Environmental Change Detection for Real-World Change Analysis

Kyusik Cho[1], Suhan Woo[1], Hongje Seong[2⋆], and Euntai Kim[1,3⋆]

[1] Yonsei University [2] University of Seoul
[3] Korea Institution of Science and Technology
{ks.cho, wsh112, etkim}@yonsei.ac.kr hjseong@uos.ac.kr
https://kyusik-cho.github.io/ECD

Abstract. Scene Change Detection (SCD) evaluates changes using predefined query-reference (*i.e.*, present-past) image pairs. However, this formulation overlooks a critical dependency: the corresponding query-reference pair is assumed to be prepared in advance. In real-world applications, such as mobile robots, future query views cannot be known in advance, and thus their corresponding reference images cannot be predefined. To remove this dependency and push change detection toward more practical applications, we introduce Environmental Change Detection (ECD). A key aspect of ECD is to avoid unrealistically predefined and aligned query-reference pairs and instead retrieve environmental cues from an uncurated image database of reference scenes. To tackle this new challenging task, we additionally introduce an initial solution that enables change detection under unknown and imperfect query-reference conditions. The main idea of our solution is to retrieve multiple reference candidates and aggregate semantically rich representations for change detection. We further construct ECD benchmark sets by reformulating three standard change detection datasets. Extensive experimental results demonstrate the efficacy of our solution in both ECD and SCD.

Keywords: Change detection · Semantic segmentation · Scene analysis

1 Introduction

Change detection in images is a fundamental problem in computer vision. Traditionally, this problem has been addressed through Scene Change Detection (SCD), which aims to identify differences between a reference image and a query image captured at distinct times. SCD has attracted increasing attention due to its potential in a wide range of practical applications, such as urban scene analysis [2,4,27], disaster damage assessment [1,26], anomaly detection [13], and warehouse management [20,21].

Despite its significance, SCD has limited applicability in real-world scenarios. This limitation arises primarily from two idealized assumptions: (1) an

⋆ Corresponding authors.

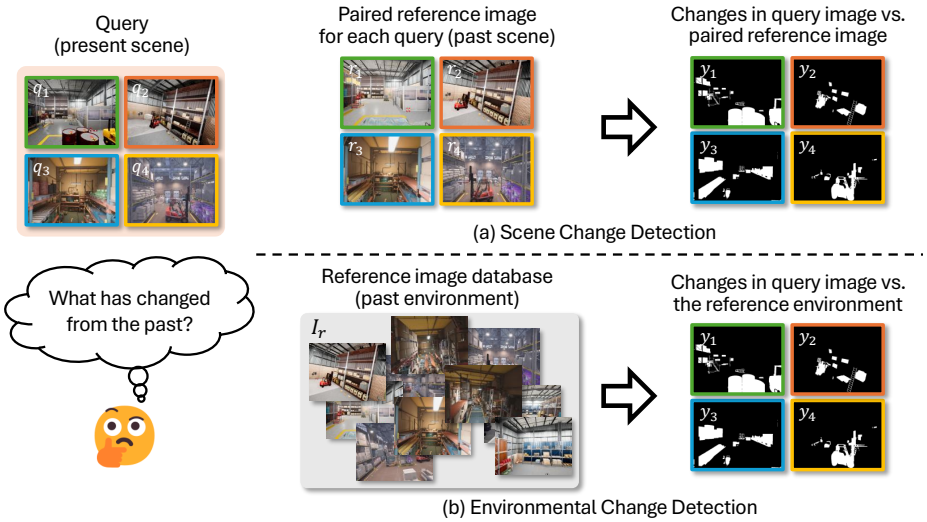

Fig. 1: Comparison between the problem settings of SCD and ECD. In the conventional SCD setting, each query image is provided with its corresponding reference image as a predefined pair. In contrast, ECD considers a more practical scenario where query-reference pairs are unknown. ECD just provides a single query image, where the database of reference images is accessible. Furthermore, ECD does not guarantee the presence of a perfectly matched pair within the database.

appropriate reference image for every query is predefined before the query is presented, and (2) each query-reference pair is captured from the same viewpoint, ensuring spatial alignment. Several recent studies, such as SimSac [21] and RSCD [14], have attempted to relax the second assumption by handling misalignments through optical flow or cross-attention. However, these approaches still rely on the predefined pairing assumption, which imposes an inherent limitation: unobservable regions caused by viewpoint mismatch impose an upper bound on achievable performance. Consequently, such extensions toward mismatch handling do not fully resolve the practicality issues of the conventional SCD formulation.

To address these limitations, we introduce Environmental Change Detection (ECD). ECD is inspired by how changes are detected in real-world scenarios. To detect changes from the past to the present, an appropriate past image does not emerge simultaneously with the query; rather, only a database of past images exists. In such cases, humans first search for relevant images from the past image database. In reality, perfectly aligned past images are exceedingly rare, and the available images often have only a limited field-of-view overlap. Under these challenging conditions, humans fill in missing information from multiple observations rather than abandoning inference for unobservable regions.

ECD is designed to reflect these common practices. In ECD, the reference information is indirectly provided by a large-scale reference database rather than

predefined pairs, and a perfectly aligned image may not exist within it. This formulation simultaneously relaxes both assumptions inherent in SCD, thereby making the task more challenging while improving its relevance to real-world applications. The comparison between SCD and ECD is illustrated in Fig. 1.

