Towards Generalizable Scene Change Detection

Supplementary Material

1. ChangeVPR Dataset

The meticulously annotated ChangeVPR dataset is designed to evaluate the robustness and generalizability of scene change detection models. It covers a diverse range of environments with challenging scenarios—significantly expanding the traditional urban or synthetic-only SCD scope. We source the images from three widely used visual place recognition datasets with different environmental characteristics: SF-XL [2] (urban), St Lucia [18] (suburban), and Nordland [24] (rural) dataset (see Fig. 1). The detailed descriptions of each split are as follows:

- **SF-XL** is a vast dataset covering San Francisco, with over 41M images collected from Google Street View. The temporal distribution of the dataset ranges from 2007 to 2020. We quantize the whole dataset using classification strategy [2] and carefully select two bi-temporal images from quantized cells for query and reference. Following [2], we consider two bi-temporal images to be of the same place (same class) if they are located in the same geographical cell (a cell with a side of 10 meters) and their heading difference is less than 30 degrees.
- **St Lucia** features video recordings from a car-mounted camera, capturing multiple drives through the St Lucia suburb of Brisbane. Following [3], among nine available drives, we use the first drive as a query and the last as reference sets. Since there are no heading labels for the image, we carefully select image pairs after sampling two bi-temporal images with a UTM distance of less than 10 meters.
- **Nordland** captures a train journey through the Norwegian countryside, capturing the same route across four seasons with frames extracted at 1 FPS. Following previous VPR works [11, 12], which use the winter traverse as queries and summer as a database, we adopt winter traverse as a query and summer as a reference. Since no heading labels exist, we follow the same process in the St Lucia split.

We manually labeled the ground-truth for scene changes, providing it as a binary image, matching the size of the input image pairs of resolution 512×512. In this binary image, each pixel value indicates whether a change occurred at the corresponding point between the bi-temporal images. Following the convention [1, 22], we define scene changes as both 2D surface alterations (e.g., changes to advertising boards) and 3D structural modifications (e.g., the appearance or disappearance of buildings, vehicles, trash bins, and pedestrians). Finally, the ChangeVPR dataset contains the binary change masks C_{t0}, C_{t1} and the intersection change mask $C_{t0 \leftrightarrow t1}$, respectively.

Figure 1. **Examples of ChangeVPR dataset.** (Top) SF-XL. (Middle) St Lucia. (Bottom) Nordland. Each row of the splits represents a query, reference, and ground-truth mask.

2. Implementation Details

2.1. Hyperparameters of GeSCF

To ensure general applicability and consistent performance on real-world applications, *we uniformly set the SAM pa-*

Method	Training	VL-CMU-CD			TSUNAMI			ChangeSim			Avg.
		$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	
SimSaC	VL-CMU-CD†	**75.5**	42.3	0.39	32.0	41.9	0.27	53.5	32.2	0.24	46.2
	ChangeSim†	74.1	17.7	0.12	17.7	20.6	0.13	**62.3**	37.3	0.29	38.2
GeSCF (Ours)	Zero-shot	75.4	**75.4**	**1.0**	**72.8**	**72.8**	**1.0**	54.8	**54.8**	**1.0**	**67.7**

Table 1. **Quantitative results of SimSaC on standard SCD datasets.** † indicates that an additional Synthetic [21] dataset is used for training.

Method	Training	SF-XL (U)			St Lucia (S)			Nordland (R)			Avg.
		$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	
SimSaC	VL-CMU-CD†	56.3	55.0	0.31	51.2	51.8	0.32	28.1	25.5	0.25	44.7
	ChangeSim†	44.3	40.5	0.12	34.3	36.7	0.04	19.4	17.5	0.05	32.1
GeSCF (Ours)	Zero-shot	**71.2**	**71.2**	**1.0**	**62.1**	**62.1**	**1.0**	**59.0**	**59.0**	**1.0**	**64.1**

Table 2. **Quantitative results of SimSaC on ChangeVPR dataset.** † indicates that an additional Synthetic [21] dataset is used for training.

