Multi-View Pose-Agnostic Change Localization with Zero Labels

Chamuditha Jayanga[1,2]   Jason Lai[3]   Lloyd Windrim[2,4]   Donald Dansereau[2,3]
Niko Suenderhauf[1,2]   Dimity Miller[1,2]

[1]QUT Centre for Robotics   [2]ARIAM Hub*   [3]University of Sydney   [4]Abyss Solutions

Abstract

Autonomous agents often require accurate methods for detecting and localizing changes in their environment, particularly when observations are captured from unconstrained and inconsistent viewpoints. We propose a novel label-free, pose-agnostic change detection method that integrates information from multiple viewpoints to construct a change-aware 3D Gaussian Splatting (3DGS) representation of the scene. With as few as 5 images of the post-change scene, our approach can learn additional change channels in a 3DGS and produce change masks that outperform single-view techniques. Our change-aware 3D scene representation additionally enables the generation of accurate change masks for unseen viewpoints. Experimental results demonstrate state-of-the-art performance in complex multi-object scenes, achieving a 1.7× and 1.6× improvement in Mean Intersection Over Union and F1 score respectively over other baselines. We also contribute a new real-world dataset to benchmark change detection in diverse challenging scenes in the presence of lighting variations. Our code and dataset will be released at github.com/PASLCD.
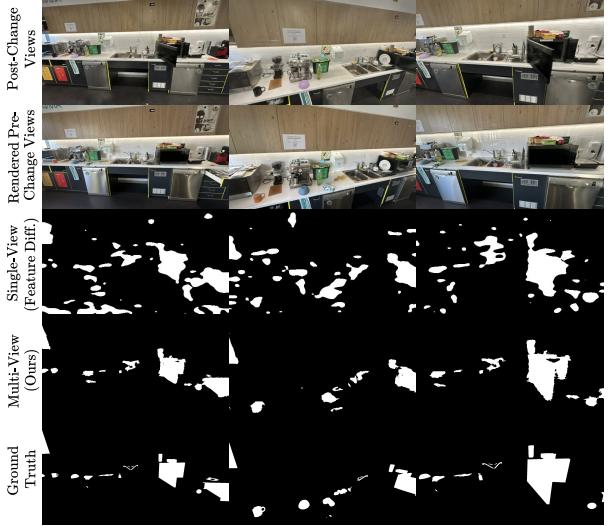
Figure 1. Our multi-view approach to visual change detection (second row from bottom) enforces consistency of the predicted changes across multiple viewpoints by embedding change information in a 3D Gaussian Splatting model of the scene. This effectively suppresses many of the false-positive detections exhibited by current single-view methods (middle row).

1. Introduction

There is increasing effort to develop autonomous agents that assist us with complex tasks, from handling daily chores to performing undesirable work. Capable autonomous agents require the ability to detect and interpret changes in their environment, enabling them to update maps and re-plan tasks, or perform applied tasks such as infrastructure or environment monitoring. Change detection remains a challenging task in 3D scenes, particularly when an agent observes the scene from two sets of views that have no constraint on the poses (i.e. consider a robot that captures images of a scene following a random trajectory at each inspection round).

Many established change detection methods rely on precise alignment between a pre-change image and post-change image to localize the change [2, 6, 7, 11], limiting their applicability to scenes without viewpoint consistency. Some approaches extend to detect changes in images with inconsistent viewpoints [16, 20, 29], but learn change viewpoint invariance by training on image pairs labeled with changes and showcasing viewpoint discrepancy. Supervised learning can have limitations for change detection, including the cost of labeling datasets and significant performance drops under distribution shift (such as environments not present in the dataset) [2, 11, 38, 40]. In this paper, we address the problem of label-free, pose-agnostic change localization, where changes are detected between a pre-change scene and a post-change scene, without labeled data for training or aligned viewpoints for observations between the scenes.