Conceptually, ECD can be viewed as a large-scale multi-reference change detection problem with many query-irrelevant references acting as distractors. To address this setting, we propose a framework that identifies and integrates query-relevant reference evidence for change detection. Directly processing the entire reference database is computationally prohibitive and exposes the detector to numerous irrelevant references. Instead, our method first uses VPR to construct a compact, query-specific subset of informative reference candidates. Then, we introduce the Environment Mosaic Composer (EMC), which reconstructs a query-oriented reference environment representation by selectively composing the most relevant feature patches for each region of the query in a mosaic fashion. After applying EMC across multiple scales, we aggregate the resulting representations to further enhance robustness. A change segmentation head is subsequently applied to produce the final change map.

Furthermore, we reformulate three standard change detection datasets and construct ECD benchmark sets. Experimental results highlight the effectiveness of the proposed approach and provide a step toward practical applications of change detection. Our contributions are summarized as follows:

- We identify practical limitations of the existing change detection approaches and introduce a new challenging task named ECD.
- We propose the first solution to ECD, which predicts changes by retrieving and integrating multiple reference candidates, inspired by human reasoning.
- We establish standard experimental settings for ECD, providing benchmark datasets and an initial baseline to guide further studies.

2 Related Work

Scene change detection. Recent research on scene change detection (SCD) has leveraged deep learning to develop a variety of approaches [12, 24, 25]. DR-TANet [6] introduced the temporal attention module into the encoder-decoder architecture. C-3PO [31] proposes a model that utilizes temporal features to distinguish three types of changes. EMPLACE [4] performs self-supervised learning using a time-interval-based triplet loss. Sachdeva and Zisserman [25] and Lee and Kim [12] utilized synthetic change data. ZSSCD [7] leverages a foundation tracking model to perform change detection without training. Meanwhile, several methods are designed to be robust against misalignment between the reference and query images. SimSac [21] incorporates optical flow estimation and a warping module to account for misalignment. RSCD [14] introduces a cross-attention-based change detection head, enabling comparison even when object positions differ. However, these methods still rely on the predefined query-reference pairs. When misalignment is introduced under such predefined pairing, unobservable

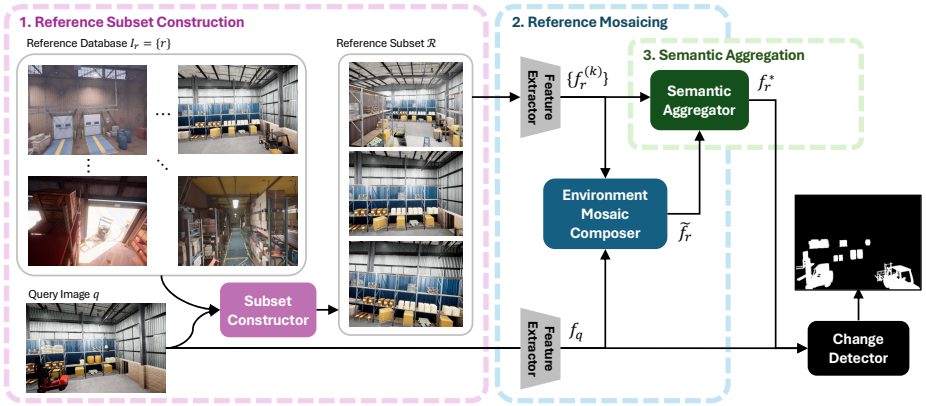

Fig. 2: Overview. Our approach first constructs a query-specific subset from the reference database. Then, EMC generates a mosaic reference environment representation that matches the perspective of the query image. Based on the mosaic representation, the semantic aggregator integrates the semantic information across the subset. Finally, both the reference features and the query features are fed into the change detector to produce the final prediction.

regions inevitably arise, making the formulation impractical. Hence, it is essential to resolve both problems simultaneously. To jointly overcome these limitations, we establish a novel setting that removes the assumptions of predefined pairing and spatial alignment.

Visual place recognition. Visual place recognition (VPR) is the task of estimating the location of a query image by retrieving visually similar images from a pre-collected, geo-tagged database. With the rapid progress of deep learning, numerous effective methods have been proposed to tackle this challenge. NetVLAD [5] is one of the first works to successfully apply deep learning to VPR, outperforming traditional SIFT-based methods [16] and becoming a common baseline in later studies [17,18,22,32,33]. DINOv2 SALAD [9] reformulates local feature aggregation in VPR as an optimal transport problem using the Sinkhorn algorithm, and employs DINOv2 [19] as the backbone, achieving competitive performance with reduced training cost. BoQ [3] trains learnable global queries that aggregate local features via cross-attention to enable accurate and fast retrieval. Meanwhile, several SCD studies have been associated with VPR. For example, SimSac [21] utilized VPR to prepare imperfect match pairs for SCD. ZeroSCD [10] leveraged a VPR backbone for feature extraction in SCD, inspired by the observation that VPR models are inherently robust to variations in style and content. Unlike previous approaches that use VPR to prepare image pairs [21] or extract useful representations [10], we use VPR as the first stage of our solution to construct a query-specific reference subset from a large-scale reference database.