Method	Training	VL-CMU-CD			TSUNAMI			ChangeSim			Avg.
		$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	$t0 \rightarrow t1$	$t1 \rightarrow t0$	TC	
C-3PO	VL-CMU-CD	**77.4**	4.5	0.02	5.6	19.4	0.02	25.5	13.6	0.09	24.3
C-3PO*		59.0	59.0	1.0	5.1	5.1	1.0	17.2	17.2	1.0	27.1
GeSCF (Ours)	Zero-shot	75.4	**75.4**	**1.0**	**72.8**	**72.8**	**1.0**	**54.8**	**54.8**	**1.0**	**67.7**

Table 3. **Quantitative results of C-3PO variants on VL-CMU-CD dataset.** * indicates the (I+A+D+E) structure; otherwise, the (I+D) structure for C-3PO.

rameters for all configurations, including the standard SCD datasets and our ChangeVPR dataset. Specifically, we employ SAM ViT-H with a point per side of 32, an NMS threshold of 0.7, a predicted IoU threshold of 0.7, and a stability score threshold of 0.7. For the adaptive threshold function in GeSCF, we adopt b_γ=0.05 and s_γ=0.1 for right-skewed distributions (γ>0.2), and b_γ=0.7 and s_γ=1.0 for left-skewed distributions (γ<−0.2). For moderate distributions (−0.2≤γ≤0.2), we employ the z-score method with a z-value of −0.52. Further, we intercept key facets from the 17th layer of the SAM ViT-H encoder. In the Geometric-Semantic Mask Matching module, we set α_t to 0.65 and change confidence score to 0.88. We conduct a linear search on a small validation set sampled from the VL-CMU-CD training set to find the optimal hyperparameters. Importantly, after setting these hyperparameters, GeSCF requires no further tuning for specific datasets——the same hyperparameters are used consistently across all settings.

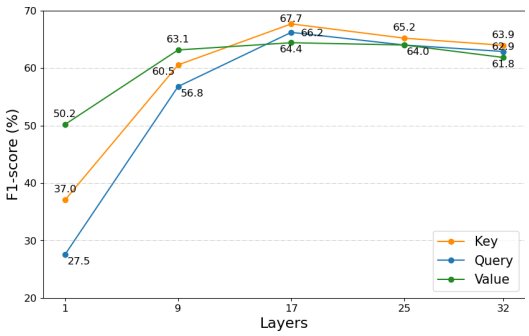
2.2. Adaptor Networks

In the absence of SAM-based baselines, we have meticulously constructed various adaptor networks based on relevant literature for ablation. As change detection necessitates processing bi-temporal features, adaptor [27] networks are commonly used when leveraging foundation models—which usually take single-image inputs—in the CD domain [7, 17]. To demonstrate our effectiveness of GeSCF in SAM's feature utilization for generalizability through ablation studies, we configured several adaptor networks on the frozen SAM ViT-H image encoder with three prominent feature processing modules in SCD: Correlation layers [8], CHVA [4], and feature merging modules [25]. Specifically, following [7], we intercept four image embeddings at different ViT blocks to obtain local and global features from the SAM ViT image encoder [10]. We evenly select features after the global attention layers of the SAM image encoder [13]; for the ViT-H model, these are the outputs of the 7th, 15th, 23rd, and 31st blocks output for the 32 blocks in total. The intercepted features are then interpolated to form a multi-scale feature pyramid, which is input to the feature processing modules. The processed image features are subsequently fed to a decoder network, following the configurations in [4, 23, 25]. As shown in the manuscript, leveraging SAM with learnable adapters significantly degrades generalizability, since the current SCD datasets are not representative enough to cover diverse real-world changes.