Recent works [15, 47] perform label-free, pose-agnostic change localization by learning a 3D representation of the scene, such as a Neural Radiance Field (NeRF) [25] or 3D Gaussian Splatting (3DGS) [12], and rendering images from the viewpoints of observed images. Changes are detected through feature-level comparisons between the observed and rendered images when using a pre-trained vision model [15, 47]. While this is a feasible approach to pose-agnostic change detection, such approaches struggle to produce accurate change maps in the presence of view-dependent feature-level inconsistencies (e.g. reflections, shadows, unseen regions) common in real-world scenarios.

The main contribution of our paper is a novel *multi-view* change detection method that is both pose-agnostic and label-free. Our multi-view approach integrates change information from multiple viewpoints by constructing a 3DGS model of the environment that encodes not only appearance but also a measure of *change*. With this approach, we can generate a change mask for any viewpoint in the scene, even those that have not yet been observed post-change. By leveraging multiple viewpoints alongside change masks that are both feature and structure-aware, our approach produces multi-view change masks robust to potential view-dependent false changes flagged at the feature level (see Fig. 1).

We make three key claims that are supported by our experiments: First, our approach achieves state-of-the-art performance, particularly in complex multi-object scenes. Second, our change-aware 3D scene representation allows us to generate change predictions for entirely unseen views in the post-change scene, which current methods are unable to do. Third, pre-trained features and the Structural Similarity Index Measure (SSIM) [41] contain complemental change information, and their combination generates robust change masks to learn a change-aware 3DGS.

Our paper additionally contributes a novel dataset encompassing 10 real-world scenes with multiple objects and diverse changes. Our dataset includes variations in lighting, indoor and outdoor settings, and multi-perspective captures, enabling a finer-grained analysis of change detection methods in realistic conditions. To evaluate our approach, we test on an existing single-object pose-agnostic change dataset and our novel dataset, comparing to existing state-of-the-art methods and demonstrating significant improvements in performance.

2. Related Work

2.1. Change Detection with 2D Images

A typical change detection scenario involves a pair of before-and-after RGB images without explicitly considering a 3D scene [2, 3, 18, 30, 32]. These images often adhere to specific conditions: the camera remains fixed, resulting in images related by an identity transform, as in surveillance footage [11, 14]; the scene is planar, as in bird's-eye view or satellite images [6, 7]; or there is minimal viewpoint shift, as in street-view scenes capturing distant buildings or objects [2, 32]. In these cases, models are generally expected to learn to identify changes between image pairs by localizing differences through segmentation [2, 5, 34].

Convolutional Neural Networks (CNNs) have been widely studied for localizing changes [6, 13, 21, 34, 38, 40]. More recently, transformer-based architectures [8] have shown the ability to learn rich, context-aware representations through attention mechanisms, advancing change detection tasks [3, 9, 36, 39, 42]. Foundation models, such as DINOv2 [27], have proven to be robust pre-trained backbones for feature extraction, enhancing change detection across diverse applications [20, 22].

2.2. 2D-3D Scene-level Change Detection

2D to 3D scene-level change detection tackles the challenging and realistic task of identifying changes in 3D scenes, where large viewpoint shifts, severe occlusions, and disocclusions are common. While detecting changes in 3D scenes from sparse 2D RGB images remains underexplored, Sachdeva *et al*. [29] recently introduced a "register-and-difference" approach that leverages frozen embeddings from a pre-trained backbone and feature differences to detect changes. Similarly, Lin *et al*. [20] proposed a cross-attention mechanism built on DINOv2 [27] to address viewpoint inconsistencies in street-view settings. However, both methods rely solely on image-to-image comparisons and do not explicitly construct a 3D representation of the scene.