3 Environmental Change Detection

3.1 Background: Scene Change Detection (SCD)

Conventional SCD datasets [2, 20, 27] consist of (r, q, y) triplets, where r and q are the pre-change (*i.e.*, reference) and post-change images (*i.e.*, query) taken at time t_0 and t_1, respectively, and y denotes the corresponding ground-truth change map. The objective is to predict the pixel-wise change mask $\hat{y}$ for the given pair (r, q), formulated as $\hat{y} = F(r, q)$, where F denotes the SCD system. This formulation is based on two idealized assumptions: (1) every query image q is paired with a corresponding reference image r, and (2) the image pair (r, q) is spatially aligned. Although these assumptions substantially simplify the change detection problem, they limit its broader applicability in practical real-world scenarios.

3.2 Problem Formulation

To increase its applicability in practical environments, we introduce Environmental Change Detection (ECD). Fig. 1 contrasts SCD with ECD and highlights their key differences. ECD relaxes the two unrealistic assumptions of SCD as follows: (1) ECD replaces the predefined paired input (r, q) with a query-database input $(\mathcal{I}_r, q)$, where q denotes the query image and $\mathcal{I}_r = \{r^{(i)}\}_{i=1}^{N_r}$ denotes a large-scale reference image database. Importantly, $\mathcal{I}_r$ is not constructed specifically for each query or sequence. Instead, the same reference database is shared across all queries and sequences and may therefore contain images captured in environments entirely different from that of the current query. The system must explore $\mathcal{I}_r$ and collect useful reference information relevant to the given q, reflecting practical usage scenarios. (2) A reference image r that is perfectly aligned with q is not guaranteed to exist in $\mathcal{I}_r$. To emulate this condition, we define $\mathcal{I}_r$ as a uniformly sampled subset of the entire collection of reference images. The goal of ECD is to predict a pixel-wise change mask $\hat{y}$ for a given q by leveraging the information within $\mathcal{I}_r$, formulated as $\hat{y} = G(\mathcal{I}_r, q)$, where G denotes the ECD system.

4 Method

The overall architecture of our ECD framework is illustrated in Fig. 2. To explore $\mathcal{I}_r$, we first employ a VPR-based retrieval scheme to construct a compact subset of relevant reference images. Next, we apply the Environment Mosaic Composer (EMC) to the constructed subset. EMC reconstructs a mosaic representation of the environment by selectively composing feature patches from multiple references to match the spatial layout of the query q. Subsequently, semantic features from the subset are aggregated based on this mosaic representation to further enrich the reference understanding. Finally, change detection is performed using the enriched reference representation and the query.

4.1 Reference Subset Construction

Instead of directly utilizing the entire reference database $\mathcal{I}_r$, we adopt a selective strategy that focuses on the most relevant subset for each query q. Our subset constructor consists of two components: a VPR module and a subset size parameter K. The VPR module computes place-level similarity between the query image q and all reference images in $\mathcal{I}_r$, and the top-K images form the reference subset $\mathcal{R}$:

$$\mathcal{R} = \{r^{(1)}, r^{(2)}, \ldots, r^{(K)}\} = \mathrm{TopK}_{\mathrm{VPR}}(q, \mathcal{I}_r). \tag{1}$$

This subset serves as a compact and relevant representation of $\mathcal{I}_r$ for estimating changes with respect to q.

According to the ECD formulation, however, each retrieved image $r^{(k)} \in \mathcal{R}$ is highly likely to be captured from a viewpoint different from that of q. Consequently, simply relying on a single top-ranked reference often results in suboptimal performance due to viewpoint misalignment and limited scene coverage. While conventional SCD approaches attempt to alleviate the mismatch between paired images, they inherently overlook the fundamental limitation that a single reference image can only represent a narrow portion of the environment. To overcome this, we introduce a new paradigm that composes the regions most relevant to the query from multiple viewpoints, thereby enabling a query-oriented understanding of the scene beyond the constraints of any single reference.

4.2 Environment Mosaic Composition

We introduce a new module, termed Environment Mosaic Composer (EMC), which reconstructs a query-oriented reference environment representation inspired by the way humans mentally reconstruct spatial layouts from fragmented observations. When no single image fully captures the query viewpoint, humans imagine the missing parts of a scene. They intuitively assemble the most plausible regions from multiple viewpoints to complete the query view, forming a coherent mental picture. Analogously, EMC divides the query feature map into local patches and selects the most relevant reference feature patch for each patch. Without performing any explicit pixel-level reconstruction, this compositional process yields a mosaic environment representation $\tilde{f}_r$. The resulting representation provides sufficient spatial correspondence cues for reliable change detection.

To realize this composition process, all query and reference images are first encoded using a frozen DINO backbone [19] to produce feature maps f_q and $\{f_r^{(k)}\}_{k=1}^{K}$. These features are then ℓ_2-normalized along the feature dimension. Next, the query feature f_q is divided into an $n \times n$ grid, forming the set of grid locations $\mathcal{P}_n$. For each grid cell $p \in \mathcal{P}_n$, we obtain a local patch $f_q[p] \in \mathbb{R}^{d \times h \times w}$, where d, h, and w denote channel, height, and width dimensions, respectively. Each query patch is compared with all possible patches from the reference features:

$$S_k^p = f_r^{(k)} \circledast f_q[p], \tag{2}$$

$$(k^*, z^*) = \arg\max_{k,z} S_k^p[z]\ , \tag{3}$$

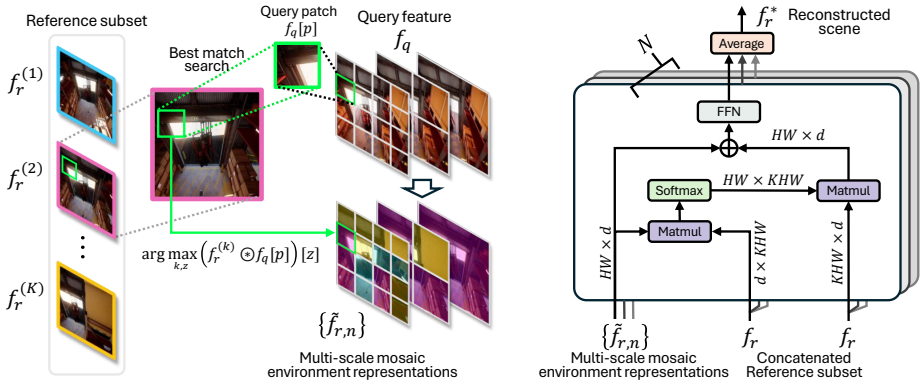

Fig. 3: Illustration of environment mosaic composer and semantic aggregator. (a) The input query feature f_q is divided into an $n \times n$ grid. EMC searches over all reference features $\{f_r^{(k)}\}_{k=1}^{K}$ and identifies the best-matching patch through a convolution-based similarity computation. By compositing the best matches across all grid cells, EMC constructs a mosaic environment representation $\tilde{f}_{r,n}$ that reflects the viewpoint of the query. This process is performed hierarchically with multiple N scales to capture both coarse and fine-grained correspondences. (b) Features from the reference subset are integrated via cross-attention, using the mosaic environment representation $\tilde{f}_{r,n}$ as the query. Semantic aggregation is performed at each grid resolution, and the resulting features are averaged to produce the final reconstructed scene f_r^*. For clarity, the illustration of the multi-head attention (MHA) is omitted.

where $\circledast$ denotes 2D convolution over the spatial dimensions, computing a sliding-window sum of local cosine similarities. The index z ranges over all valid stride-1 locations in $f_r^{(k)}$ at which a reference patch with the same spatial dimensions as $f_q[p]$ can be extracted. The patch with the highest similarity $f_r^{(k^*)}[z^*]$ is then copied from the reference map and placed into the mosaic, composing a query-oriented reference environment representation:

$$\tilde{f}_{r,n}[p] = f_r^{(k^*)}[z^*] \, . \tag{4}$$

This process is repeated for all $p \in \mathcal{P}_n$ and across multiple grid resolutions n. Consequently, multi-scale mosaic environment representations $\{\tilde{f}_{r,n}\}$, which reflect the spatial layout of f_q, are obtained. An illustration of EMC is provided in Fig. 3a.

4.3 Semantic Aggregation and Change Segmentation

Each mosaic environment representation $\tilde{f}_{r,n}$ provides strong spatial alignment with the query feature f_q, but the compositional process may introduce spatial discontinuities at patch boundaries, which can lead to minor semantic inconsistencies across adjacent regions. To complement such inconsistencies and enrich the contextual understanding of the reconstructed scene, we incorporate a

semantic aggregator that refines information from multiple references through cross-attention. The strong spatial correspondence of $\tilde{f}_{r,n}$ makes it particularly suitable to serve as the query in the cross-attention mechanism. Meanwhile, the reference features $\{f_r^{(k)}\}_{k=1}^K$ offer complementary contextual cues and are thus employed as the keys and values. We implement this process using a multi-head attention [30], followed by dropout, a residual connection, and a two-layer feed-forward network (FFN) with ReLU activation:

$$f_{r,n}^* = \mathrm{FFN}(\mathrm{Dropout}(\mathrm{MHA}(\tilde{f}_{r,n}, f_r, f_r)) + \tilde{f}_{r,n}) \ , \tag{5}$$

where $f_r = \mathrm{concat}(f_r^{(1)}, \ldots, f_r^{(K)})$.

The resulting representation $f_{r,n}^*$ is semantically enriched and spatially aligned with f_q at grid resolution n. Finally, we generate the reconstructed scene f_r^* by averaging the representations $f_{r,n}^*$. An illustration of the semantic aggregator is provided in Fig. 3b.

To predict the change, we feed f_r^* and f_q into a change detection module. Following RSCD [14], we adopt a change detection module consisting of a shallow cross-attention block followed by convolutional layers for pixel-level change prediction. The architectural details are provided in the supplementary material.

5 Experiments

5.1 Datasets

We select three SCD datasets, VL-CMU-CD [2], PSCD [27], and ChangeSim [20], and restructure them for ECD. The key modification in our restructuring is the removal of the conventional query-reference pairing assumption. Specifically, the dataset is no longer composed of paired tuples (r, q) but instead two independent sets: a set of reference images $\mathcal{I}_r$ and a set of query images $\mathcal{I}_q$. To further eliminate the unrealistic assumption that a spatially aligned reference image r always exists in $\mathcal{I}_r$ for every $q \in \mathcal{I}_q$, we define $\mathcal{I}_r$ as a subset sampled from the entire collection of pre-change (reference) images of the original dataset. The sampling is performed by applying a database stride s, resulting in $\mathcal{I}_r$ being downsampled to approximately $1/s$ of its original size. This increases both the practicality and difficulty of the benchmark, approximating realistic scenarios in which past environments are sparsely captured with camera-mounted vehicles. We generally write the restructured dataset as VL-CMU-CD^s, PSCD^s, and ChangeSim^s, and when a specific value is used, we indicate it explicitly (*e.g.*, ChangeSim^5 for $s = 5$). In our experiments, due to the maximum sequence lengths of VL-CMU-CD and PSCD (20 and 15, respectively), we consider s up to 10; however, if experiments are conducted on datasets with longer sequences, this setting can be naturally extended beyond 10.

