Towards Generalizable Scene Change Detection

Jae-Woo Kim[1] and Ue-Hwan Kim[1*]
[1]Department of AI Convergence
Gwangju Institute of Science and Technology, Gwangju, South Korea
kjw01124@gm.gist.ac.kr, uehwan@gist.ac.kr

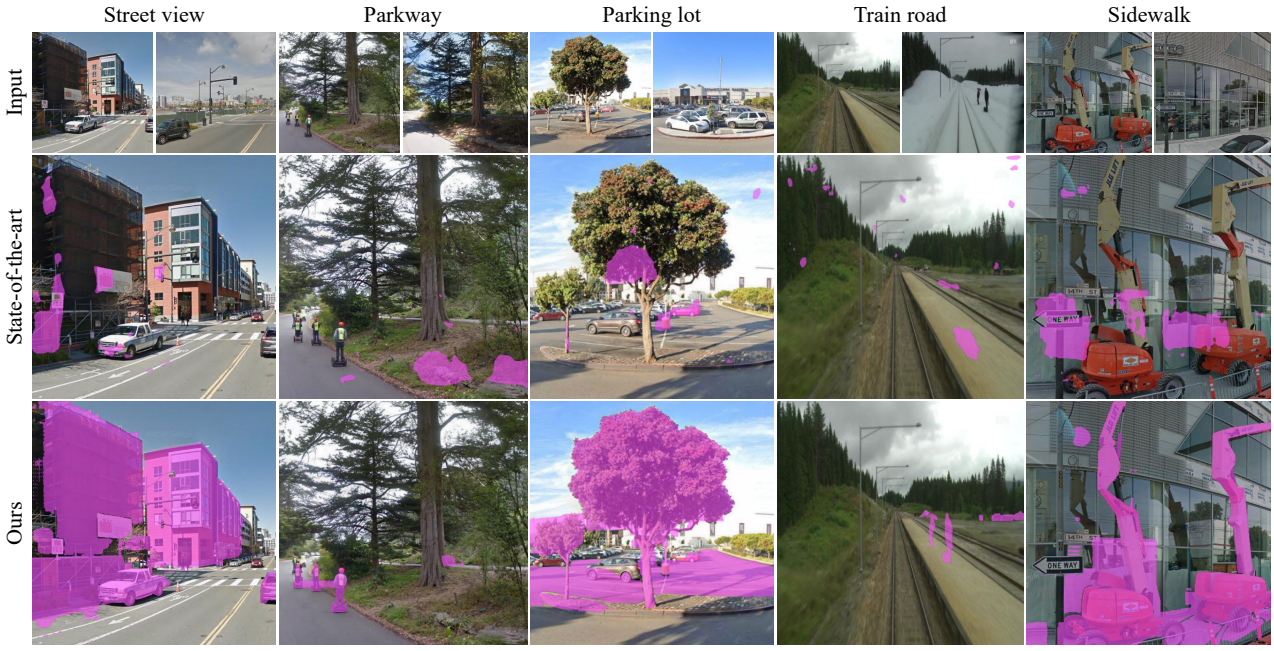

Figure 1. **Comparative results of the current state-of-the-art model and our GeSCF under various unseen environments (ChangeVPR).** GeSCF outperforms with precise boundaries and edges, where the state-of-the-art model hardly captures changes.

Abstract

While current state-of-the-art Scene Change Detection (SCD) approaches achieve impressive results in well-trained research data, they become unreliable under unseen environments and different temporal conditions; in-domain performance drops from 77.6% to 8.0% in a previously unseen environment and to 4.6% under a different temporal condition—calling for generalizable SCD and benchmark. In this work, we propose the Generalizable Scene Change Detection Framework (GeSCF), which addresses unseen domain performance and temporal consistency—to meet the growing demand for anything SCD. Our method leverages the pre-trained Segment Anything Model (SAM) in a zero-shot manner. For this, we design Initial Pseudo-mask Generation and Geometric-Semantic Mask Matching—seamlessly turning user-guided prompt and single-image based segmentation into scene change detection for a pair of inputs without guidance. Furthermore, we define the Generalizable Scene Change Detection (GeSCD) benchmark along with novel metrics and an evaluation protocol to facilitate SCD research in generalizability. In the process, we introduce the ChangeVPR dataset, a collection of challenging image pairs with diverse environmental scenarios—including urban, suburban, and rural settings. Extensive experiments across various datasets demonstrate that GeSCF achieves an average performance gain of 19.2% on existing SCD datasets and 30.0% on the ChangeVPR dataset, nearly doubling the prior art performance. We believe our work can lay a solid foundation for robust and generalizable SCD research.

Code and dataset are available here.
* Corresponding author

1. Introduction

Scene Change Detection (SCD) [37] is a pivotal technology that enables a wide range of applications: visual surveillance [54], anomaly detection [20], mobile robotics [31], and autonomous vehicles [21]. The ability to accurately identify meaningful changes in a scene across different time steps—despite challenges such as illumination variations, seasonal changes, and weather conditions—plays a key role in the system's effectiveness and reliability.

Recent SCD models have achieved remarkable improvement by leveraging deep features [3, 48] and advancing model architectures [11, 35, 44, 50]. However, this progress raises a fundamental question: ***"Can these models detect arbitrary real-world changes beyond the scope of research data?"*** Our findings, as shown in Fig. 1, indicate that their reported effectiveness does not hold in real-world applications. Specifically, we observe that they produce inconsistent change masks when the input order is reversed and exhibit significant performance drops when deployed to unseen domains with different visual features. It is because current SCD methods rely heavily on their training datasets, which are often limited in size [3, 42], sparse in coverage [3, 42, 44], and predominantly synthetic [25, 34, 35] due to the costly change annotation.

To address these challenges, we introduce the ***Generalizable Scene Change Detection Framework (GeSCF)***, the first zero-shot scene change detection method that operates robustly regardless of the temporal input order and environmental conditions. Our approach builds upon the Segment Anything Model (SAM) [23], a pioneering vision foundation model for image segmentation. While SAM excels at segmenting *anything* within a single image, directing SAM to identify and segment *changes* between two input images presents a significant challenge. This difficulty arises because SAM is designed for promptable interactive segmentation, relying on user-guided prompts and single-image inputs, whereas scene change detection necessitates processing image pairs to identify changes. To bridge this gap, we propose two innovations: the Initial Pseudo-mask Generation and the Geometric-Semantic Mask Matching. By analyzing the localized semantics of the SAM's feature space, we effectively binarize pixel-level change candidates with zero additional cost. Additionally, we leverage the geometric properties of SAM's class-agnostic masks and the semantics of mask embeddings to refine the final change masks—incorporating object-level information as well.

