view Scene Change Detection) that we collected and annotated.

VL-CMU-CD contains 1,362 registered perspective image pairs. Following prior work, we resize images to 512×512 and split them into 933 pairs for training and 429 pairs for testing. Following [38], the training set is further augmented to 3,732 pairs through rotation.

PSCD consists of 770 panoramic street-view image pairs with only minor viewpoint variations. Following the official protocol [15], we resize each image to 2048×512, and then apply a sliding window cropping along the width with stride 112 and window size 512×512. Each cropped patch is further augmented through rotation. We adopt 5-fold cross-validation, where each training set contains 36,960 image pairs of size 512×512, and each test set contains 154 images of size 2048×512, serving as the aligned test set. To evaluate performance under significant viewpoint differences, we construct an unaligned version of the test set to simulate real-world conditions: for each original test image at time t_0, the corresponding t_1 image is randomly flipped horizontally or vertically. Since PSCD provides manually annotated change disentanglement masks $\mathbf{M}_0, \mathbf{M}_1$, we apply

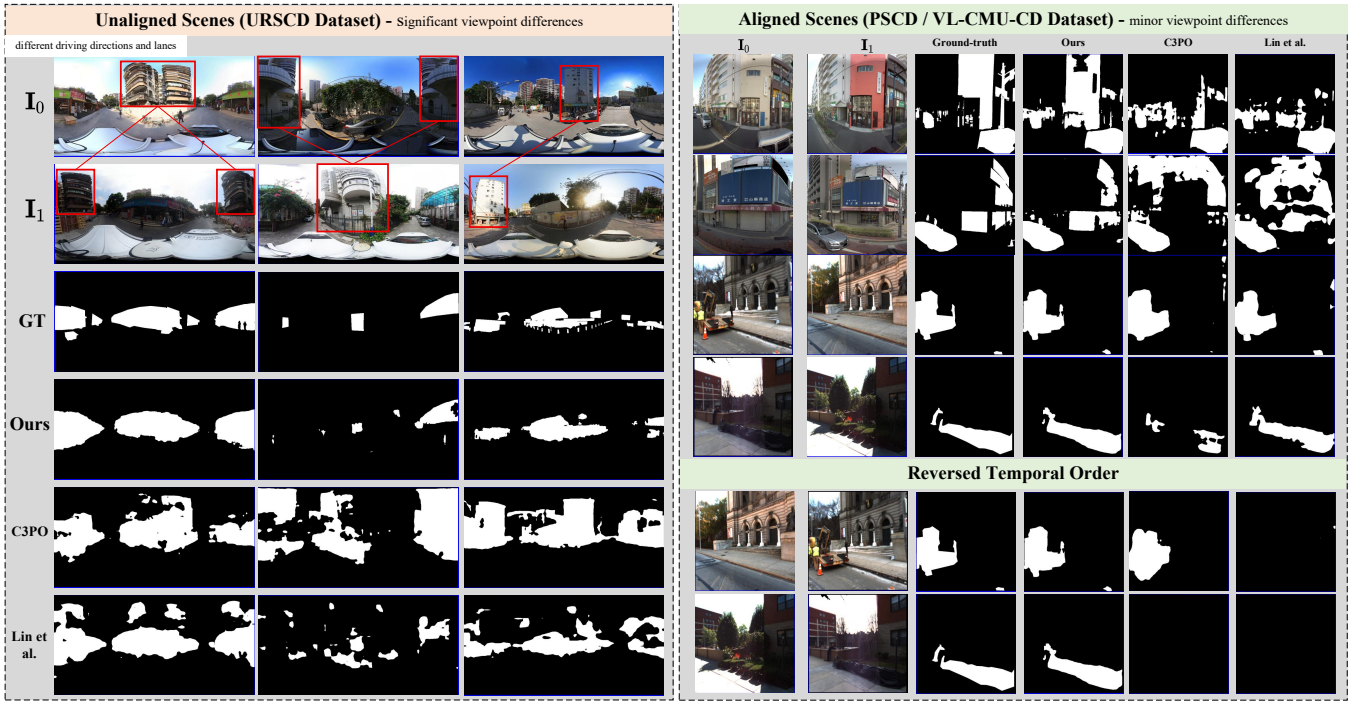

Fig. 3. **Qualitative Comparison Results.** Our method performs robustly under both varying degrees of viewpoint differences and different temporal conditions, whereas previous methods fail to handle these scenarios simultaneously.

