A Implementation Details

A.1 ECD Setup Construction

We describe the detailed procedure for preparing the experimental setup used in ECD. Starting from the standard SCD datasets, which consist of triplets (r, q, y), we first collect all reference images r and build the database. The query images q and their corresponding labels y are used without modification from the SCD setup. To remove the assumption that a spatially aligned reference image r always exists for every query q, we restrict the size of the database by applying a database stride s. Specifically, only every s-th image in the sequence from the original database is included in the reference database $\mathcal{I}_r$.

Given that several SCD datasets (*e.g.*, VL-CMU-CD [1] and ChangeSim [6]) are composed of multiple sequences, we apply the stride on a per-sequence basis. Thus, only every s-th image within each sequence is included in the reference database $\mathcal{I}_r$. For PSCD [9], where each panorama image is divided into 15 crops, we treat the set of cropped images from a single panorama as a sequence and apply the stride accordingly.

Formally, $\mathcal{I}_r$ is defined as:

$$\mathcal{I}_r = \left\{ r^{(m,j)} \;\middle|\; j \in \{1, 1+s, 1+2s, \dots\},\ j \le L_m \ \text{ for each sequence } m \right\}, \quad (1)$$

where L_m denotes the length of sequence m. This striding constraint reduces the spatial and temporal coverage of the reference database and removes the guarantee that a ground-truth reference-query pair exists.

The examples of the original and strided reference database are illustrated in Fig. 1. As shown in the figure, increasing the stride sparsifies the reference set, reduces environmental coverage, and makes retrieval inherently more challenging. This construction reflects real-world situations in which past environments are only sparsely recorded, such as camera-mounted vehicles that capture scenes intermittently. The stride s also acts as a controllable difficulty parameter, enabling consistent reproduction of experimental conditions and supporting systematic expansion of the benchmark (*e.g.*, ChangeSim^1, ChangeSim^5, ChangeSim^{10}, ...).

A.2 Architectural Details

VPR module. We use a pre-trained BoQ model [2] without any architectural modifications. The BoQ feature extractor is a pretrained ResNet-50 [4].

Semantic aggregator. We employ multi-head attention with 6 heads and a dropout rate of 0.1. The feed-forward network (FFN) that follows the attention layer has a dimensionality of 384.

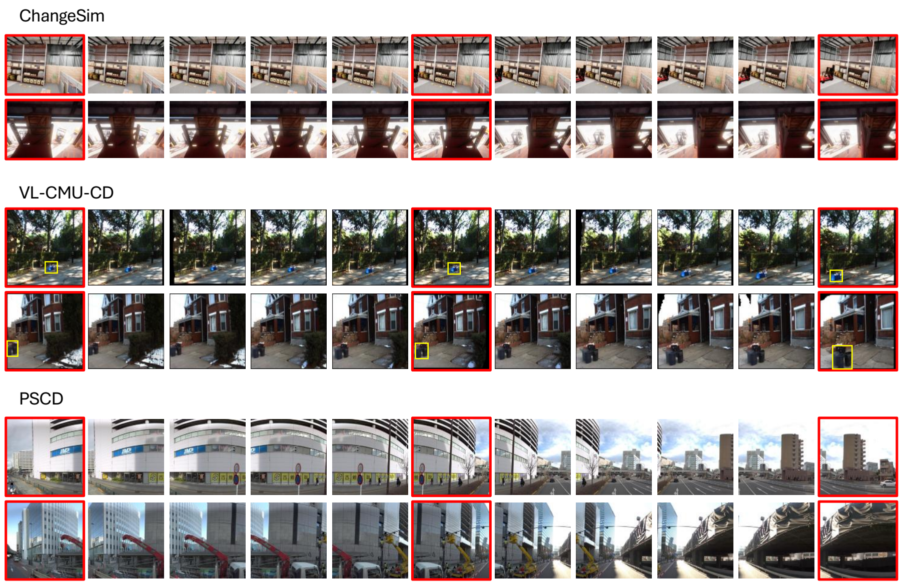

Fig. 1: Example images from the reference database. The figure compares the full reference images used in the conventional SCD setting with the strided reference database adopted in the proposed ECD setup. In the SCD setting, all images are adopted as the reference, whereas in the ECD setup, only a subset of images is selected. The images outlined in red boxes indicate the reference images selected for the ECD reference database using a database stride of $s = 5$. By applying a stride, ECD reduces the spatial and temporal density of the reference set, thereby intentionally introducing misalignment between queries and available references. Best viewed when zoomed in.