Furthermore, we introduce the Generalizable Scene Change Detection (GeSCD) benchmark by developing novel metrics and an evaluation protocol to foster SCD research in generalizability; most conventional SCD methods have focused on individual benchmarks separately rather than the generalizability to unseen domains and temporal consistency. We believe our GeSCD can meet the growing needs for developing *anything* SCD in the era of diverse *anything* models [23, 51, 56, 58, 61, 66] with strong zero-shot capabilities. Concretely, our GeSCD performs extensive cross-domain evaluations to rigorously test the generalizability across diverse environments and assesses temporal consistency quantitatively. Our dual-focused evaluation strategy not only ensures the robustness and reliability of a method but also sets a new benchmark in the SCD field. In the process of designing GeSCD, we collect the ChangeVPR dataset, which comprises carefully annotated images sourced from three prominent Visual Place Recognition (VPR) datasets. This comprehensive dataset includes urban, suburban, and rural environments under challenging conditions, significantly expanding the traditional scope of SCD domains.

To summarize, our contributions are as follows:

1. **Problem Formulation.** We introduce GeSCD, a novel task formulation in scene change detection. To the best of our knowledge, this is the first comprehensive effort to address the generalization problem and temporal consistency in SCD research.
2. **Model Design.** To tackle the GeSCD task, we propose GeSCF, the first zero-shot scene change detection model. GeSCF exhibits complete temporal consistency and demonstrates strong generalizability over previous SCD models that are tightly coupled to their training datasets—ours resulting in a substantial performance gain in unseen domains.
3. **Benchmark Set up.** We present new evaluation metrics, the ChangeVPR dataset, and an evaluation protocol that effectively measures an SCD model's generalizability. These contributions provide a solid foundation to guide and inspire future research in the field.

2. Related Works

Segment Anything Model. Segment Anything Model (SAM) [23] has set a new standard in image segmentation and made significant strides in various computer vision domains: medical imaging [63], camouflaged object detection [46], salient object detection [27], image restoration [61], image editing [56], and video object tracking [58]. By leveraging geometric prompts—such as points or bounding boxes—SAM showcases exceptional zero-shot transfer capabilities across diverse segmentation tasks and unseen image distributions. While SAM demonstrates impressive abilities, its potential for zero-shot scene change detection remains largely unexplored. In this work, we extend SAM's utility beyond single-image segmentation by introducing a novel, training-free approach that guides SAM to detect changes between a pair of natural images.

Change Detection. In the field of Change Detection (CD), the research falls broadly into three areas based on data characteristics: remote sensing CD, video sequence CD,

and natural scene CD—the focus of our work. Remote sensing CD [5–7, 10, 19, 32, 41] involves detecting surface changes over time using data captured by satellite or aerial platforms, providing a high-altitude perspective of phenomena such as urbanization, deforestation, and disaster damage. Additionally, video sequence CD focuses on segmenting frames into foreground and background regions, usually corresponding to moving objects [2, 28]. In contrast to these CDs, natural scene CD aims to detect localized changes from a ground-level perspective such as movement of vehicles [3], pedestrians [44], the appearance and disappearance of objects [3, 42, 44], and significant background changes like the construction or demolition of buildings [42, 44]. Moreover, the task inherently involves misaligned and noisy images due to the nature of data acquisition, as images are often captured by cameras mounted on moving vehicles or robots [26]. Overall, our work concentrates natural scene CD [3, 11, 25, 35, 43, 44, 48, 50]; for the remainder of this paper, we will refer to natural Scene CD as SCD.

Scene Change Detection (SCD). In exsiting SCD benchmarks, most methods are supervised [3, 11, 35, 43, 44, 48, 50] or semi-supervised [25], heavily optimized and evaluated on specific training datasets—leading to low generalizability. Although several works have proposed self-supervised pre-training strategies [39] or leveraged temporal symmetry [50], they still exhibit significant performance gaps when deployed to unseen data. Moreover, the symmetric structure relies on a specific prior knowledge of the domain, rendering it impractical for unknown domains without the proper inductive bias. In contrast, our GeSCF stands as a unified, training-free framework displaying robust performance on unseen data while preserving symmetric architecture for all settings.

Segment Anything with Change Detection. Previous works have primarily leveraged SAM in the remote sensing CD with the Parameter-Efficient Fine-Tuning (PEFT) strategy [57]. For instance, several works leveraged SAM variants [62, 64] with learnable adaptors [57], fine-tuning adaptor networks and change decoders tailored to specific datasets [13, 29]. In contrast to these approaches, our method is the first solid SAM-integrated framework in SCD, leveraging SAM's internal byproducts to effectively binarize change candidates without any guidance and learnable parameters. Moreover, we further exploit valuable prior information [1, 12, 18, 56, 61] from SAM's class-agnostic masks—enabling robust zero-shot SCD across a broad range of domains for the first time.

3. GeSCF

3.1. Motivation and Overview

Despite the abundance of web-scale data [8, 33, 36] and the advent of various zero-shot generalizable models [22, 23, 60], current SCD still suffers from the curse of datasets due to the costly nature of change annotation [65]. Therefore, our research motivation arises from ***how SCD can benefit from recent web-scale trained models*** like SAM. By addressing this question, we aim to overcome the long-standing obstacle of creating a generalizable SCD model—culminating in our proposed GeSCF model.

Fig. 2 gives an overview of the GeSCF pipeline. GeSCF handles the technical gap between SAM designed for promptable interactive segmentation with single-image inputs and SCD for identifying changes with image pairs through two key stages: Initial Pseudo-mask Generation and Geometric-Semantic Mask Matching. First, we intercept and correlate feature facets—one of query, key, and value—from the image encoder to obtain rich, multi-head similarity maps; then, we transform these similarity maps into a binary pseudo-mask by adaptively thresholding the low-similarity pixels using a *skewness*-based algorithm. Finally, we elaborate the pseudo-mask by leveraging the geometric properties of SAM's class-agnostic masks; then, we further validate these masks by comparing the semantic similarities of the corresponding mask embeddings between bi-temporal images, ensuring that the detected changes are meaningful and contextually accurate.