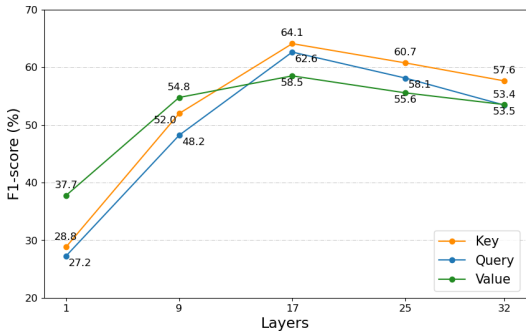
3. More Experiments

3.1. More Quantitative Comparisons

Comparison with SimSaC. In Tabs. 1 and 2, we provide additional comparison results between our GeSCF and SimSaC [21]. We were unable to perform full assessments due to the incomplete release of the necessary code and the Synthetic dataset [21] required for model training. Instead, we leveraged the pre-trained models avail-



F1-score on standard SCD datasets



F1-score on ChangeVPR dataset

Figure 2. **Comparative analysis of GeSCF performance using various facets and layers across standard SCD datasets (VL-CMU-CD [1], TSUNAMI [22], and ChangeSim [20]) and the ChangeVPR dataset.** The key facet from the intermediate layer achieves the best performance among other choices, highlighting that our facet and layer selection is quantitatively a reasonable design choice.

Backbone	SCD Datasets		ChangeVPR	
	F1-score	mIoU	F1-score	mIoU
DINOv2 (ViT-B)	60.8	43.6	50.7	32.8
SAM (ViT-B)	61.5	42.9	53.0	34.4
DINOv2 (ViT-L)	62.3	45.3	50.9	35.9
SAM (ViT-L)	64.8	47.6	59.4	40.4
SAM (ViT-H)	**67.7**	**51.3**	**64.1**	**46.3**

Table 4. **Performance comparison of GeSCF using different ViT backbones.** We maintain the default GeSCF configuration, modifying only the backbone during the initial pseudo-mask generation and semantic similarity matching processes.

able on VL-CMU-CD [1] and ChangeSim [20], along with the Synthetic dataset [21] accessible from the official website. The results align with those presented in the main paper, demonstrating that our GeSCF shows superior robustness on unseen domains and achieves comparable performance on seen domains, all while ensuring complete temporal consistency. Overall, our GeSCF surpasses SimSaC with an exceptional margin, achieving an average improvement of 21.5% on standard SCD datasets and 19.4% on the ChangeVPR dataset.

Comparison with C-3PO Variant. C-3PO [25] introduces a series of model variants, each incorporating an inductive bias tailored to specific training domains. For example, the (I+D) structure is utilized for the VL-CMU-CD dataset, whereas the (I+A+D+E) structure is employed for the TSUNAMI and ChangeSim datasets. However, applying models with the appropriate prior knowledge in real-world scenarios remains challenging, as it extends beyond the well-explored research datasets. The discrepancy between the model's built-in assumptions and the application domain can result in suboptimal performance (see Tab. 3). Notably, the performance of C-3PO (I+A+D+E) on the VL-CMU-CD dataset is significantly hindered due to the incorrect assumption, even though it maintains complete temporal consistency. In contrast, our GeSCF provides a robust and unified framework that consistently performs well, maintaining temporal consistency across diverse application domains.

3.2. More Qualitative Evaluations

We present extensive results that demonstrate the effectiveness of our approach across various change scenarios in both seen and unseen domains (see Figs. 3 and 4). Unlike conventional SCD methods, which are confined to detecting changes within specific training datasets, our GeSCF captures changes across diverse environments without requiring SCD supervision. Moreover, GeSCF accurately detects meaningful changes beyond dataset biases, achieving performance comparable to or surpassing other supervised in-domain baselines. This capability addresses the long-standing issue of dataset dependency in the SCD field—establishing a robust foundation for a truly applicable and versatile *anything* SCD.

3.3. Exploring Design Choices in GeSCF

Our GeSCF builds upon the Segment Anything Model [14] (SAM) leveraging the intermediate key facets of SAM ViT image encoder during the initial pseudo-mask generation process, and utilizes the final mask embeddings for the Semantic Similarity Matching (SSM). To validate these design choices, we conduct comprehensive ablation studies.