Change detection head. The change detection head operates on the reconstructed scene f_r^*, and the query feature f_q. The architecture of the change detection head is identical to that of RSCD [5]. First, two cross-attentions are applied in both directions from f_r^* to extract information corresponding to f_q, and vice versa. The resulting features are concatenated and passed through a sequence of layers: a 3×3 convolution that halves the channel dimension, followed by a 1×1 convolution to produce a single-channel change prediction map, and finally an upsampling layer. This process produces the final change detection output.

A.3 Computational Resources

The experiments were performed on a system equipped with the Intel Xeon Gold 6426Y CPU and a single NVIDIA RTX A5000 GPU. The software environment consisted of Python 3.10.16 and PyTorch 2.1.0, with CUDA version 12.2. The total GPU memory usage during the experiments remained below 24 GB. Training was completed in under one day, whereas PSCD is reserved exclusively for testing and does not participate in training.

Table 1: The accuracy of the VPR module. Each value is the average accuracy (%) computed on the training, validation, and test splits.

Database stride s	ChangeSim^s		VL-CMU-CD^s		PSCD^s		Average (Strict)	Average (Coarse)
	Strict	Coarse	Strict	Coarse	Strict	Coarse		
1	53.45	99.35	32.59	84.45	93.08	97.90	59.71	93.90
3	51.93	99.29	23.27	84.39	32.51	96.98	35.91	93.56
5	48.35	99.32	18.64	84.13	19.86	88.04	28.95	90.50
10	42.77	98.66	12.65	83.74	13.28	68.29	22.90	83.56

B VPR Retrieval Analysis

We evaluate the top-1 retrieval accuracy of the VPR module to assess its effectiveness in retrieving relevant reference images under different database stride s values. To support this analysis, we define two levels of matching criteria: *strict match* and *coarse match*. Note that these criteria are used exclusively for analysis in this section and are not incorporated into the main pipeline.

A *strict match* is defined as follows: for VL-CMU-CD^s and PSCD^s, it corresponds to the exact ground-truth reference pair; for ChangeSim^s, it refers to retrieved images within 1 meter and 10 degrees of the query, following SimSac [7], which only considers distance. A *coarse match* is defined more loosely: for VL-CMU-CD^s and PSCD^s, it includes any image from the same sequence as the ground-truth; for ChangeSim^s, it refers to retrieved images within 25 meters and 45 degrees of the query, following the retrieval metric [8].

The results are presented in Tab. 1. As the database stride s increases, the retrieval accuracy for strict matches drops sharply, whereas coarse match rates remain relatively stable. This indicates that strictly relying on GT-aligned references becomes increasingly impractical at larger strides. Instead, models must leverage nearby images to support accurate predictions. Notably, for VL-CMU-CD^s, the top-1 retrieval accuracy is lower compared to other datasets. This is because there is minimal viewpoint variation within each sequence, making many reference images visually similar and difficult to distinguish.

C Additional Analyses

Reference subset size. As shown in Tab. 2, the performance generally improves as the size of the reference subset K (*i.e.*, the number of reference images) increases, since a larger subset provides richer contextual cues. However, the performance degrades when K becomes too large, due to the inclusion of irrelevant or noisy references. This degradation becomes more pronounced as the database stride increases, because a larger stride yields a sparser reference database, making it easier for unrelated references to appear in $\mathcal{R}$ even with smaller K. Considering the trade-off between accuracy and computational efficiency, we set $K = 3$ as the default configuration.