3.2. Preliminary

Since our GeSCF exploits a set of image features intercepted from the SAM image encoder, we first recap how such features are obtained.

Feature Facets. The SAM image encoder employs a Vision Transformer (ViT) architecture [15], consisting of multi-head self-attention layers and multi-layer perceptrons within each ViT block [15, 49]. In the multi-head self-attention layer of the l-th ViT block, the query, key, and value facets are denoted by $\mathbf{QKV}_l \in \mathbb{R}^{3\times N\times H\times W\times C}$, where N, H, W, and C represent the number of heads, height, width, and channel dimensions of facets, respectively.

Image Embedding and Mask Embedding. Similarly, we extract the image embedding $\mathbf{E}_l \in \mathbb{R}^{H\times W\times C}$ from the final multi-layer perceptron layer of the l-th ViT block. Furthermore, given the image embedding $\mathbf{E}_l$ and an arbitrary binary mask $\overline{\mathbf{m}}$, we compute the mask embedding $\mathcal{M}_l$ by averaging image embedding $\mathbf{E}_l$ over all spatial positions where the binary mask $\overline{\mathbf{m}}$ is non-zero, obtaining a single vector representation of the masked image.

3.3. Initial Pseudo-mask Generation

As demonstrated in [23], SAM preserves semantic similarities among mask embeddings within the same natural scene. Furthermore, as observed in prior studies [4, 9], attention maps can capture semantically meaningful objects in images. Building on these foundational insights, we expand

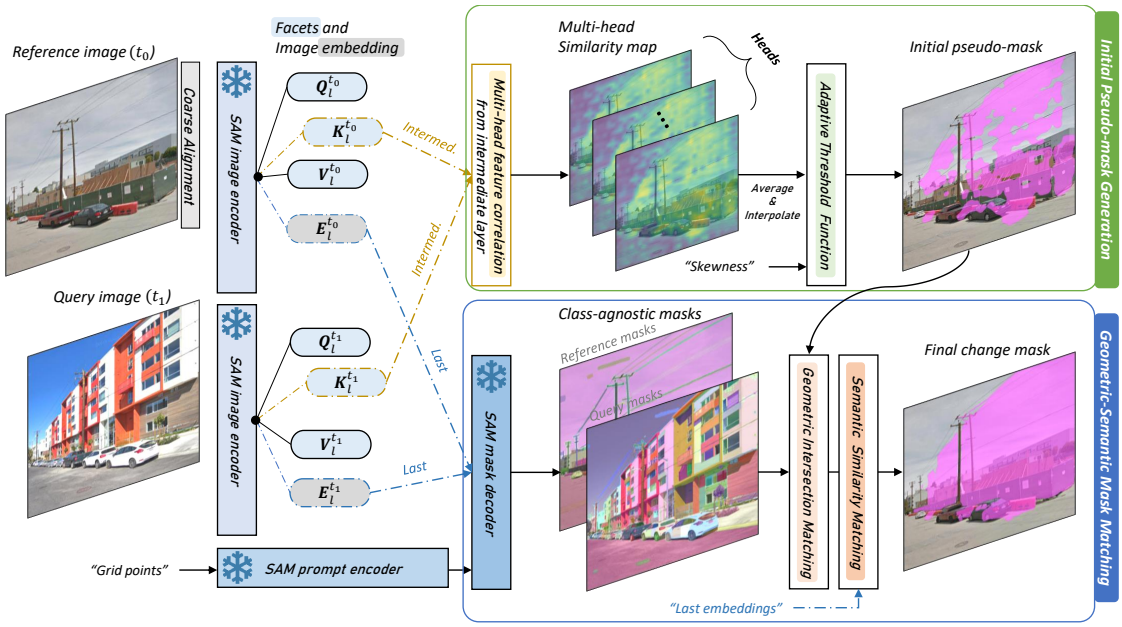

Figure 2. **Illustration of the proposed GeSCF pipeline.** The GeSCF pipeline consists of two major steps: (1) initial pseudo-mask generation and (2) geometric-semantic mask matching. First, we intercept facet features from the SAM image encoder and correlate them to obtain multi-head similarity maps, which are then converted into pseudo-masks using an adaptive threshold function. Next, SAM's class-agnostic masks and last image embeddings are utilized to refine these pseudo-masks based on both geometric and semantic information.

SAM's feature space utilization by extending its application to bi-temporal images and leveraging multi-head feature facets from various layers rather than relying solely on single-image embeddings [23].

Multi-head Feature Correlation. Given a bi-temporal RGB image pair, we first estimate a coarse transformation [24, 26, 47] relating the bi-temporal images. Then, we intercept the feature facets $\mathbf{F}_{l,n}^{t_0}$, $\mathbf{F}_{l,n}^{t_1}$ (one of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$) for the n-th head from the l-th ViT block and compute multi-head feature correlation as follows:

$$\overline{\mathbf{S}}_{l,n}^{t_0 \leftrightarrow t_1} = \mathbf{F}_{l,n}^{t_0} {}^{\top} \mathbf{F}_{l,n}^{t_1}, \tag{1}$$

$$\mathbf{S}_{l}^{t_0 \leftrightarrow t_1} = \xi \left(\sum_{n=1}^{N} \overline{\mathbf{S}}_{l,n}^{t_0 \leftrightarrow t_1} \right), \tag{2}$$

where $\xi(\cdot)$ performs spatial reshaping and bilinear interpolation to the input image size. The correlation in Eq. (1) is a commutative operation that generates identical similarity maps $\overline{\mathbf{S}}_{l,n}$ even with the reversed order. Subsequently, the similarity maps are averaged over the head dimension—leveraging the semantic diversity across the N heads.

Facet and Layer Selection. As shown in Fig. 3 , the similarity map of SAM's facets effectively highlights semantic changes while remaining relatively unaffected by visual variations, such as seasonal or illumination differences. Concretely, our feature selection strategy is guided by the following principles: (a) semantic changes should appear prominently in the similarity map compared to other visual variations; (b) the similarity values in changed regions should clearly contrast with surrounding areas; and (c) artifacts in unchanged regions should be minimized or appear faint. We empirically observe that all facets satisfy principle (a); however, principles (b) and (c) are particularly noticeable when using the key (or query) facet over the value facet. Moreover, we find that these principles hold more strongly in intermediate layers than in the initial or last layers. Consequently, we leverage the similarity map of key facets from intermediate layers as inputs to the subsequent steps in the pseudo-mask generation. For the detailed quantitative analysis, please refer to the supplementary material.