Tab. 1 summarizes the statistics of the test sets in the restructured ECD datasets under different database strides. To characterize the geometric difficulty induced by reference sparsification, we report two overlap measures: reference-reference (r-r) overlap and query-reference (q-r) overlap. We define overlap as

Table 1: Statistics of the restructured ECD datasets under different database strides.

	VL-CMU-CD^s						PSCD^s						ChangeSim^s					
	Database size			Overlap (%)		avg.	Database size			Overlap(%)		avg.	Database size			Overlap(%)		avg.
s	$\lvert\mathcal{I}_q\rvert$	$\lvert\mathcal{I}_r^s\rvert$	$\lvert\mathcal{I}_r^s\rvert/\lvert\mathcal{I}_q\rvert$	r-r	q-r	seq. len.	$\lvert\mathcal{I}_q\rvert$	$\lvert\mathcal{I}_r^s\rvert$	$\lvert\mathcal{I}_r^s\rvert/\lvert\mathcal{I}_q\rvert$	r-r	q-r	seq. len.	$\lvert\mathcal{I}_q\rvert$	$\lvert\mathcal{I}_r^s\rvert$	$\lvert\mathcal{I}_r^s\rvert/\lvert\mathcal{I}_q\rvert$	r-r	q-r	seq. len.
1	429	429	1.00	97.28	96.90	7.94	1155	1155	1.00	74.55	100.00	15	8212	5910	0.72	98.51	96.11	738.75
3	429	162	0.38	96.45	95.99	3.00	1155	385	0.33	23.66	81.34	5	8212	1973	0.24	95.72	95.13	246.63
5	429	106	0.25	94.65	94.54	1.96	1155	231	0.20	0.00	62.80	3	8212	1185	0.14	93.37	94.54	148.13
10	429	66	0.15	92.08	93.86	1.22	1155	154	0.13	0.00	42.80	2	8212	596	0.07	88.47	94.44	74.50

the percentage of one image that is geometrically visible in another image. The r-r overlap measures the average geometric overlap between reference images in the downsampled reference database, while the q-r overlap measures the overlap between each query and its top-1 reference. For PSCD, the overlap is computed directly from the panoramic crop geometry. For VL-CMU-CD and ChangeSim, it is estimated using a RANSAC-based homography [8] fitted to LoFTR matches [28]. As the stride increases, both the reference database size and the overlap generally decrease, indicating that a single reference image provides increasingly limited geometric coverage.

VL-CMU-CD^s is a restructured dataset derived from VL-CMU-CD [2], which captures changes in urban street scenes. The dataset primarily represents changes such as the disappearance of buildings, roads, and vehicles caused by long temporal intervals. Following the previous work [14], we resize the images to 504×504, and split the training dataset into 3,324 queries for training, 408 for validation.

PSCD^s is a restructured dataset of PSCD [27], which consists of panoramic urban scene images. Following the previous work, we crop each panoramic image into 15 sub-images with a size of 224×224 and consider them as a sequence. PSCD^s is used only for testing purposes, and evaluations are performed using the best model obtained from training and validation on VL-CMU-CD^s.

ChangeSim^s is a restructured dataset derived from ChangeSim [20], which simulates an industrial warehouse environment. The dataset includes changes such as the appearance, disappearance, and rotation of various objects. It contains 13,225 coarsely aligned image pairs for training and 8,212 for testing. We split the original training set into 8,754 training and 4,471 validation samples. Each image is resized to 224×224. In ChangeSim, each reference image is sampled from long sequences. As a result, multiple query images may share the same reference image, and some reference images may not be used at all. Therefore, rather than collecting reference images from the paired structure, we directly utilize the full t_0 sequences as a pre-change database before applying s.

5.2 Experimental Setup

Training details. Following previous work [14], we use the Adam optimizer [11] with a learning rate of 1×10^{-4} and apply a cosine learning rate decay schedule [15]. The model is trained for 100 epochs with a batch size of 4, including 10 warm-up epochs, and using the weighted cross-entropy loss. The VPR model [3] and DINO feature extractor [19] are kept frozen throughout the pipeline. We set

Table 2: Experimental results on the ECD benchmark. Our method consistently outperforms the state-of-the-art change detector across all datasets and database stride s settings. For a fair comparison, we feed reference images retrieved from the VPR model [3] to RSCD.

s	Method	Dataset			Average
		VL-CMU-CD^s	PSCD^s	ChangeSim^s	
1	RSCD [14]	0.619	0.300	0.369	0.429
	Ours	**0.704**	**0.331**	**0.392**	**0.476**
3	RSCD [14]	0.548	0.240	0.365	0.385
	Ours	**0.622**	**0.280**	**0.396**	**0.433**
5	RSCD [14]	0.468	0.203	0.360	0.344
	Ours	**0.546**	**0.225**	**0.388**	**0.386**
10	RSCD [14]	0.395	0.224	0.354	0.324
	Ours	**0.450**	**0.245**	**0.388**	**0.361**

Table 3: Experimental result under the SCD setup. Results obtained under the standard SCD setting of RSCD [14]. Inference time (FPS) is measured on the VL-CMU-CD dataset using a single NVIDIA RTX A5000 GPU. * are taken from the RSCD paper and are scaled based on the FPS of RSCD for a fair comparison.