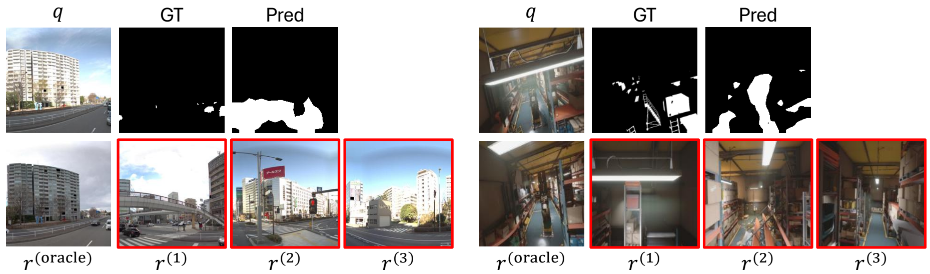

Fig. 2: Failure Analysis. The first stage of our solution relies on constructing a reference subset using the VPR module. Consequently, when the VPR retrieval fails, the subsequent change detection process also fails due to the lack of relevant reference cues. These observations reveal the limitations of the current pipeline and motivate future research on more robust or retrieval-independent reference database exploration strategies.

EMC grid resolution. In Tab. 3, we study how the choice of grid size in EMC affects performance. 'Hier.' denotes a hierarchical configuration: hierarchical $\times 2$ averages representations obtained from 1×1 and 2×2 grids, while hierarchical $\times 3$ additionally incorporates a 4×4 grid level. We observe that hierarchical configurations consistently outperform fixed single-scale grids, with hierarchical $\times 3$ achieving the highest average performance across both dataset stride settings.

Overlap-based adaptive striding. To consider different viewpoint characteristics across datasets, we evaluate overlap-based adaptive striding. Specifically, we construct the reference database by selecting the first frame whose r-r overlap falls below a predefined threshold. Tab. 4a reports the performance under this adaptive striding strategy. The table shows that the performance gaps across datasets are reduced. This provides an alternative criterion for selecting a stride when balancing performance and database size.

PSCD fine-tuning protocol. Although RSCD [5] primarily treats PSCD [9] as a test-only benchmark, it additionally reports results under a PSCD fine-tuning protocol. For a more comprehensive evaluation, we adopt the same protocol and present the corresponding results in Tab. 4b. Our method also achieves higher performance than RSCD under this setting.

D Failure Analysis

To further complement our evaluation, we present representative failure cases in Fig. 2. These examples show typical errors, mainly caused by inaccurate retrieval in the VPR stage. The first stage of our solution relies on constructing a reference subset using the VPR module. When VPR retrieval fails, the subsequent change detection process also fails. These failure cases motivate future research on alternative ways of exploring the environment database.

Table 2: Analysis experiment on the size of the reference subset K. Performance improves as K increases, but degrades when excessive references are included.

s	K	VL-CMU-CD^s	PSCD^s	ChangeSim^s	Average
1	1	0.6282	0.3300	0.3889	0.4490
	3	0.7043	0.3308	0.3924	0.4758
	5	0.7100	**0.3397**	0.3940	0.4812
	10	**0.7281**	0.3255	**0.3950**	**0.4829**
5	1	0.4908	**0.2678**	0.3855	0.3814
	3	**0.5456**	0.2246	0.3879	**0.3860**
	5	0.5203	0.2325	**0.4010**	0.3846
	10	0.5235	0.2316	0.3984	0.3845

Table 3: Analysis experiment on the grid size of the environment mosaic composer. Hierarchical settings generally show improved performance, with 'Hier. ×3' achieving the best result.

s	Grid size	VL-CMU-CD^s	PSCD^s	ChangeSim^s	Average
1	1×1	0.6929	0.3339	0.3794	0.4687
	2×2	0.6884	0.2651	0.3849	0.4461
	4×4	0.5645	0.2623	**0.3930**	0.4066
	Hier. ×2	0.6957	**0.3446**	0.3826	0.4743
	Hier. ×3	**0.7043**	0.3308	0.3924	**0.4758**
5	1×1	0.4755	**0.2562**	0.3790	0.3702
	2×2	0.4986	0.1790	0.3863	0.3546
	4×4	0.4089	0.1197	0.3851	0.3046
	Hier. ×2	0.5181	0.2230	0.3868	0.3760
	Hier. ×3	**0.5456**	0.2246	**0.3879**	**0.3860**

E Visualization and Analysis of EMC

We provide visual examples in Figs. 3 to 5 to showcase the effectiveness of our proposed EMC in producing hierarchical mosaic environment representations. In the visual examples, q denotes the query image, $r^{(\text{oracle})}$ is the ground truth reference image, and $r^{(k)} \in \mathcal{R}$ are the retrieved reference subset $\mathcal{R}$. Note that $r^{(\text{oracle})}$ is not guaranteed to be included in the reference database depending on the query and database stride, but is shown here for illustrative purposes. The retrieved reference images are visually marked as follows: green borders indicate a strict match, blue borders indicate a coarse match, and red borders indicate incorrect matches.