Key Facets from the Intermediate Layer. As explained in the main paper, we chose to use key facets of the intermediate layer for the multi-head feature correlation and generate initial pseudo-masks. As we will see here, this choice provides the best change detection performance among other alternatives (see Fig. 2). While all facets perform best in the intermediate layer, the key facet outperforms the query and value facets in the F1-score. Specifically, on standard SCD

Figure 3. **More qualitative examples on the seen domain (standard SCD datasets).** Although GeSCF does not learn dataset biases, it can accurately segment meaningful changes with comparable performance or even better than other in-domain baselines. Moreover, our GeSCF can effectively segment unannotated semantic changes highlighted with yellow bounding boxes.

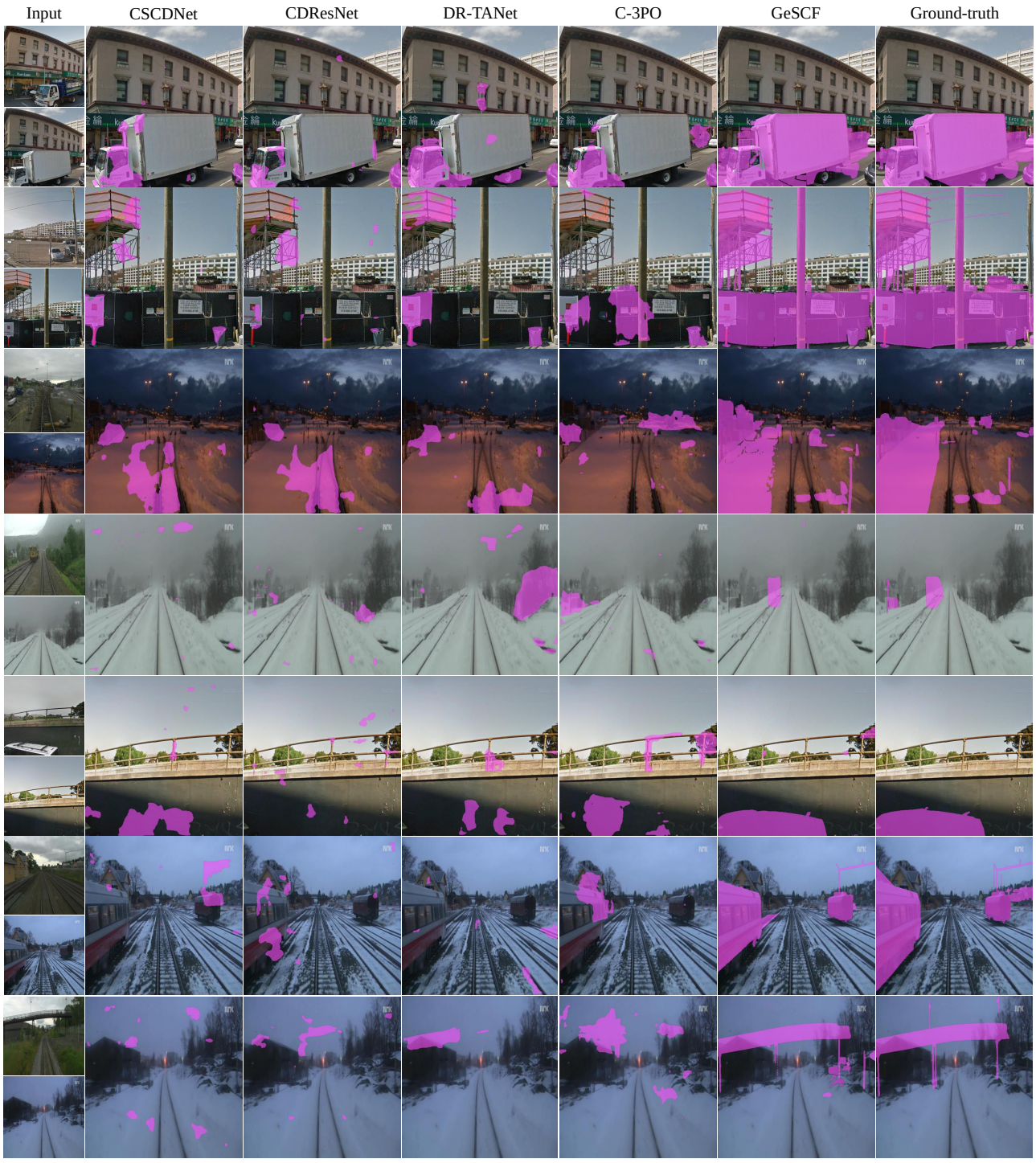


Figure 4. **More qualitative examples on the unseen domain (ChangeVPR).** Our GeSCF produces accurate and sharp masks with exceptional generalizability, capturing changes across diverse and challenging scenarios more effectively than other existing baselines.

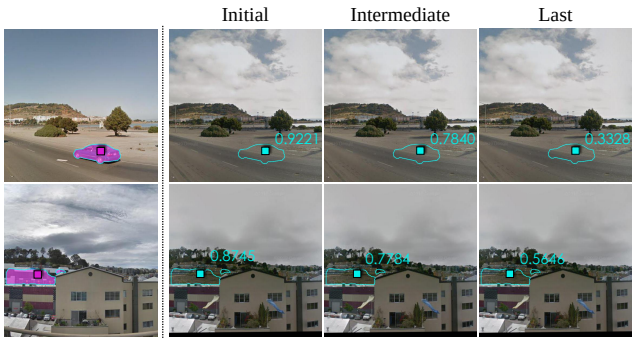


Figure 5. **Cosine similarity between bi-temporal mask embeddings depending on the layer.** The object-level semantic difference is more pronounced in the last layer compared to the initial and intermediate layers.

Layer	Score	SCD Datasets		ChangeVPR	
		F1-score	mIoU	F1-score	mIoU
Initial	0.88	61.1	43.5	61.2	42.4
Intermed.		67.3	51.0	63.8	46.0