Adaptive Threshold Function. To create a binary mask from an arbitrary similarity map, the most straightforward approach is to apply a fixed threshold (e.g., $p = 0.5$) [17], where similarity scores below this threshold are classified as changes. However, this fixed threshold approach is inherently limited, as it fails to account for the relative nature of "*change*" within each similarity map. For instance, a similarity score of 0.7 may signify a change if all other values are close to one, whereas it may not indicate a change if most values are near zero. Consequently, the perception of "*change*" is context-dependent, necessitating an adaptive thresholding method that accounts for the relative distribution of similarity scores within each map.

Therefore, a critical factor to binarize the similarity map

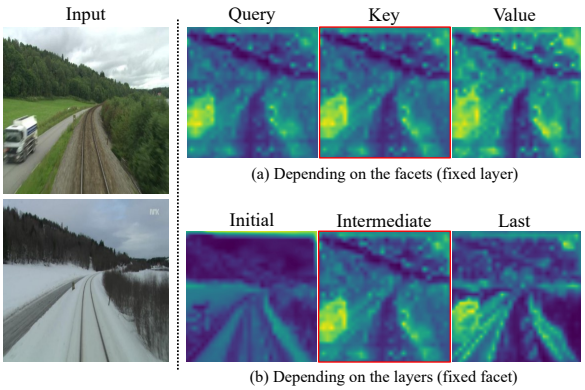

Figure 3. **Visualization of the similarity map depending on the facets and the layers.** We use key facets from the intermediate layer of the SAM ViT image encoder, which is highlighted with a red bounding box.

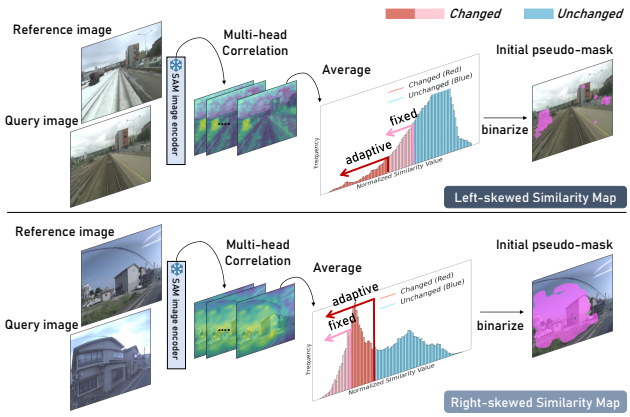

Figure 4. **Illustration of the adaptive thresholding process.** We dynamically adjust the threshold based on the skewness of the distribution to generate the initial pseudo-masks.

is the skewness (γ) [38] of the similarity distribution (see Fig. 4). For right-skewed distributions, where the majority of pixels display lower similarity scores with a long tail of higher scores, a lower threshold is required to capture a larger portion of the distribution as changes. Conversely, for left-skewed distributions, where most scores are high with a small tail of lower values, a higher threshold is needed to avoid false positives. To this end, we propose an adaptive threshold function for dynamic adjustment based on the skewness of the distribution—enabling a more accurate and context-sensitive method for pseudo-mask generation, as detailed in the following formulation:

$$\mathbf{F}(\gamma) = b_\gamma + c \cdot \text{sign}(\gamma) \cdot s_\gamma \cdot \gamma, \tag{3}$$

where b_γ is the baseline threshold, c is a normalization constant, s_γ is the skewness sensitivity factor, γ is the skewness of the distribution, and $\text{sign}(\cdot)$ adjusts the threshold direction based on the skewness. By applying the computed adaptive threshold to the similarity map—normalized using the mean absolute deviation (MAD) [38]—we obtain the initial pseudo-mask essential for the subsequent Geometric-Semantic Mask Matching.

3.4. Geometric-Semantic Mask Matching

Building upon the initial pseudo-masks, we elevate our focus to detecting object-level changes by using SAM's class-agnostic object proposals. This transition from pixel-wise analysis to object-level consideration enables a more comprehensive and interpretable change mask. Here, we introduce two matching strategies: geometric intersection matching (GIM) and semantic similarity matching (SSM).

Geometric Intersection Matching. The main idea of this strategy is to select SAM masks by evaluating their overlap with the initial pseudo-masks. We calculate the intersection ratio α between each SAM mask and the pseudo-mask, retaining only those masks with $\alpha > \alpha_t$, where α_t is a threshold. Since changes can occur bi-temporally, we apply the process for both t_0 and t_1 images—maintaining commutativity for the proposed GeSCF.

Semantic Similarity Matching. Although GIM provides reasonable masks for potential object-level changes, a set of unchanged regions is included due to noise in the initial pseudo-masks. To address this issue, we perform semantic verification of the overlapping regions by calculating the cosine similarity between the mask embeddings of the bi-temporal images. Specifically, for each overlapping mask $\overline{m}_o$, we extract the corresponding mask embeddings $\mathcal{M}_{l,o}^{t_0}$ and $\mathcal{M}_{l,o}^{t_1}$ from the l-th image embeddings $\mathbf{E}_l$ at times t_0 and t_1, respectively. We then compute a change confidence score using the cosine similarity $c(\mathcal{M}_{l,o}^{t_0}, \mathcal{M}_{l,o}^{t_1})$ which allows further refinement of the masks selected by GIM and generates the final change mask $\mathbf{Y}_{\text{pred}}$. Through a layer-wise analysis, we empirically observe that semantic differences are more pronounced in the last layer compared to the initial and intermediate layers, thereby utilizing final image embeddings for our SSM process[1].