Method	Dataset			Average	Inference time (ms)
	VL-CMU-CD	PSCD	ChangeSim		
DR-TANet [6]	0.607	0.023	-	-	31.84*
C-3PO [31]	**0.795**	0.048	-	-	23.54*
RSCD [14]	**0.795**	0.337	0.344	0.492	31.14
Ours	0.791	**0.347**	**0.372**	**0.503**	36.21

the reference subset size to $K = 3$ by default, and define the EMC multi-scale grid resolution as $n \in \{1, 2, 4\}$.

Evaluation metrics. Following previous work, we employ the F1-score as the evaluation metric for all datasets.

5.3 Experimental Results

The experimental results are shown in Tab. 2. As shown in the table, our method consistently outperforms the state-of-the-art change detector, RSCD [14], across all datasets and database strides. Additional implementation details are provided in the supplementary material. In the most favorable case, database stride $s = 1$, our method achieves an average performance of 0.476, significantly surpassing the RSCD [14] of 0.429. As the reference database becomes sparser (*i.e.*, as the database stride increases), the overall performance of both RSCD [14] and our method degrades. However, the performance gap between the two remains

Table 4: Ablation study. We evaluate the impact of two key components in our method, the semantic aggregator and EMC. Results across three datasets show that combining both components yields the best results.

s	Method		Dataset			Average
	Agg.	EMC	VL-CMU-CD^s	PSCD^s	ChangeSim^s	
1			0.6107	0.2999	0.3688	0.4265
	✓		0.6112	0.3281	0.3621	0.4338
		✓	0.6990	0.2029	**0.4036**	0.4352
	✓	✓	**0.7043**	**0.3308**	0.3924	**0.4758**
5			0.4680	0.2027	0.3601	0.3436
	✓		0.4537	0.2023	0.3712	0.3424
		✓	0.5304	0.1285	**0.3952**	0.3514
	✓	✓	**0.5456**	**0.2246**	0.3879	**0.3860**

consistent. These results validate that our feature extraction and aggregation process enables effective change detection even under significant viewpoint misalignments.

In addition, we also report the results under the conventional SCD setting in Tab. 3. For evaluation, we follow the experimental protocol of RSCD [14], which represents the current state of the art in SCD. To adapt our ECD pipeline to this setting, we set the database stride to $s = 1$ and replace the reference subset with the ground-truth reference pair. Even under this idealized condition, our method achieves the best average performance while remaining competitive with existing SCD approaches, demonstrating its strong capability for robust change detection.

5.4 Ablation Study and Analysis Experiments

We first conduct a component-wise ablation study to quantify the contribution of the main components of our method. We then present a focused analysis of EMC from three complementary perspectives. First, we compare EMC with alternative misalignment mitigation strategies based on single-reference warping. Second, we compare EMC with naive multi-view fusion strategies to examine whether the gain comes simply from using more references. Finally, we analyze the sensitivity of EMC to the quality of the first-stage VPR retrieval by replacing the retrieved references with oracle or distractor references. Additional analyses of the remaining design components are provided in the supplementary material.

Ablation study. In Tab. 4, we present a component-wise ablation study to evaluate the impact of each component of our proposed method. When the environment mosaic composer is disabled, feature aggregation is performed using the features from the top-1 VPR result. Conversely, when feature aggregation is disabled, the averaged mosaic environment representation $\tilde{f}_r$ is directly fed into

Table 5: Analysis experiments.

(a) Alternative misalignment mitigation modules.

VL-CMU-CD^s	s=1	s=5
w/o align	0.619	0.468
Homography	0.641	0.512
Optical flow	0.604	0.467
EMC (Ours)	**0.704**	**0.546**

(b) Alternative multi-view fusion strategies.

VL-CMU-CD^s	s=1	s=5
Best single	0.619	0.468
Average	0.586	0.382
Concatenate	0.614	0.457
EMC (Ours)	**0.704**	**0.546**

(c) Retrieval sensitivity analysis. *O: Oracle, D: distractor*

VL-CMU-CD^1	F1	Patch acc.
1 O + 2 VPR	**0.716**	**97.71**
3 VPR	0.704	97.57
2 VPR + 1 D	0.695	96.93
1 VPR + 2 D	0.663	95.58

the change detection module. If both components are removed, features from the top-1 VPR result are used for change detection. The results show that combining both aggregation and alignment yields the highest average scores, demonstrating their complementary nature.

Analysis of misalignment mitigation. Our method employs a mosaic-based design to address misalignment. Beyond this choice, two straightforward alternatives can be considered: optical-flow-based warping and homography-based warping. To analyze the effectiveness of EMC, we construct two variants by replacing EMC with (i) RAFT-based optical-flow [29] and (ii) a RANSAC-based homography [8,28], and report the results in Tab. 5a. This comparison examines whether conventional single-reference geometric warping can provide benefits comparable to those of the proposed mosaic-based composition. The results show that homography-based warping has only a marginal effect, whereas optical-flow-based warping degrades performance. In contrast, EMC achieves the strongest performance among the compared misalignment mitigation strategies.