Last	0.72	65.5	47.9	63.0	44.7
Last	0.95	66.7	50.0	63.3	45.2
Last	0.88	**67.7**	**51.3**	**64.1**	**46.3**

Table 5. **Performance comparison of GeSCF using different mask embedding layers and change confidence score in semantic similarity matching.**

Method	Training	Resolu.	Perturbation Scale				Avg.
			0%	5%	15%	25%	
CSCDNet	in-domain	256×256	71.9	70.5	70.3	69.7	70.6
CDResNet			70.1	69.2	68.6	67.9	69.0
DR-TANet			70.1	69.4	68.3	67.0	68.7
C-3PO			72.8	72.2	71.9	71.1	72.0
GeSCF	zero-shot	256×256	**74.9**	**74.2**	**72.8**	**71.4**	**73.3**

Table 6. **Quantitative comparison (F1-score) on VL-CMU-CD.**

Method	Training	CDnet 2014			Avg.
		Bus	Tram	Boats	
3DCD [16]	in-domain	0.79	0.75	**0.88**	0.81
GeSCF	zero-shot	**0.80**	**0.81**	0.86	**0.82**

Table 7. **Quantitative comparison (F1-score) on CDnet 2014.**

datasets, the key facet attains an F1-score of 67.7%, which is 1.5% higher than the query facet (66.2%) and 3.3% higher than the value facet (64.4%). Similarly, on the ChangeVPR dataset, utilizing the key facet results in an F1-score of 64.1%, outperforming the query facet by 1.5% (62.6%) and the value facet by 5.6% (58.5%). These results emphasize that our selection strategy is quantitatively a well-justified choice.

Different ViT Backbones. We incorporate various ViT backbones of different sizes from SAM and DINOv2 [19] (see Tab. 4). The results show that ViTs of SAM variants outperform their DINOv2 counterparts and larger ViT yields superior performance compared to smaller models.