Although $\{\tilde{f}_{r,n}\}$ are feature representations, we visualize them as images by projecting the matched feature patches onto their spatial coordinates. For improved visibility, we also provide a colored version in which the color indicates the origin of each patch from the reference images. Specifically, $r^{(1)}$, $r^{(2)}$, and $r^{(3)}$

Table 4: Analysis experiments.

(a) Performance under adaptive striding based on r-r overlap thresholds.

overlap	VL-CMU-CD	PSCD	ChangeSim
95.0%	0.4420	-	0.3709
85.0%	0.3696	-	0.3646
75.0%	0.3543	0.2684	0.3581

(b) SCD results under the PSCD finetuning setting.

Method	$\text{PSCD}^{\text{finetune}}$
DR-TANet [3]	0.190
C-3PO [10]	0.433
RSCD [5]	0.442
Ours	**0.446**

are colored in cyan, magenta, and yellow, respectively. To aid in understanding viewpoint alignment, corresponding regions are highlighted with black or orange boxes.

Fig. 3 presents examples of mosaic environment representations. These results show that our method can imitate the query view by selecting high-similarity patches for each query patch from multiple references. In particular, the mosaic environment representations on the 4×4 grid effectively approximate the target viewpoints.

Fig. 4 illustrates a case involving changes to a large object. Because EMC operates at the patch-level similarity, alignment accuracy may degrade at finer grid resolutions, especially when a patch is entirely altered. Even so, the hierarchical design maintains robust alignment at coarser grid levels. Thus, large changes and substantial scene variations are still handled effectively. These results highlight the complementary roles of fine- and coarse-level alignment, validating our hierarchical design.

Fig. 5 depicts a challenging case with a large database stride s, where no correctly aligned reference images are available and many incorrect distractors are included in the reference subset. Even under these conditions, EMC effectively suppresses the influence of unrelated images and reconstructs the view primarily using the relevant match.

Overall, these results provide strong evidence that EMC can operate reliably across a wide range of conditions. However, the outputs may show semantic inconsistencies, particularly near patch boundaries. Therefore, we leverage the strong spatial alignment provided by the mosaic representation to employ the following semantic aggregator to integrate the reference features.

F Qualitative Results

We present additional qualitative results across various datasets in Figs. 6 to 8. We provide prediction results for each query image q under two different dataset stride s settings, to illustrate how spatial information and prediction difficulty vary depending on the sparsity of the reference database.

As in the previous section, $r^{(\text{oracle})}$ is the ground truth reference image, and $r^{(k)} \in \mathcal{R}$ are the retrieved reference subset $\mathcal{R}$. The borders of the reference images are colored to indicate match quality: green for strict matches, blue for coarse matches, and red for incorrect matches.

Our method achieves accurate predictions even without exact viewpoint matches by leveraging nearby frames. Moreover, it demonstrates greater robustness to viewpoint changes compared to the baseline. Even when the VPR retrieval includes incorrect matches, most notably observed in the PSCD[5], our model can still produce accurate predictions, thanks to effective attention mechanisms.

G Broader Impacts

SCD and the proposed ECD can be applied to a wide range of real-world scenarios, including natural disaster assessment, urban development monitoring, and automated warehouse management. These technologies have the potential to enhance safety, efficiency, and environmental awareness. However, their deployment may raise privacy concerns and enable intrusive surveillance applications. Care must be taken to ensure that such systems are developed and used in a way that respects individual rights and complies with legal and ethical standards.

Fig. 3: Examples of mosaic environment representations. Given a query image q, EMC generates mosaic environment representations at multiple grid levels using reference images $r^{(1)}$, $r^{(2)}$, and $r^{(3)}$. $r^{(\text{oracle})}$ is shown for reference purposes. At finer grid levels, the generated mosaics closely imitate the viewpoint of the query. The borders of the reference images are colored to indicate match quality: green for strict matches, blue for coarse matches, and red for incorrect matches.