Related to the field of scene-level change detection is pose-agnostic anomaly detection. Anomaly detection typically leverages unsupervised learning to build a normality model from a set of 2D images, tagging images inconsistent with this model as anomalies during inference [19, 44–46]. Recently, Zhou *et al*. [47] introduced a pose-agnostic anomaly detection dataset consisting of small-scale scenes containing single toy LEGO objects. Closely related to our work, OmniPoseAD [47] and SplatPose [15] explore this dataset to build 3D object representations of a scene containing a faultless object. OmniPoseAD employs NeRFs [24] to model the object, using coarse-to-fine pose estimation with iNeRF [43] to render a matching viewpoint, and generates anomaly scores by comparing multi-scale features from a pre-trained CNN. SplatPose replaces NeRF with 3DGS [12] and directly learns rigid transformations for each Gaussian, bypassing iNeRF. Both methods leverage a 3D scene representation, but only consider anomaly detection on a single per-view image basis – we extend beyond these works by leveraging multiple views and the 3D scene representation to learn more robust multi-view change masks.

2.3. Learning a 3D Representation

Learning a 3D representation of a scene has been used by prior works to enable pose-agnostic, unsupervised change detection [15, 47]. Complex geometries can be represented as continuous implicit fields using coordinate-based neural networks. For example, signed distance fields [28, 37] capture the distance of each point to object surfaces, while occupancy networks [23] indicate whether points lie within an object. Recent advances in high-fidelity scene representations, such as NeRFs [25] and variants [4, 10, 26], model scenes by regressing a 5D plenoptic function [1], outputting view-independent density and view-dependent radiance for photorealistic novel view synthesis.

In contrast to implicit fields, 3DGS [12] provides an explicit scene representation using anisotropic 3D Gaussians, enabling high-quality, real-time novel view synthesis. Each Gaussian is defined by a center position μ and covariance matrix Σ, calculated from a scaling matrix S and rotation matrix R as $\Sigma = RSS^T R^T$. Additionally, an opacity factor α and color component c, modeled with spherical harmonics, are learned to capture view-dependent appearance. To initialize, 3DGS uses Structure-from-Motion (SfM) with COLMAP [35] to estimate camera poses and create a sparse point cloud from multi-view images. Gaussian parameters and color components are then optimized by comparing rendered views with ground truth images using a combination of L_1 loss and a D-SSIM loss term [12].

3. Methodology

An overview of our proposed approach for pose-agnostic, label-free change detection is shown in Fig. 2. We construct a 3D Gaussian Splatting (3DGS) [12] representation for the pre-change (*reference*) scene, allowing us to render pre-change images from novel viewpoints (Sec. 3.2). After collecting images from the post-change (*inference*) scene, we compare to corresponding rendered pre-change images and compute feature and structure-aware change masks (Sec. 3.3). We then learn an *updated* 3DGS for the post-change scene that also embeds Gaussian-specific change channels for reconstructing change masks, leveraging the multiple views from the 3D scene (Sec. 3.4). This change-aware 3DGS can be queried for any pose to generate a multi-view change mask of the scene (Sec. 3.5). We additionally introduce a data augmentation strategy to increase the number of change masks used to learn our change-aware 3DGS (Sec. 3.6).

3.1. Problem Setup

A set of n_{ref} images are collected from a reference scene, $\mathcal{I}_{\text{ref}} = \{I_{\text{ref}}^k\}_{k=1}^{n_{\text{ref}}}$. Changes in this scene then occur, including structural changes (addition, removal, or movement of objects) and surface-level changes (changes to texture or color of objects, drawings on surfaces). "Distractor" or irrelevant visual changes can also occur, such as the changing of lighting, shadows, or reflections in the scene. A set of n_{inf} images are collected from the scene post-change, referred to as the inference scene, $\mathcal{I}_{\text{inf}} = \{I_{\text{inf}}^k\}_{k=1}^{n_{\text{inf}}}$. Our objective is to generate a set of segmentation masks $\mathcal{M} = \{M^k\}_{k=1}^{n_{\text{inf}}}$ for all images in $\mathcal{I}_{\text{inf}}$ that localizes all relevant changes between the reference and inference scenes while disregarding distractor changes.