4. GeSCD

Since the introduction of the first SCD benchmark [42] in 2015, the generalizability to unseen domains and temporal consistency of predictions have not been consistently or comprehensively addressed in the SCD field. Most traditional SCD approaches train and evaluate models on individual datasets separately [3, 11, 25, 35, 43, 44, 48, 50]. Only [39] performed cross-domain evaluation, however, it still suffers from domain gaps and limited training datasets. Moreover, in contrast to the remote sensing CD [65], the

[1]Please refer to the supplementary material for the analysis.

temporal consistency is often overlooked in model design [3, 11, 25, 35, 43, 44, 48] or training objectives [3, 11, 25, 35, 43, 44, 48, 50] in the SCD field. The temporal symmetry proposed in [50] is unsuitable for real-world applications since it assumes perfect inductive bias of the application domain, which is not feasible in real-world scenarios. Further, as we are in the age of diverse *anything* models [23, 51, 56, 58, 61, 66] with strong zero-shot prediction and generalizability, the necessity of *anything* SCD model that can adapt to various change scenarios has become increasingly important in the research community.

Based on these requirements, we propose *GeSCD*—a novel task approach that addresses the generalizability of broader scenarios and temporal consistency of SCD models. By pioneering new metrics, an evaluation dataset, and a comprehensive evaluation protocol, our approach fulfills the critical need for SCD research that is genuinely applicable and effective across diverse settings.

Metrics. To evaluate the environmental and temporal robustness of the methods simultaneously, we report conventional metrics (e.g., Intersection over Union and F1-score) for both temporal directions in contrast to previous methods that typically report performance for a single temporal direction [3, 11, 25, 35, 43, 44, 48, 50]. Furthermore, we propose the Temporal Consistency (TC) metric by measuring the union intersection between $t0 \rightarrow t1$ and $t1 \rightarrow t0$ predictions as follows:

$$\text{Temporal Consistency (TC)} = \frac{\mathbf{Y}_{\text{pred}}^{t0\rightarrow t1} \cap \mathbf{Y}_{\text{pred}}^{t1\rightarrow t0}}{\mathbf{Y}_{\text{pred}}^{t0\rightarrow t1} \cup \mathbf{Y}_{\text{pred}}^{t1\rightarrow t0}}, \quad (4)$$

where the proposed TC score indicates how much the SCD algorithms can generate consistent change masks in bi-directional orders.

Evaluation Datasets. First, we consider three standard SCD datasets with different characteristics: VL-CMU-CD [3], TSUNAMI [42], and ChangeSim [34]. These datasets represent urban environments in the USA, disaster-impacted urban areas in Japan, and industrial indoor settings within simulation environments, respectively. Furthermore, for the quantitative evaluation on broader unseen domains, we create a new dataset named the *ChangeVPR*. The ChangeVPR comprises 529 image pairs collected from the SF-XL (urban, U) [8], St Lucia (suburban, S) [30], and Nordland (rural, R) [45] datasets, which are widely used in the Visual Place Recognition (VPR) research. We carefully sampled image pairs from each dataset to reflect various SCD challenges such as weather conditions and seasonal changes, and hand-labeled the ground-truth change masks for each image pair. We employ ChangeVPR only for the evaluation to assess unseen domain performance. Please refer to the supplementary material for more details.

Evaluation Protocol. We perform extensive cross-domain evaluation for SCD models. First, we train each model on the three standard SCD datasets; the training stage results in three distinctive models for each method. Then, we assess each of the three distinctive models in two stages. First, we evaluate the three models on the three conventional SCD datasets; this process totals nine assessments. If the method of interest does not involve training, we evaluate the model on the three datasets (three assessments). Next, we evaluate the models on the proposed ChangeVPR dataset. Since we use ChangeVPR only for evaluation, this stage contains three assessments per model (due to three splits of ChangeVPR). As a result, the proposed protocol estimates performance on both seen and unseen domains—ensuring a thorough evaluation of SCD performance in various environmental scenarios with arbitrary real-world changes.

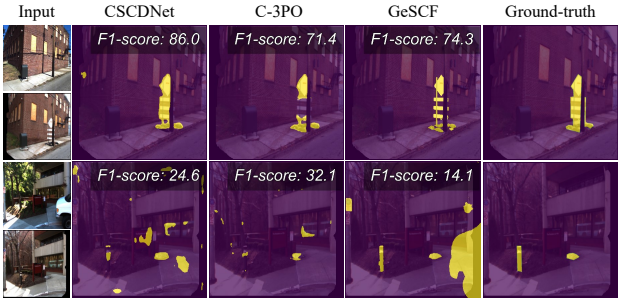

Figure 5. **Qualitative results on VL-CMU-CD dataset with F1-scores.** Our model generates reasonable change masks and does not display annotation bias.

5. Experiments

5.1. Comparative Studies

We compare our method with four state-of-the-art SCD models: CSCDNet [44], CDResNet [44], DR-TANet [11], and C-3PO [50]. For C-3PO, we adopt the (I+D) structure for the VL-CMU-CD and the (I+A+D+E) structure for the TSUNAMI and ChangeSim datasets, following the configurations proposed in the paper. To verify the temporal consistency, we additionally train all models with bi-temporal objective [65]; the performance of these bi-temporally trained models has not been reported in the literature.

Quantitative Comparison. Tab. 1 displays the quantitative comparison results on the standard SCD datasets. The results attest that GeSCF outperforms the baselines with a large margin in every unseen domain (off-diagonal results) and on-par performance on seen domains (diagonal results), along with complete TC score (TC=1.0) for all settings—demonstrating its superior generalizability over diverse environments and temporal conditions. The current best performance on the VL-CMU-CD dataset drops sharply from 77.6% to 4.6% when the temporal order is reversed and to 8.0% when deployed to the unseen