Fig. 4: Examples of mosaic environment representations under large object changes. When a substantial change spans the entire area of a patch, the mosaic environment representation at fine grid levels may fail to align accurately. However, the mosaic environment representation at coarse grid levels remains robust and provides strong evidence for the effectiveness of the hierarchical approach.

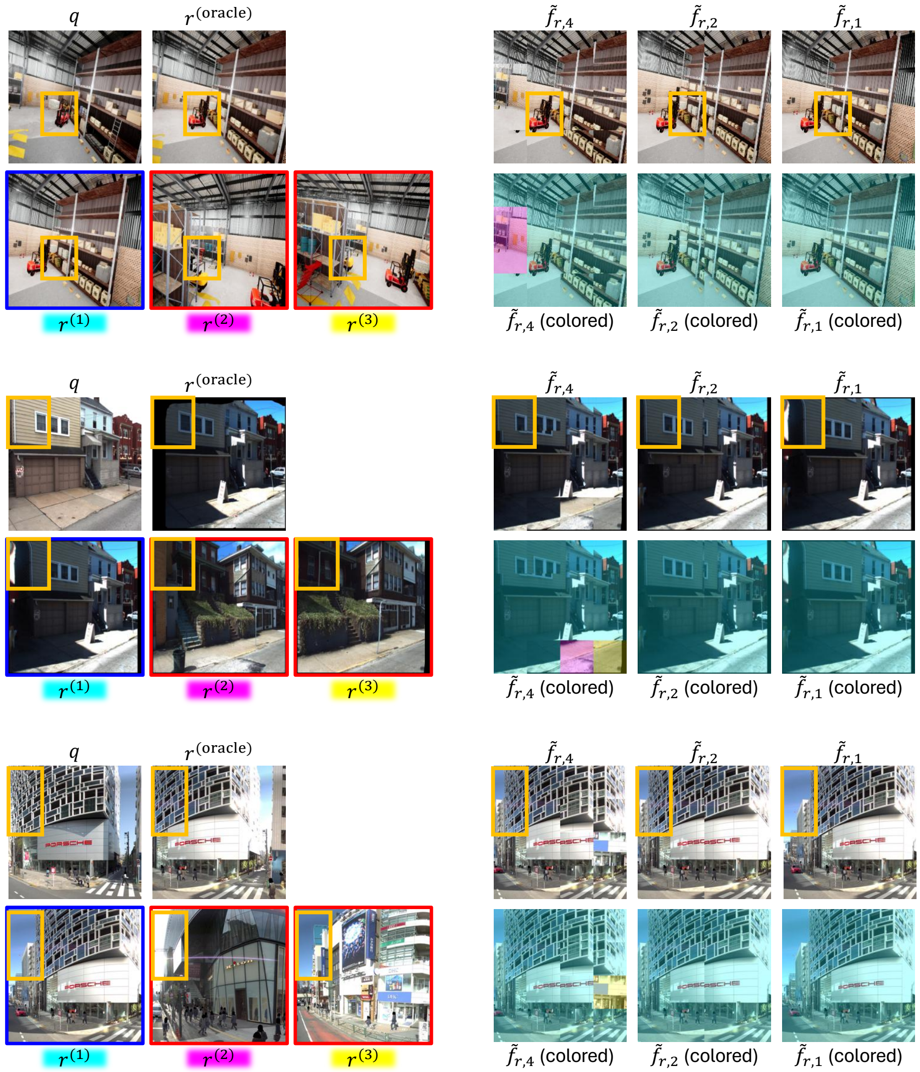

Fig. 5: Examples of mosaic environment representations under a large database stride and incorrect retrieval. Even when the reference subset $\mathcal{R}$ contains multiple incorrect references, EMC effectively suppresses the influence of unrelated images and reconstructs the mosaic environment representation primarily using the most relevant match.