the same flipping to $\mathbf{M}_1$ and take the union of $\mathbf{M}_0$ and the flipped $\mathbf{M}_1$ to generate the union change mask $\mathbf{M}$ for training. We report results on both the aligned and unaligned versions of the test set.

URSCD contains 686 panoramic street-view image pairs with manually annotated change masks $\mathbf{M}_0, \mathbf{M}_1$. The original image resolution is 2048×1024. Among them, 98 pairs exhibit significant viewpoint differences (e.g., different capture positions, or differences in driving lanes and directions), which we use as the unaligned test set. URSCD contains more diverse change scenarios and generally larger viewpoint differences. Notably, due to the large temporal gaps between paired images, pedestrians and vehicles are almost certainly different, making changes involving these entities reduce to standard segmentation tasks. Therefore, URSCD excludes changes related to pedestrians and vehicles. Similar to PSCD, we resize the images to 1024×512, and then apply sliding window cropping with stride 128 and window size 512×512, followed by rotation augmentation. We adopt 5-fold cross-validation, where each training set contains 9,400 image pairs of size 512×512, and each aligned version test set contains 118 image pairs of size 1024×512. Likewise, results are reported on both aligned and unaligned versions of the test set.

B. Evaluation Metrics

Following previous work [38], [28], [16], we adopt F1-score as the evaluation metric. F1-score is calculated upon Precision and Recall (TP : True Positives; FP : False Positives; FN : False Negatives):

$$\text{F1-Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{10}$$

where $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$ and $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$. In binary change detection task, changed regions represent positives, while background (unchanged regions) represent negatives.

C. Implementation Details

We use pretrained DINO-V2 (base) [12] as our backbone. For SPCA (Eq. 8), we set $k = 4$. We optimize our model using the Adam optimizer with an initial learning rate of 10^{-4} and employ a cosine annealing schedule for learning rate decay. For reproducibility, batch size is set to 4 for all datasets. For training hardware, all experiments are conducted using a single NVIDIA H100 GPU.

D. Comparative Studies

We conduct systematic evaluation of model performance under significant viewpoint differences, covering both simulated and real conditions. In contrast, existing methods are typically tested only under minor viewpoint variation, such as small-angle rotations ($< 15°$) or limited translations (< 50 pixels), which fail to reflect the challenges of real-world environments. Fig. 3 illustrates an intuitive comparison between aligned and unaligned scenes: the former corresponds to minor viewpoint differences, while the latter comes from our URSCD dataset, showcasing real-world examples of significant viewpoint differences. Moreover, existing research commonly neglects evaluation of temporal condition dependency. However, a robust and reliable change detection algorithm

should focus on detecting changes themselves, rather than relying on other unpredictable external conditions.

Method	Backbone	F1-Score(%)		
		Align	Reverse	Avg.
FC-EF	U-Net	43.8	18.3	31.1
Siam-Conc	U-Net	64.5	23.0	43.8
Siam-Diff	U-Net	65.2	31.8	48.5
DR-TANet	ResNet-18	68.8	7.4	38.1
CSCDNet	ResNet-18	75.7	4.8	40.3
ChangeNet	ResNet-50	60.3	2.4	31.4
C-3PO	VGG-16	78.2	15.2	46.7
Lin et al.[14]	DINO-V2	79.2	3.7	41.5
FR-CDNet(Ours)	DINO-V2	**80.3**	**80.3**	**80.3**