3.2. Building a 3D Reference Scene Representation

Given the reference scene images $\mathcal{I}_{\text{ref}}$, we utilise COLMAP [35] to perform SfM and obtain camera poses for all images, $\mathcal{P}_{\text{ref}} = \{P_{\text{ref}}^k\}_{k=1}^{n_{\text{ref}}}$. We then use $\mathcal{P}_{\text{ref}}$ and $\mathcal{I}_{\text{ref}}$ to construct a 3DGS representation of the reference scene, $3DGS_{\text{ref}}$, following the pipeline described in [12]. We assume that the number, quality and viewpoints of images in $\mathcal{I}_{\text{ref}}$ is sufficient to build a 3DGS [12] representation.

3.3. Generating Feature and Structure-Aware Change Masks

Given the inference scene images $\mathcal{I}_{\text{inf}}$, we acquire corresponding camera poses $\mathcal{P}_{\text{inf}} = \{P_{\text{inf}}^k\}_{k=1}^{n_{\text{inf}}}$ by registering $\mathcal{I}_{\text{inf}}$ to the same SfM reconstruction built from $\mathcal{I}_{\text{ref}}$ using COLMAP [35]. This ensures $\mathcal{P}_{\text{ref}}$ and $\mathcal{P}_{\text{inf}}$ share a reference frame, assuming that the magnitude of appearance change is not so severe that COLMAP [35] is unable to make the registration (i.e. inference scene is extremely dark).

We then render a new image set, $\mathcal{I}_{\text{ren}}$, from our $3DGS_{\text{ref}}$ with the exact poses from our inference scene $\mathcal{P}_{\text{inf}}$. Comparing images from $\mathcal{I}_{\text{ren}}$ with the corresponding pose-aligned image in $\mathcal{P}_{\text{inf}}$, we can now generate change masks.

Feature-Aware Change Mask: We extract a feature-aware change mask by leveraging a pre-trained visual foundation model $\mathcal{H}$ (specifically DINOv2 [27]). We test $\mathcal{H}$ with $\mathcal{I}_{\text{ren}}$ and $\mathcal{I}_{\text{inf}}$ to produce a dense feature set $\{(f_{\text{ren}}^k, f_{\text{inf}}^k)\}_{k=1}^{n_{\text{inf}}}$ for each pose-aligned image pair. These feature maps are defined by the image height h, width w, patch size s of the foundation model, and embedding dimension d, $f \in \mathbb{R}^{\frac{h}{s} \times \frac{w}{s} \times d}$. We then compute a preliminary feature-aware change mask D^k between f_{ren}^k and f_{inf}^k across the embedding dimension d as follows:

$$D^k = \sum_{j=1}^{d} |f_{\text{ren}}^{k,j} - f_{\text{inf}}^{k,j}| \in \mathbb{R}^{\frac{h}{s} \times \frac{w}{s}}. \qquad (1)$$

We then normalize D^k values to range between 0 and 1 and apply bicubic interpolation to create a feature-aware change mask with the original image dimensions. We create our final feature-aware change mask, M_F^K, by masking all change values below 0.5 to equal zero – this can remove potential low-value false changes flagged in the feature change mask.

1. Building a 3D Reference Scene Representation   **3. Embedding Change Channels in a 3D Inference Scene Representation**

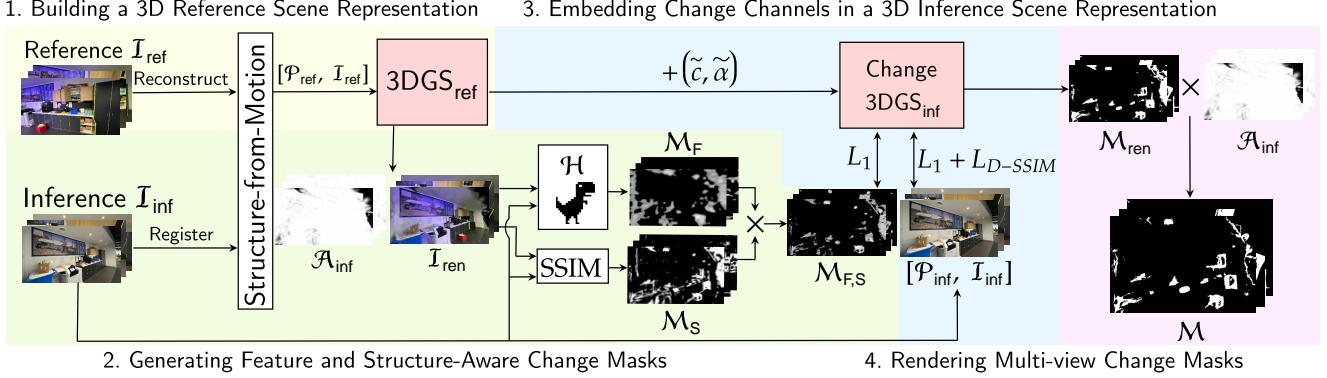**2. Generating Feature and Structure-Aware Change Masks**   **4. Rendering Multi-view Change Masks**