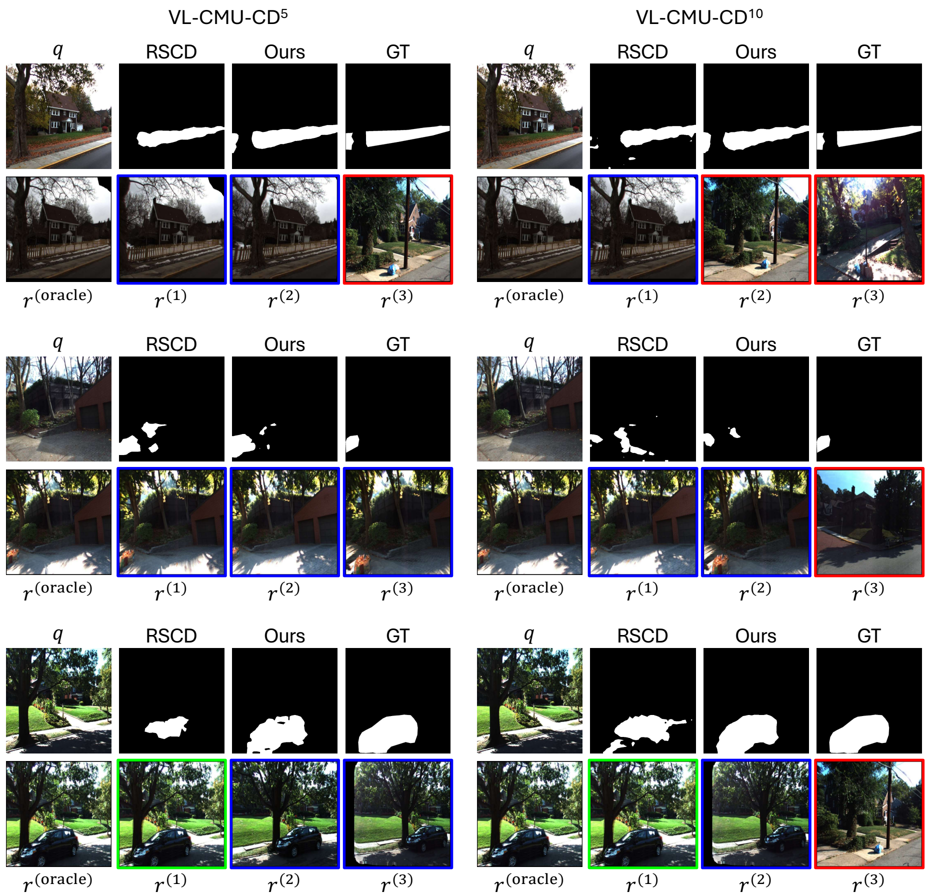

Fig. 6: Qualitative results on VL-CMU-CD^5 and VL-CMU-CD^{10}.