TABLE I

QUANTITATIVE RESULTS ON VL-CMU-CD. *Reverse represents reversed temporal order on aligned test set. Avg. represents average metric across different temporal conditions. Our method is not affected by temporal conditions.*

Method	F1-Score(%)				
	Avg1.	Unalign	Align	Reverse	Avg2.
FC-EF	41.7	44.0	39.3	36.0	37.7
Siam-Conc	44.7	45.3	44.0	42.4	43.2
Siam-Diff	46.8	47.3	46.2	45.3	45.8
DR-TANet	53.2	51.2	55.1	36.6	45.9
CSCDNet	55.1	55.1	55.0	36.6	45.8
ChangeNet	52.0	55.2	48.7	30.5	39.6
C-3PO	59.9	58.6	61.2	61.2	61.2
Lin et al.[14]	60.0	60.5	59.4	47.7	53.6
FR-CDNet(Ours)	**67.4**	**69.3**	**65.5**	**65.5**	**65.5**

TABLE II

QUANTITATIVE RESULTS ON URSCD. *Reverse represents reversed temporal order on aligned test set. Avg1. represents the average metric across aligned and unaligned test set. Avg2. represents average metric across different temporal conditions. Our method consistently maintains robust performance under all settings.*

We compare the proposed FR-CDNet with a series of representative methods, including both classical convolution-based architectures and recent transformer-based solutions: FC-EF [37], Siam-Conc [37], Siam-Diff [37], DR-TANet [16], CSCDNet [15], ChangeNet [27], C-3PO [38] and the recent method Lin et al.(2025) [14]. To ensure fairness, we follow the official training specifications for each method.

Tables I-III report quantitative results on three benchmark datasets. The results demonstrate that FR-CDNet significantly outperforms existing methods in unaligned scenes (Unalign column) and under unseen temporal conditions (Reverse column), while maintaining performance comparable to or slightly better than state-of-the-art methods in aligned scenes (Align column). This verifies both stability and universality of our method. Existing methods generally exhibit dependency on temporal conditions, especially on VL-CMU-CD dataset, where their performance drops drastically by $25.5\% \sim 75.5\%$ F1-score when the temporal order is reversed. In contrast, FR-CDNet remains stable and outperforms the second-best method by 48.5% F1-score.

Method	F1-Score(%)				
	Avg1.	Unalign	Align	Reverse	Avg2.
FC-EF	50.7	40.1	61.2	59.7	60.5
Siam-Conc	54.8	41.2	68.3	67.6	68.0
Siam-Diff	56.0	43.3	68.6	68.4	68.5
DR-TANet	63.1	47.7	78.5	75.6	77.1
CSCDNet	61.8	45.4	78.1	76.3	77.2
ChangeNet	57.9	54.1	61.6	58.3	60.0
C-3PO	69.6	57.4	**81.7**	**81.7**	**81.7**
Lin et al.[14]	66.9	59.3	74.4	72.9	73.7
FR-CDNet(Ours)	**78.6**	**75.5**	81.6	81.6	81.6

TABLE III

QUANTITATIVE RESULTS ON PSCD. *Our method maintains detection performance in both aligned and unaligned scenes, while other methods fail to do so.*

In unaligned scenes, existing methods also struggle to maintain detection performance. For instance, on PSCD dataset, performance under unaligned setting generally decreases by $14.9\% \sim 30.8\%$. It is worth noting that although Lin et al. outperform C-3PO in unaligned scenes, they fall significantly behind in aligned scenes (-7.3%). In contrast, FR-CDNet demonstrates robust performance in both settings: it leads the second-best method by 16.3% F1-score in unaligned setting while maintaining performance on par with SOTA in aligned setting. On URSCD dataset, since the two test sets are independent and the unaligned version contains a larger proportion of changed areas, the overall scores in Unalign column are generally higher than those in Align. FR-CDNet also shows consistent superiority on this dataset, outperforming the second-best method by 4.3% and 8.8% under aligned and unaligned scenes, respectively, further verifying its cross-scene robustness.