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	
Uni-temporal Training											
CSCDNet	VL-CMU-CD	77.4	4.5	0.02	5.6	19.4	0.02	25.5	13.6	0.09	24.3
CDResNet	VL-CMU-CD	74.7	2.1	0.01	2.3	4.6	0.0	25.4	13.2	0.12	20.4
DR-TANet	VL-CMU-CD	74.2	1.6	0.01	2.6	6.7	0.0	27.5	18.0	0.19	21.8
C-3PO	VL-CMU-CD	**77.6**	4.6	0.02	8.0	34.3	0.02	30.4	21.0	0.16	29.3
CSCDNet	TSUNAMI	17.0	22.1	0.40	83.6	60.3	0.46	27.8	30.5	0.29	40.2
CDResNet	TSUNAMI	13.2	18.4	0.42	82.8	51.5	0.38	26.3	29.0	0.31	36.9
DR-TANet	TSUNAMI	14.8	17.3	0.53	82.0	57.6	0.44	24.4	26.8	0.35	37.2
C-3PO	TSUNAMI	27.0	27.0	1.0	**84.2**	**84.2**	1.0	34.3	34.3	1.0	48.5
CSCDNet	ChangeSim	20.1	3.5	0.04	2.5	6.6	0.01	43.1	18.9	0.31	15.8
CDResNet	ChangeSim	17.5	3.1	0.02	4.6	9.9	0.01	41.3	18.2	0.18	15.8
DR-TANet	ChangeSim	20.3	4.9	0.05	5.0	8.3	0.01	40.3	18.0	0.18	16.1
C-3PO	ChangeSim	1.6	1.6	1.0	0.3	0.3	1.0	15.8	15.8	1.0	5.9
Bi-temporal Training											
CSCDNet	VL-CMU-CD	64.2	64.8	0.61	8.5	7.5	0.12	24.0	23.9	0.42	32.2
CDResNet	VL-CMU-CD	60.6	61.7	0.58	5.0	3.8	0.18	17.7	14.1	0.29	27.2
DR-TANet	VL-CMU-CD	45.6	61.9	0.40	3.5	1.2	0.02	18.0	23.0	0.26	25.5
C-3PO	VL-CMU-CD	58.4	63.5	0.63	2.5	1.4	0.07	21.2	23.8	0.51	28.5
CSCDNet	TSUNAMI	20.9	21.4	0.56	82.4	82.8	0.82	30.0	29.4	0.46	44.5
CDResNet	TSUNAMI	17.6	17.1	0.63	81.7	81.8	0.81	28.9	29.0	0.48	42.7
DR-TANet	TSUNAMI	18.1	17.4	0.61	81.1	82.0	0.76	28.6	28.7	0.46	42.7
C-3PO	TSUNAMI	27.5	27.5	1.0	84.1	84.1	1.0	33.7	33.7	1.0	48.4
CSCDNet	ChangeSim	0.2	0.1	0.86	4.7	1.0	0.69	19.5	18.6	0.49	7.4
CDResNet	ChangeSim	4.0	2.0	0.75	1.1	0.5	0.53	19.8	20.7	0.48	8.0
DR-TANet	ChangeSim	7.5	3.7	0.48	2.6	1.8	0.13	27.9	21.4	0.32	10.8
C-3PO	ChangeSim	0.4	0.4	1.0	1.2	1.2	1.0	18.0	18.0	1.0	6.5
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 (F1-score) on standard SCD datasets.**

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	
Uni-temporal Training											
CSCDNet	VL-CMU-CD	28.6	23.1	0.07	11.7	15.3	0.27	14.3	5.8	0.33	16.5
CDResNet	VL-CMU-CD	22.9	18.9	0.08	13.5	18.0	0.18	9.3	6.7	0.39	14.9
DR-TANet	VL-CMU-CD	22.5	19.6	0.05	15.5	20.8	0.19	11.8	8.2	0.33	16.4
C-3PO	VL-CMU-CD	38.7	32.8	0.07	21.7	25.7	0.14	16.2	14.2	0.28	24.9
CSCDNet	TSUNAMI	38.9	39.6	0.55	15.3	15.4	0.55	18.8	20.8	0.35	24.8
CDResNet	TSUNAMI	36.6	37.4	0.59	13.7	14.0	0.66	16.4	19.9	0.33	23.0
DR-TANet	TSUNAMI	35.2	36.1	0.62	13.4	13.4	0.76	17.1	16.7	0.44	22.0
C-3PO	TSUNAMI	50.0	50.0	1.0	28.4	28.4	1.0	23.2	23.2	1.0	33.9
CSCDNet	ChangeSim	32.8	28.1	0.08	19.3	19.6	0.09	6.7	7.8	0.29	19.1
CDResNet	ChangeSim	26.6	23.4	0.07	15.2	14.2	0.15	7.1	6.4	0.3	15.5
DR-TANet	ChangeSim	27.5	25.0	0.08	15.9	15.9	0.15	10.6	13.8	0.21	18.1
C-3PO	ChangeSim	4.0	4.0	1.0	3.1	3.1	1.0	0.8	0.8	1.0	2.6
Bi-temporal Training											
CSCDNet	VL-CMU-CD	24.1	25.4	0.40	5.6	5.1	0.68	2.2	3.0	0.71	10.9
CDResNet	VL-CMU-CD	23.1	23.3	0.37	7.0	6.7	0.69	1.3	1.2	0.80	10.4
DR-TANet	VL-CMU-CD	17.0	17.3	0.27	9.0	4.1	0.56	2.2	2.4	0.66	8.7
C-3PO	VL-CMU-CD	23.3	23.2	0.5	7.9	6.9	0.77	3.6	4.6	0.79	11.6
CSCDNet	TSUNAMI	42.8	42.9	0.67	17.1	17.1	0.64	20.4	21.0	0.57	26.9
CDResNet	TSUNAMI	40.0	40.1	0.71	13.9	13.8	0.76	19.8	19.5	0.58	24.5
DR-TANet	TSUNAMI	38.6	38.7	0.71	13.9	14.0	0.78	20.6	17.5	0.55	23.9
C-3PO	TSUNAMI	50.0	50.0	1.0	28.0	28.0	1.0	24.3	24.3	1.0	34.1
CSCDNet	ChangeSim	13.3	12.1	0.31	0.9	1.1	0.82	0.7	1.0	0.84	4.9
CDResNet	ChangeSim	11.4	9.6	0.26	2.2	1.5	0.76	1.3	0.7	0.92	4.5
DR-TANet	ChangeSim	12.8	12.5	0.12	8.3	6.1	0.30	7.5	6.3	0.40	8.9
C-3PO	ChangeSim	6.0	6.0	1.0	2.1	2.1	1.0	1.0	1.0	1.0	3.0
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 (F1-score) on the ChangeVPR dataset.**

TSUNAMI dataset. Remarkably, our GeSCF even outperforms the supervised baselines for the ChangeSim dataset by an F1-score of 54.8% in a zero-shot manner. Overall, the average performance of our method on standard SCD datasets surpasses that of the second-best by a substantial margin (+19.2%). Furthermore, the majority of baselines do not display a complete TC score which is crucial for the system's reliability. The temporal consistency of C-3PO does not hold in VL-CMU-CD settings as shown in its low TC scores. We find that simple bi-temporal training does not guarantee complete temporal consistency and unseen domain performance, even worsening the uni-

temporal performances—indicating different data characteristics from [65] and the significance of our commutative architecture. Tab. 2 presents the quantitative comparison results on the proposed ChangeVPR dataset. The results consistently show that GeSCF outperforms the baselines across all unseen domains with a $+30.0\%$ margin (*nearly doubling* the prior art performance), extending beyond the conventional urban-/synthetic-only SCD scope. Notably, GeSCF demonstrates exceptional performance in Nordland [45] split, a challenging dataset with severe seasonal variation (summer-winter).