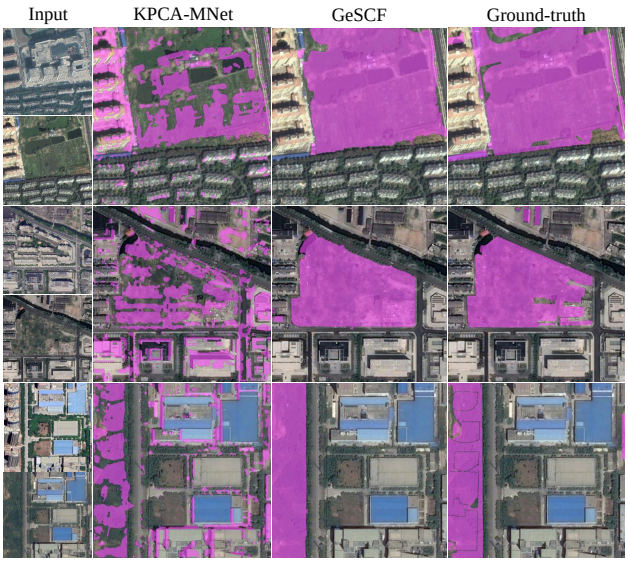


Figure 6. **Qualitative results of GeSCF on the SECOND (`test`) benchmark.** Our GeSCF can also perform zero-shot remote sensing CD, demonstrating its versatility and potential across different CD domains.

Mask Embeddings from the Last Layer. Previous studies [10] have empirically demonstrated the ViT's transition from low-level feature encoding in early layers to capturing more global, semantic representations in deeper layers. We observe that this hierarchical feature specialization also exists in SAM ViT (see Fig. 5). Additionally, we perform quantitative ablation studies regarding mask embedding layers and change confidence scores (see Tab. 5). Our results show that utilizing mask embeddings from the final layer yields superior performance compared to embeddings from the initial or intermediate layers. Furthermore, a high change confidence score can result in the inclusion of irrelevant masks, while a low score becomes overly strict, hindering the detection of actual changes. Therefore, our change confidence score acts as a semantic lower bound within SAM's latent space for accurate change detection.

3.4. Regarding Registration and Resolution

To assess robustness against perspective and resolution variations, we simulate perspective distortions by applying random perturbations of varying scales (up to a 64-pixel offset) to images downscaled to half of their original 512×512 resolution on the VL-CMU-CD dataset (see Tab. 6). While all baselines are trained on full-resolution images, GeSCF maintains a zero-shot setting, augmented with a simple SuperPoint [6] + RANSAC [9] registration module.

3.5. Beyond Scene Change Detection

Although our research primarily focuses on natural scene CD, with each CD domain typically focusing on its specialized area rather than integrating other CD fields [15], we have discovered that our framework can also work as a zero-shot remote sensing CD framework (see Fig. 6 for qualitative examples). Despite SAM being exclusively trained on natural images from photographers, SAM's feature space and our proposed feature strategies remain effective when applied to remote sensing data. Furthermore, we also evaluate our proposed GeSCF against the video sequence CD method from [16] on the CDnet 2014 [26] benchmark (see Tab. 7). The experimental results further confirm the exceptional generalizability of GeSCF. Although these results are preliminary, they indicate that the representations from SAM may be useful for different CD domains.

4. Training Objectives

To comprehensively evaluate temporal consistency, we train all baselines using both uni-temporal and bi-temporal training objectives [28].

Uni-temporal Objective is a standard binary cross-entropy loss, a widely adopted loss function for binary classification tasks. It quantifies the discrepancy between the predicted change probabilities and the ground-truth, effectively guiding the network to accurately distinguish between change and no-change regions. The uni-temporal objective is calculated as follows:

$$\mathcal{L}_{bce} = -[y \log(p) + (1-y) \log(1-p)], \tag{1}$$

where $y \in 0, 1$ specifies the ground-truth class and $p \in [0, 1]$ denotes predicted probability for positive class.