Finally, it should be noted that we did not introduce viewpoint variation data augmentation (e.g., applying different random rotations to image pairs) during training, as existing methods consistently exhibit performance degradation under such settings, particularly suffering over a 10% performance drop in aligned scenes. This observation is also reported in [14].

E. Qualitative Comparison

Fig. 3 presents qualitative comparisons between state-of-the-art methods and our FR-CDNet under different settings. GT (Ground-Truth) denotes the union change mask label, which represents the union of changed regions in the image pair. The visualization results clearly confirm the robustness of our method in handling various degrees of viewpoint differences. The clearly bounded and temporally consistent change detection results (Reverse Order) further demonstrate the cross-temporal-condition stability of our approach.

F. Ablation Study: Cross Fusion and Nc-MTF

Tables IV and V validate the generality and transferability of the Cross Fusion architecture, as well as the impact of the non-commutativity design constraint on change disentanglement.

Model	Unalign			Align			Reverse		
	M	M_0	M_1	M	M_0	M_1	M	M_0	M_1
FC-EF(CF)	**40.5**	26.5	30.1	**63.8**	46.3	48.6	**63.8**	48.6	46.3
Siam-Conc(CF)	**41.4**	25.3	26.6	**68.9**	32.1	35.4	**68.9**	35.4	32.1
Siam-Diff(CF)	**44.8**	27.0	29.1	**68.7**	46.9	51.4	**68.7**	51.4	46.9
DR-TANet(CF)	**58.1**	52.6	51.9	**80.0**	68.6	70.8	**80.0**	70.8	68.6
CSCDNet(CF)	**46.9**	21.1	23.8	**79.8**	36.0	37.6	**79.8**	37.6	36.0
C-3PO(CF)	**65.2**	58.1	58.7	**83.6**	73.8	76.0	**83.6**	76.0	73.8
Lin et al.[14]	**63.9**	59.5	58.5	73.5	61.2	62.1	**73.5**	62.1	61.2
FR-CDNet(WS)	**75.5**	**68.3**	**70.7**	**81.6**	72.4	73.9	**81.6**	73.9	72.4
FR-CDNet(S)	74.5	67.9	70.5	81.3	**73.0**	**74.4**	81.3	**74.4**	**73.0**

TABLE IV

ABLATION STUDY FOR CROSS FUSION ON PSCD. *We migrate the baseline models into the proposed Cross Fusion (CF) architecture and additionally report the metrics of change disentanglement (M_0, M_1), in comparison with Table III. (S) denotes the supervised strategy trained directly with disentanglement labels, while (WS) denotes the weakly supervised strategy that relies on the Combine Operator.*

Fig. 4. **Qualitative results of change disentanglement by our method.**

Model	Fusion Op	PSCD(Aligned)		
		M	M_0	M_1
C3PO(base)	MTF	81.7	-	-
C3PO+CF	MTF	82.1	58.2	62.3
C3PO+CF	Nc-MTF	83.6	**73.8**	**76.0**
Ours	MTF	81.5	70.8	72.6
Ours	Nc-MTF	81.6	**72.4**	**73.9**

TABLE V

ABLATION STUDY FOR NC-MTF. *We compare our non-commutative Nc-MTF with the original commutative MTF proposed in C-3PO [38]. +CF means migrating to Cross Fusion architecture.*

Table IV reports the performance of mainstream baseline models migrated into our proposed Cross Fusion architecture (on PSCD dataset). Since the feature fusion operator MTF in C-3PO does not satisfy the non-commutative constraint of Cross Fusion, we replace it with Nc-MTF (Sec. III-B). For methods such as DR-TANet and CSCDNet, which employ two different backbones by default, we modify them to use a shared-weight backbone. The migrated models gain the capability for change disentanglement. Therefore, their performance on the disentangled outputs (M_0, M_1 columns) is also reported. The last row shows the result of supervised disentanglement strategy trained directly with labeled masks (S) as a reference.