Qualitative Comparison. As illustrated in Fig. 1, GeSCF adeptly adapts to various unseen images, accurately delineating changed objects—significantly outperforming the current state-of-the-art model. Moreover, as a zero-shot framework, GeSCF does not learn annotation biases in a dataset (see Fig. 5). For instance, GeSCF generates a more accurate and interpretable mask for the object and identifies unannotated changes—displaying resilience to erroneous GTs. These results suggest that GeSCF's outputs are more reasonable than baselines, even though its quantitative results are lower for some samples in the in-domain settings.

5.2. Ablation Studies

To understand the effectiveness of our contributions, we conducted comprehensive ablation studies across three standard SCD datasets and ChangeVPR (see Tab. 3). We utilized baselines trained on TSUNAMI [42], which demonstrated superior average performance over others.

Fine-tuning Strategy. We fine-tuned various SCD-based adapter networks on the frozen SAM image encoder, involving correlation layers [14], CHVA [11], and the feature merging module [50]. Our experiments reveal that fine-tuning the adaptor model on relatively small SCD datasets significantly undermines its zero-shot capabilities, resulting in a substantial performance drop compared to our proposed GeSCF. Notably, our initial pseudo-masks alone achieve an F1-score that surpasses the fine-tuned variants.

Effectiveness of the Proposed Modules. We validated the effectiveness of each proposed module, including the Initial Pseudo-mask Generation with adaptive threshold function, Geometric-Semantic Mask Matching with geometric intersection matching (GIM), and semantic similarity matching (SSM). The findings suggest that each introduced module positively influences GeSCF's overall performance, with the initial pseudo-mask generation (adaptive) and GIM demonstrating significant enhancement.

5.3. Beyond Scene Change Detection

Although our research primarily focuses on natural scene CDs and each CD field tends to focus on its own [25], we discovered that GeSCF can also be applied to zero-shot remote sensing CD. Following the standard evaluation metrics of the remote sensing domain, we evaluated our model against several unsupervised methods [6, 16, 40, 52, 53, 55] on the SECOND [59] benchmark (see Tab. 4). Remarkably, our GeSCF outperforms other baselines with 48.0% in F1-score and a high precision of 52.2%—underscoring its zero-shot potential for different CD domains.

Method	Metric	
	F1-score	mIoU
Uni-temporal Training		
SAM [23] + Corr. layers [14]	28.7	19.7
SAM [23] + CHVA [11]	30.4	21.1
SAM [23] + MTF [50] + MSF [50]	40.1	30.2
Bi-temporal Training		
SAM [23] + Corr. layers [14]	30.0	21.1
SAM [23] + CHVA [11]	31.0	22.0
SAM [23] + MTF [50] + MSF [50]	40.3	30.3
GeSCF (Ours)		
Pseudo. only (fixed)	43.0	26.3
Pseudo. only (adaptive)	53.9	34.9
Pseudo. + GIM	65.0	47.6
Pseudo. + GIM + SSM (**ALL**)	**65.9**	**48.8**

Table 3. **Ablation study results.**

Method	Metric		
	F1-score	Precision	Recall
ISFA [52]	32.9	29.8	36.8
DSFA [16]	33.0	24.2	**51.9**
DCAE [6]	33.4	35.7	31.4
OBCD [55]	34.3	29.6	40.7
KPCA-MNet [53]	36.7	29.5	48.5
DCVA [40]	36.8	29.6	48.7
GeSCF (Ours)	**48.0**	**52.2**	44.4

Table 4. **Quantitative comparison with unsupervised remote sensing CD methods on the SECOND (`test`) benchmark.**

6. Conclusion

In this study, we attempted to address the generalization issue and temporal bias of scene change detection using the Segment Anything Model for the first time to the best of our knowledge; we define GeSCD along with the novel metrics, evaluation dataset (ChangeVPR), and an evaluation protocol to effectively assess the generalizability and temporal consistency of SCD models. Furthermore, we proposed GeSCF, a generalizable zero-shot approach for the SCD task using the localized semantics of the SAM that does not require costly SCD labeling. Our extensive experiments demonstrated that the proposed GeSCF significantly outperforms existing SCD models on standard SCD datasets and overwhelms them on the ChangeVPR dataset while achieving complete temporal consistency. We expect our methods can serve as a solid step towards robust and generalizable Scene Change Detection research.

Acknowledgement. This research was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.RS-2022-II220926, Development of Self-directed Visual Intelligence Technology Based on Problem Hypothesis and Self-supervised Methods; No.RS-2022-II220907, Development of AI Bots Collaboration Platform and Self-organizing AI; No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)), National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1C1C1009989), and 'Project for Science and Technology Opens the Future of the Region' program through INNOPOLIS FOUNDATION founded by Ministry of Science and ICT (No.2022-DD-UP-0312-02-101).