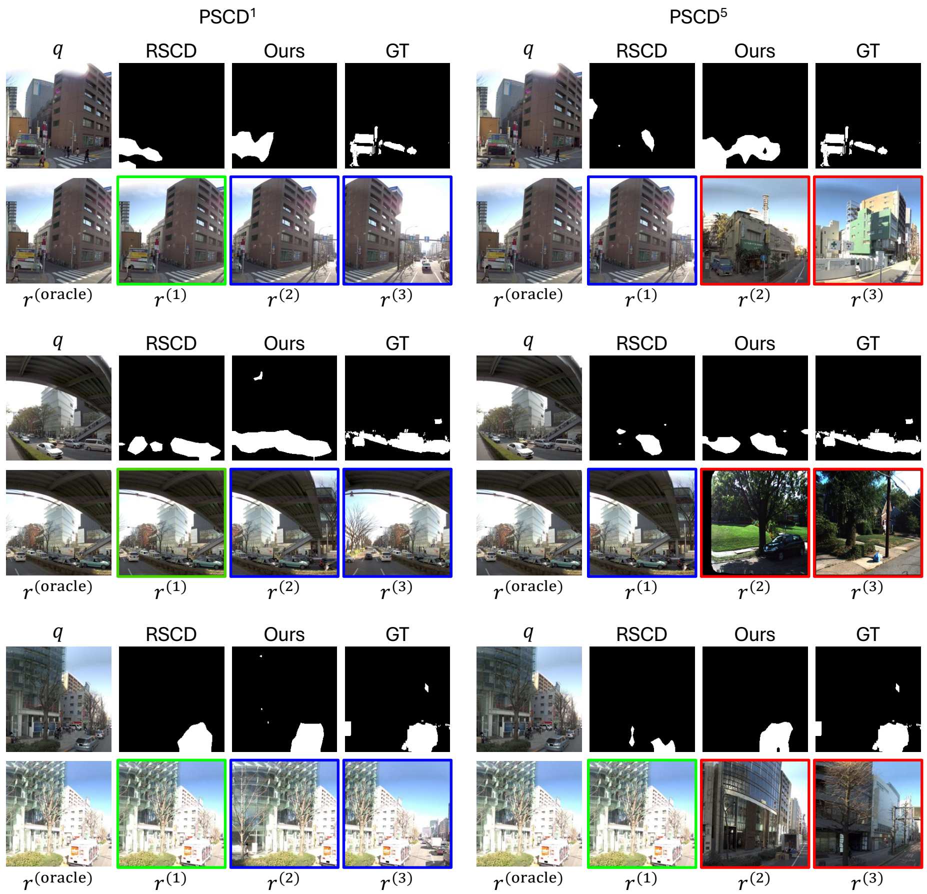

Fig. 7: Qualitative results on PSCD^1 and PSCD^5.

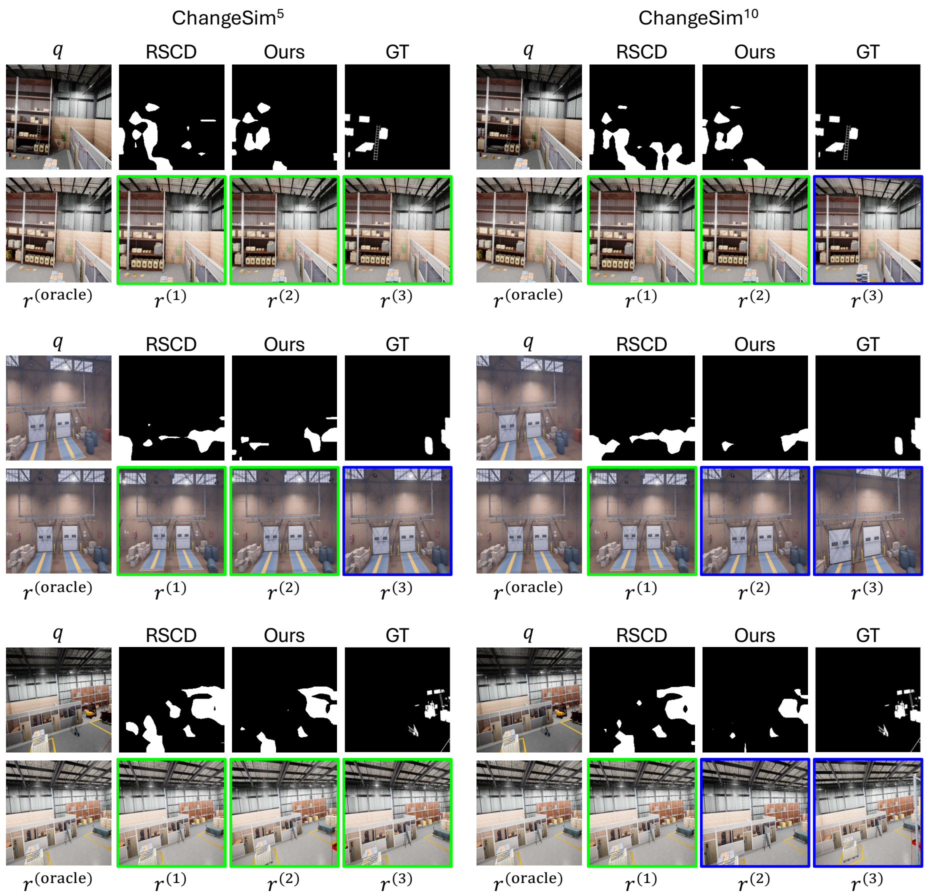

Fig. 8: Qualitative results on ChangeSim^5 and ChangeSim^{10}.

References

1. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. Autonomous Robots **42**, 1301–1322 (2018)
2. 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)
3. 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)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. 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)
6. 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)
7. 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)
8. Pion, N., Humenberger, M., Csurka, G., Cabon, Y., Sattler, T.: Benchmarking image retrieval for visual localization. In: 2020 International Conference on 3D Vision (3DV). pp. 483–494. IEEE (2020)
9. 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)
10. Wang, G.H., Gao, B.B., Wang, C.: How to reduce change detection to semantic segmentation. Pattern Recognition **138**, 109384 (2023)