A horizontal comparison with Table III demonstrates that Cross Fusion architecture consistently brings performance improvements—for example, +10.4% for DR-TANet and +7.8% for C-3PO in unaligned setting. Moreover, Cross Fusion eliminates temporal-condition dependency, achieving consistent results between the Align and Reverse columns. For change disentanglement, the weakly supervised strategy (WS) utilizing the Combine Operator (Eq. 5) achieves comparable performance to supervised strategy, validating its effectiveness. Qualitative results in Fig. 4 further confirm that our method achieves correct disentanglement.

Table V demonstrates the importance of the non-commutative constraint for change disentanglement. When the commutative MTF operator is used, the disentanglement performance drops by 15.6% and 13.7%, respectively. FR-CDNet does not exhibit significant performance degradation in this context, since the feature remapping function SPCA alters the original features, producing heterogeneous outputs across the two branches if using MTF which to some extent simulates non-commutative constraint.

G. Ablation Study: SPCA

Table VI validates the transferability and effectiveness of SPCA. We embed standard Cross Attention (CA) and our proposed SPCA, respectively, as the feature remapping function into C-3PO equipped with Cross Fusion. The results show that by introducing spatial priors, SPCA significantly improves performance in unaligned scenes while preserving detection capability in aligned scenes, whereas standard Cross Attention severely damages performance in the aligned setting.

Model	VL-CMU-CD		PSCD	
	Align	Reverse	Unalign	Align
C3PO	78.2	15.2	57.4	81.7
+CF	78.7	78.7	65.2	83.6
+CF+CA	66.4	66.4	74.6	77.4
+CF+SPCA	76.8	76.8	75.6	81.5

TABLE VI

ABLATION STUDY FOR SPCA. *CA means standard Cross Attention, CF means Cross Fusion.*

V. CONCLUSION

We introduce FR-CDNet, a transferable scene change detection architecture that simultaneously handles viewpoint variations ranging from minor to significant, along with different temporal conditions, in a unified manner. Without requiring additional annotations, FR-CDNet achieves effective temporal disentanglement of change entities, significantly reducing labeling costs and lowering the barrier for semantic- and instance-level change detection. Extensive experiments demonstrate state-of-the-art performance across multiple settings, including both aligned and unaligned scenes under various temporal conditions. Moreover, ablation studies verify the strong transferability of the proposed architecture,

consistently improving mainstream models and offering a new perspective for future research.

REFERENCES

[1] G. Wu, Y. Zheng, Z. Guo, Z. Cai, X. Shi, X. Ding, Y. Huang, Y. Guo, and R. Shibasaki, "Learn to recover visible color for video surveillance in a day," in *European conference on computer vision*. Springer, 2020, pp. 495–511.

[2] U. Nehmzow, *Mobile robotics: a practical introduction*. Springer Science & Business Media, 2012.

[3] C. Choe, S. Lee, and N. Sung, "Scene change detection for robotic patrol system," in *2024 Eighth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2024, pp. 114–115.

[4] J. Janai, F. Güney, A. Behl, A. Geiger, *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and trends® in computer graphics and vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[5] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang, "Registration based few-shot anomaly detection," in *European conference on computer vision*. Springer, 2022, pp. 303–319.

[6] S. Stent, R. Gherardi, B. Stenger, and R. Cipolla, "Detecting change for multi-view, long-term surface inspection." in *BMVC*, 2015, pp. 127–1.

[7] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Autonomous Robots*, vol. 42, no. 7, pp. 1301–1322, 2018.

[8] K. Sakurada, T. Okatani, and K. Deguchi, "Detecting changes in 3d structure of a scene from multi-view images captured by a vehicle-mounted camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 137–144.