References

[1] Mohsen Ahmadi, Ahmad Gholizadeh Lonbar, Abbas Sharifi, Ali Tarlani Beris, Mohammad Javad Nouri, and Amir Sharifzadeh Javidi. Application of segment anything model for civil infrastructure defect assessment. *ArXiv*, abs/2304.12600, 2023. 3

[2] Thangarajah Akilan, Q.M.Jonathan Wu, Amin Safaei, Jie Huo, and Yimin Yang. A 3d cnn-lstm-based image-to-image foreground segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21:959–971, 2020. 3

[3] 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. 2, 3, 5, 6

[4] Shirzad Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ArXiv*, abs/2112.05814, 2021. 3

[5] W. G. C. Bandara and Vishal M. Patel. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *ArXiv*, abs/2204.08454, 2022. 3

[6] Luca Bergamasco, Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised change detection using convolutional-autoencoder multiresolution features. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 8

[7] Maximilian Bernhard, Niklas Strauß, and Matthias Schubert. Mapformer: Boosting change detection by using pre-change information. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16791–16800, 2023. 3

[8] 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. 3, 6

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 3

[10] Chao Chen, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bo Wang. Saras-net: Scale and relation aware siamese network for change detection. *ArXiv*, abs/2212.01287, 2022. 3

[11] 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, 3, 5, 6, 8

[12] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-Lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *ArXiv*, abs/2305.05803, 2023. 3

[13] 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. 3

[14] 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. 8

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 3

[16] Bo Du, Lixiang Ru, Chen Wu, and Liangpei Zhang. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57:9976–9992, 2018. 8

[17] Yukuko Furukawa, Kumiko Suzuki, Ryuhei Hamaguchi, Masaki Onishi, and Ken Sakurada. Self-supervised simultaneous alignment and change detection. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6025–6031, 2020. 4

[18] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (sam). *ArXiv*, abs/2304.07764, 2023. 3

[19] Fan Hao, Zongfang Ma, Hong peng Tian, Hao Wang, and Di Wu. Semi-supervised label propagation for multi-source remote sensing image change detection. *Comput. Geosci.*, 170:105249, 2023. 3

[20] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael W. Spratling, and Yanfeng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[21] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *Found. Trends Comput. Graph. Vis.*, 12:1–308, 2017. 2

[22] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2023. 3

[23] 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 In-*

ternational Conference on Computer Vision (ICCV), pages 3992–4003, 2023. 2, 3, 4, 6, 8

[24] Daniel Koguciuk, E. Arani, and Bahram Zonooz. Perceptual loss for robust unsupervised homography estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4269–4278, 2021. 4

[25] 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. 2, 3, 5, 6, 8

[26] Dong Li, Lineng Chen, Chengzhong Xu, and Hui Kong. Umad: University of macau anomaly detection benchmark dataset. *ArXiv*, abs/2408.12527, 2024. 3, 4

[27] Jun Ma, Yuting He, Feifei Li, Li-Jun Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15, 2023. 2

[28] 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. 3

[29] 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. 3

[30] 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. 6

[31] Ulrich Nehmzow, Jasna Kuljis, Ray J. Paul, and Peter Thomas. Mobile robotics: A practical introduction: History, design, analysis and examples. 2000. 2

[32] Hyeoncheol Noh, Jingi Ju, Min seok Seo, Jong chan Park, and Dong geol Choi. Unsupervised change detection based on image reconstruction loss. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1351–1360, 2022. 3

[33] 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. 3

[34] 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. 2, 6

[35] 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, 5, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 3

[37] Richard J. Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14:294–307, 2005. 2

[38] Kandethody M. Ramachandran, Ramachandran Tsokos, and Chris P. Tsokos. Mathematical statistics with applications. 2009. 5

[39] Vijaya Raghavan T. Ramkumar, E. Arani, and Bahram Zonooz. Differencing based self-supervised pretraining for scene change detection. In *CoLLAs*, 2022. 3, 5

[40] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57:3677–3693, 2019. 8

[41] Sudipan Saha, M. Shahzad, Patrick Ebel, and Xiao Xiang Zhu. Supervised change detection using prechange optical-sar and postchange sar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 8170–8178, 2022. 3

[42] 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. 2, 3, 5, 6, 8

[43] Ken Sakurada, Weimin Wang, Nobuo Kawaguchi, and Ryosuke Nakamura. Dense optical flow based change detection network robust to difference of camera viewpoints. *ArXiv*, abs/1712.02941, 2017. 3, 5, 6

[44] 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, 3, 5, 6

[45] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. 2013. 6, 8

[46] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection. *ArXiv*, abs/2304.04709, 2023. 2

[47] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10247–10266, 2021. 4

[48] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P. Balamuralidhar. Changenet: A deep learning architecture for visual change detection. In *ECCV Workshops*, 2018. 2, 3, 5, 6

[49] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and

Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 3

[50] Guo-Hua Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation. *Pattern Recognition*, 138:109384, 2023. 2, 3, 5, 6, 8

[51] Zhenyu Wang, Yali Li, Xi Chen, Ser Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11433–11443, 2023. 2, 6

[52] Chen Wu, Bo Du, and Liangpei Zhang. Slow feature analysis for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52:2858–2874, 2014. 8

[53] Chen Wu, Hongruixuan Chen, Bo Du, and Liangpei Zhang. Unsupervised change detection in multitemporal vhr images based on deep kernel pca convolutional mapping network. *IEEE Transactions on Cybernetics*, 52:12084–12098, 2019. 8

[54] Guangming Wu, Yinqiang Zheng, Zhiling Guo, Zekun Cai, Xiaodan Shi, Xin Ding, Yifei Huang, Yimin Guo, and Ryosuke Shibasaki. Learn to recover visible color for video surveillance in a day. In *European Conference on Computer Vision (ECCV)*, pages 495–511. Springer, 2020. 2

[55] Pengfeng Xiao, Xue liang Zhang, Dongguang Wang, Minglei Yuan, Xuezhi Feng, and Maggi Kelly. Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. *Isprs Journal of Photogrammetry and Remote Sensing*, 119:402–414, 2016. 8

[56] Defeng Xie, Ruichen Wang, Jiancang Ma, Chen Chen, H. Lu, D. Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing. *ArXiv*, abs/2304.14006, 2023. 2, 3, 6

[57] 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. 3

[58] Jinyu Yang, Mingqi Gao, Zhe Li, Shanghua Gao, Fang Wang, and Fengcai Zheng. Track anything: Segment anything meets videos. *ArXiv*, abs/2304.11968, 2023. 2, 6

[59] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. 8

[60] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024. 3

[61] Tao Yu, Runsen Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *ArXiv*, abs/2304.06790, 2023. 2, 3, 6

[62] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong-Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *ArXiv*, abs/2306.14289, 2023. 3

[63] Yizhe Zhang, Tao Zhou, Peixian Liang, and Da Chen. Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *ISIC/Care-AI/MedAGI/DeCaF@MICCAI*, 2023. 2

[64] Xu Zhao, Wen-Yan Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *ArXiv*, abs/2306.12156, 2023. 3

[65] 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. 3, 5, 6, 8

[66] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *ArXiv*, abs/2304.06718, 2023. 2, 6