Analysis of multi-view fusion. We further compare EMC with alternative multi-view fusion strategies by replacing EMC with feature averaging and feature concatenation. The results in Tab. 5b show that merely increasing the number of reference images does not necessarily improve performance. Naively mixing features can be harmful, especially at larger strides, where the retrieved subset is more likely to contain inaccurate references. In contrast, EMC improves performance by spatially selecting reliable reference evidence rather than simply aggregating more reference images.

Analysis of retrieval sensitivity. Finally, we analyze the sensitivity of EMC to the quality of the first-stage VPR retrieval. To this end, we replace the top VPR results with oracle and distractor references and report the results in Tab. 5c. As expected, oracle references improve performance, while distractor references degrade performance by reducing valid pre-change evidence. For further analysis, we also report patch accuracy, which measures whether EMC selects patches from informative same-sequence references. Even under distractor injection, EMC maintains high patch accuracy, indicating that it largely avoids distractor patches and continues to exploit the remaining relevant references.

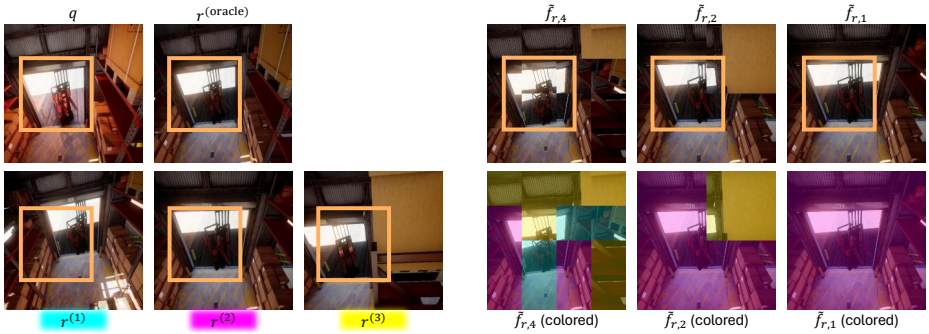

Fig. 4: Visualizations of constructing a mosaic $\tilde{f}_{r,n}$ from three reference images $r^{(1)}, r^{(2)}, r^{(3)}$. Appropriate patches are selected from each reference to form a mosaic-view similar to the ground-truth reference $r^{(\text{oracle})}$, particularly at the finer grid resolution ($n = 4$). For improved visibility, we also provide a colored version in which the color indicates the origin of each patch from the reference images.

5.5 Visualization of EMC

We present visual examples of the mosaic environment representations $\{\tilde{f}_{r,n}\}$ in Fig. 4. EMC operates on the features of the query q and retrieved reference subset $\{r^{(1)}, r^{(2)}, r^{(3)}\}$, while $r^{(\text{oracle})}$ is not included in the reference database but is shown for illustrative purposes. Although $\{\tilde{f}_{r,n}\}$ are feature representations, they are visualized as images by projecting the matched feature patches onto their spatial coordinates. The visualization confirms that EMC can successfully imitate query views by selecting high-similarity patches from references for each query patch. Additional examples and detailed analyses are provided in the supplementary material.

5.6 Qualitative Results

We present qualitative results in Fig. 5 under a challenging evaluation setting based on VL-CMU-CD^s. To construct a more practically challenging scenario beyond the standard benchmark, we retain only a single reference image from each test sequence ($s = \text{inf}$) and further augment the resulting reference database with the $\mathcal{R}$1M distractor set [23]. The $\mathcal{R}$1M distractor set comprises 1,001,001 distractor images for VPR, thereby enabling us to simulate ECD with a large-scale reference database containing diverse and unrelated images.

For the given query q, the retrieved reference subset includes both true reference and distractor images. Leveraging the environment mosaic composer and semantic aggregator modules, our method selectively extracts relevant information from this subset, enabling accurate predictions. Intermediate visualizations illustrating how these modules suppress incorrect VPR matches and compensate for camera viewpoint variations are provided in the supplementary material, along with additional qualitative examples on the standard benchmark datasets.

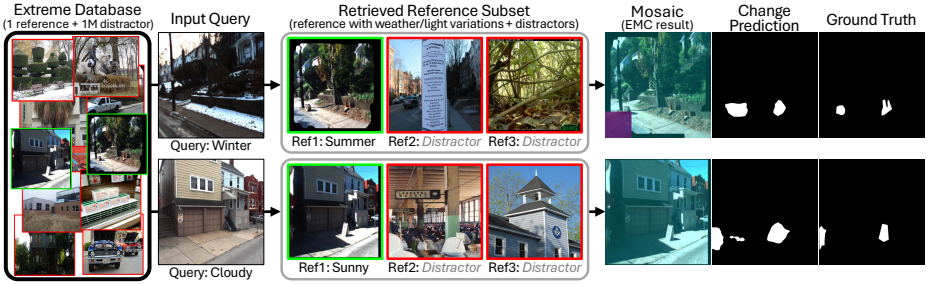

Fig. 5: Qualitative results in a challenging setting with a large-scale reference database. To simulate a less constrained real-world scenario, we restrict the reference database to one reference image per sequence and augment it with $\mathcal{R}$1M distractors. For each query, the retrieved reference subset contains both true references and distractors, yet our method selectively extracts discriminative cues to achieve accurate predictions.