[9] H. M. Badland, S. Opit, K. Witten, R. A. Kearns, and S. Mavoa, "Can virtual streetscape audits reliably replace physical streetscape audits?" *Journal of urban health*, vol. 87, no. 6, pp. 1007–1016, 2010.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[13] K. Cho, D. Y. Kim, and E. Kim, "Zero-shot scene change detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2509–2517.

[14] C.-J. Lin, S. Garg, T.-J. Chin, and F. Dayoub, "Robust scene change detection using visual foundation models and cross-attention mechanisms," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8337–8343.

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

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

[17] S. Lee and J.-H. Kim, "Semi-supervised scene change detection by distillation from feature-metric alignment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1226–1235.

[18] J.-M. Park, U.-H. Kim, S.-H. Lee, and J.-H. Kim, "Dual task learning by leveraging both dense correspondence and mis-correspondence for robust change detection with imperfect matches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 749–13 759.

[19] Y. Jie, H. He, P. Wang, A. Yue, X. Mi, Z. Ma, K. Xing, S. Cao, Z. Cui, and C. Jiang, "Sfda-ocdet: Spatial-frequency difference attention network for remote sensing object-level change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[20] J. Ma, B. Li, H. Li, S. Meng, R. Lu, and S. Mei, "Remote sensing change detection by pyramid sequential processing with mamba," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[21] G. Wang, H. Chen, T. Qiao, J. Wang, and W. Liu, "Resolution-difference embedded network for cross-resolution remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[22] Z. Li, C. Tang, X. Hu, N. Li, S. Xiang, C. Li, C. Li, and X. Liu, "Boosting remote sensing change detection via hard region mining," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[23] Y. Wen, F. Liu, Q. Cao, and G. Niu, "Semantic-based motion detection method for unmanned aerial vehicle data transmission," in *ICC 2024-IEEE International Conference on Communications*. IEEE, 2024, pp. 1697–1702.

[24] L. Trinh, A. Anwar, and S. Mercelis, "Seadsc: A video-based unsupervised method for dynamic scene change detection in unmanned surface vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 840–847.

[25] M. Mandal, V. Dhar, A. Mishra, S. K. Vipparthi, and M. Abdel-Mottaleb, "3dcd: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos," *IEEE Transactions on Image Processing*, vol. 30, pp. 546–558, 2020.

[26] K. Sakurada, W. Wang, N. Kawaguchi, and R. Nakamura, "Dense optical flow based change detection network robust to difference of camera viewpoints," *arXiv preprint arXiv:1712.02941*, 2017.

[27] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "Changenet: A deep learning architecture for visual change detection," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[28] Y. Lei, D. Peng, P. Zhang, Q. Ke, and H. Li, "Hierarchical paired channel fusion network for street scene change detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 55–67, 2020.

[29] A. Bauer, J.-C. Krabbe, and A. Kummert, "Changesam: Adapting the segment anything model to street scene image change detection," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2024, pp. 1672–1677.

[30] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[31] T. Celik and K.-K. Ma, "Unsupervised change detection for satellite images using dual-tree complex wavelet transform," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 3, pp. 1199–1210, 2009.

[32] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 1, pp. 33–37, 2008.

[33] C. Huo, Z. Zhou, H. Lu, C. Pan, and K. Chen, "Fast object-level change detection for vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 118–122, 2009.

[34] J. Im, J. Jensen, and J. Tullis, "Object-based change detection using correlation image analysis and image segmentation," *International journal of remote sensing*, vol. 29, no. 2, pp. 399–423, 2008.

[35] T. P. Nguyen, C. C. Pham, S. V.-U. Ha, and J. W. Jeon, "Change detection by training a triplet network for motion feature extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 433–446, 2018.

[36] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puig, and Y. Ruichek, "Bscgan: Deep background subtraction with conditional generative adversarial networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4018–4022.

[37] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

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

[39] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Computer Vision and Image Understanding*, vol. 187, p. 102783, 2019.

[40] 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)*. IEEE, 2021, pp. 8578–8585.