Bi-temporal Objective [28] is specifically designed for binary change detection by leveraging paired images captured at different time steps. By jointly optimizing with both temporal directions, it complements the uni-temporal objective and further refines the model's temporal consistency. The bi-temporal objective is formulated as follows:

$$\mathcal{L}_{bce}^{t0 \leftrightarrow t1} = m\mathcal{L}_{bce}^{t0 \rightarrow t1} + n\mathcal{L}_{bce}^{t1 \rightarrow t0}, \tag{2}$$

where m and n are set to 0.5 and $t0 \rightarrow t1$ (or $t1 \rightarrow t0$) represents the concatenation order of the input images. Note that $\mathcal{L}_{bce}^{t0 \leftrightarrow t1}$ is equal to $\mathcal{L}_{bce}^{t0 \rightarrow t1}$ (or $\mathcal{L}_{bce}^{t1 \rightarrow t0}$) if the model entails temporally symmetric architectures.

5. Limitations

Generalizable scene change detection is an emerging and challenging task essential for advancing the SCD community. We are the first to define ***what is generalizable scene change detection*** and to introduce a straightforward yet powerful framework, complemented by a broader domain evaluation dataset and a comprehensive evaluation protocol. However, certain limitations persist, including the need for demonstration across more diverse domains, such as various indoor environments, and challenges related to the bias of SAM. These limitations leave room for subsequent research and interdisciplinary studies [5, 13] to enhance the robustness of our framework further.

References

[1] Pablo Fernández Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42:1301 – 1322, 2016. 1, 3

[2] Gabriele Berton, Carlos German Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4868–4878, 2022. 1

[3] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlos German Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5386–5397, 2022. 1

[4] Shuo Chen, Kailun Yang, and Rainer Stiefelhagen. Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 502–509, 2021. 2

[5] Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhuo Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4081–4091, 2024. 7

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–33712, 2017. 6

[7] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2023. 2

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[9] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 6

[10] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *ArXiv*, abs/2212.06727, 2022. 2, 6

[11] Stephen Hausler, Adam Jacobson, and Michael Milford. Multi-process fusion: Visual place recognition using multiple image processing methods. *IEEE Robotics and Automation Letters*, 4:1924–1931, 2019. 1

[12] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14136–14147, 2021. 1

[13] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *ArXiv*, abs/2306.01567, 2023. 2, 7

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 3

[15] Seonhoon Lee and Jong-Hwan Kim. Semi-supervised scene change detection by distillation from feature-metric alignment. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1224, 2024. 7

[16] Murari Mandal, Vansh Dhar, Abhishek Mishra, Santosh Kumar Vipparthi, and Mohamed Abdel-Mottaleb. 3dcd: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos. *IEEE Transactions on Image Processing*, 30:546–558, 2020. 6, 7

[17] Liye Mei, Zhaoyi Ye, Chuan Xu, Hongzhu Wang, Ying Wang, Cheng Lei, Wei Yang, and Yansheng Li. Scd-sam: Adapting segment anything model for semantic change detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. 2

[18] Michael Milford and Gordon F. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24:1038–1053, 2008. 1

[19] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 6

[20] Jin-Man Park, Jae-Hyuk Jang, Sahng-Min Yoo, Sun-Kyung Lee, Ue-Hwan Kim, and Jong-Hwan Kim. Changesim: Towards end-to-end online scene change detection in industrial indoor environments. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8578–8585, 2021. 3

[21] Jin-Man Park, Ue-Hwan Kim, Seonhoon Lee, and Jong-Hwan Kim. Dual task learning by leveraging both dense correspondence and mis-correspondence for robust change detection with imperfect matches. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13739–13749, 2022. 2, 3

[22] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *British Machine Vision Conference (BMVC)*, 2015. 1, 3

[23] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6861–6867, 2018. 2

[24] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. 2013. 1

[25] Guo-Hua Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation. *Pattern Recognition*, 138:109384, 2023. 2, 3

[26] Yi Wang, Pierre-Marc Jodoin, Fatih Murat Porikli, Janusz Konrad, Yannick Benezeth, and Prakash Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 393–400, 2014. 7

[27] Lingling Xu, Haoran Xie, S. Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *ArXiv*, abs/2312.12148, 2023. 2

[28] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15173–15182, 2021. 7