6 Limitations

Although ECD eliminates the need for predefined query-reference image pairs, it still requires sufficient field-of-view overlap between the query images and the images in the reference database to enable reliable change detection. Additionally, this setting may increase the computational burden because relevant reference evidence must be discovered at inference time, whereas conventional SCD directly provides a predefined reference image for each query. Representative failure cases are provided in the supplementary material.

7 Conclusion

In this paper, we introduce a new task termed Environmental Change Detection (ECD), which removes two unrealistic assumptions commonly made in conventional Scene Change Detection (SCD): the existence of predefined query-reference pairs and the availability of perfectly aligned reference images. ECD establishes a more challenging yet practical setting that better reflects real-world conditions. We also propose the first solution to ECD. Rather than processing the entire reference database, our method first constructs a compact reference subset by retrieving images relevant to the given query. It then performs mosaic composition and semantic aggregation to reconstruct a reference scene corresponding to the query viewpoint. The resulting representation preserves rich environmental information while remaining spatially aligned with the query, thereby providing a reliable basis for change detection. Our method performs effectively in ECD and outperforms the evaluated SCD baselines. Further experimental analyses demonstrate that our approach handles spatial misalignment more effectively than the alternative strategies. We believe this work takes a significant step toward practical change detection under less constrained and more realistic conditions.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-01025, Development of core technology for mobile manipulator for 5G edge-based transportation and manipulation), and Korea Evaluation Institute Of Industrial Technology (KEIT) grant funded by the Korea government(MOTIE) (No.20023455, Development of Cooperate Mapping, Environment Recognition and Autonomous Driving Technology for Multi Mobile Robots Operating in Large-scale Indoor Workspace).

References

1. Ahn, K., Han, S., Park, S., Kim, J., Park, S., Cha, M.: Generalizable disaster damage assessment via change detection with vision foundation model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 27784–27792 (2025)
2. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Autonomous Robots **42**, 1301–1322 (2018)
3. Ali-Bey, A., Chaib-draa, B., Giguère, P.: Boq: A place is worth a bag of learnable queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17794–17803 (2024)
4. Alpherts, T., Ghebreab, S., van Noord, N.: Emplace: Self-supervised urban scene change detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1737–1745 (2025)
5. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5297–5307 (2016)
6. Chen, S., Yang, K., Stiefelhagen, R.: Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. In: IEEE Intelligent Vehicles Symposium (IV). pp. 502–509. IEEE (2021)
7. Cho, K., Kim, D.Y., Kim, E.: Zero-shot scene change detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2509–2517 (2025)
8. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
9. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17658–17668 (2024)
10. Kannan, S.S., Min, B.C.: Zeroscd: Zero-shot street scene change detection. arXiv preprint arXiv:2409.15255 (2024)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference for Learning Representations (2015)
12. Lee, S., Kim, J.H.: Semi-supervised scene change detection by distillation from feature-metric alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1226–1235 (2024)

13. Li, D., Chen, L., Xu, C.Z., Kong, H.: Umad: University of macau anomaly detection benchmark dataset. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5836–5843. IEEE (2024)
14. Lin, C.J., Garg, S., Chin, T.J., Dayoub, F.: Robust scene change detection using visual foundation models and cross-attention mechanisms. In: IEEE International Conference on Robotics and Automation (2025)
15. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: International Conference for Learning Representations (2017)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
17. Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., Yuan, C.: Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16772–16782 (2024)
18. Lu, F., Zhang, X., Ye, C., Dong, S., Zhang, L., Lan, X., Yuan, C.: Supervlad: Compact and robust image descriptors for visual place recognition. Advances in Neural Information Processing Systems **37**, 5789–5816 (2024)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Park, J.M., Jang, J.H., Yoo, S.M., Lee, S.K., Kim, U.H., Kim, J.H.: Changesim: Towards end-to-end online scene change detection in industrial indoor environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8578–8585. IEEE (2021)
21. Park, J.M., Kim, U.H., Lee, S.H., Kim, J.H.: 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 (CVPR). pp. 13749–13759 (2022)
22. Qiu, Q., Zhang, S., Gao, H., Yang, H., Ying, H., Wang, W., He, X.: Emvp: Embracing visual foundation model for visual place recognition with centroid-free probing. Advances in Neural Information Processing Systems **37**, 120928–120950 (2024)
23. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5706–5715 (2018)
24. Ramkumar, V.R.T., Bhat, P., Arani, E., Zonooz, B.: Self-supervised pretraining for scene change detection. In: Proceedings of the Conference on Neural Information Processing Systems. pp. 6–14 (2021)
25. Sachdeva, R., Zisserman, A.: The change you want to see. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3993–4002 (2023)
26. Sakurada, K., Okatani, T.: Change detection from a street image pair using cnn features and superpixel segmentation. In: British Machine Vision Conference (BMVC) (2015)
27. Sakurada, K., Shibuya, M., Wang, W.: Weakly supervised silhouette-based semantic scene change detection. In: 2020 IEEE International conference on robotics and automation (ICRA). pp. 6861–6867. IEEE (2020)
28. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8922–8931 (2021)

29. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wang, G.H., Gao, B.B., Wang, C.: How to reduce change detection to semantic segmentation. Pattern Recognition **138**, 109384 (2023)
32. Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13648–13657 (2022)
33. Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., Wang, H.: R2former: Unified retrieval and reranking transformer for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19370–19380 (2023)