Figure 2. An overview of our proposed approach for multi-view pose-agnostic change detection. We leverage a 3DGS representation of the pre-change (*reference*) scene to build feature and structure-aware change masks given images of the post-change (*inference*) scene. We embed this information as additional change channels into the representation, which can be used to render multi-view change masks.

Structure-Aware Change Mask: Alongside our feature-aware change mask, we additionally generate a structure-aware change mask by leveraging the Structural Similarity Index Measure (SSIM) [41]. The SSIM quantifies the similarity between two spatially aligned image signals based on luminance, contrast, and structure components of the images. It is typically used as a metric for the visual quality of images, for example used in image reconstruction to measure the quality of the reconstruction [12]. We observe that the SSIM can also serve as a meaningful measure of change between two images, which is complementary to the feature-level change extracted from a pre-trained model. We generate our structure-aware change masks by applying the SSIM to the pairs of $\mathcal{I}_{\text{ren}}$ and $\mathcal{I}_{\text{inf}}$, and binarizing the output to filter for low-similarity, high visual change values,

$$M_S^k = \mathbf{1}(\text{SSIM}(I_{\text{ren}}^k, I_{\text{inf}}^k) \leq 0.5), \qquad (2)$$

where $\mathbf{1}$ is the indicator function.

Combined Candidate Change Mask: We combine the feature-aware and structure-aware change masks by element-wise multiplication to create the final change masks that filter for detected changes in both the features and pixel level:

$$M_{\text{F,S}}^k = \{M_F^k \cdot M_S^k\}_{k=1}^{n_{\text{inf}}}. \qquad (3)$$

Next, we describe how the individual per-view change masks $M_{\text{F,S}}^k$ are combined and fused through the 3DGS model – making our approach multi-view.

3.4. Embedding Change Channels in a 3D Inference Scene Representation

A core contribution of our method is that we move beyond change masks generated by individual images to create change masks that leverage our 3D reference scene representation, i.e. multi-view change masks. We achieve this by learning a new 3DGS representation for the inference scene that also contains change information from our feature and structure-aware change masks $\mathcal{M}_{\text{F,S}}$. We embed this change information directly into a 3DGS by learning two additional channels per Gaussian – a change magnitude $\tilde{c}$ (i.e. the level of change each Gaussian captures in the scene) and a change opacity factor $\tilde{\alpha}$ (which allows us to model which Gaussians contribute to the pixel change values in $\mathcal{M}_{\text{F,S}}$). Using these new change parameters, we can then render a change mask from the 3DGS alongside RGB images using the standard rasterization process [12].

To achieve this, we create a new change-aware 3DGS for the inference scene, Change-3DGS$_{\text{inf}}$, that is initialized with the learnt Gaussians from 3DGS$_{\text{ref}}$. For each Gaussian, we add an additional two parameters to model change in the scene $(\tilde{c}, \tilde{\alpha})$. We then re-optimize Change-3DGS$_{\text{inf}}$ given $\mathcal{I}_{\text{inf}}$, $\mathcal{P}_{\text{inf}}$ and $\mathcal{M}_{\text{F,S}}$, following the standard optimization pipeline described in [12] while including an additional L_1 loss term to learn the change channel values. For best performance of our method, Change-3DGS$_{\text{inf}}$ is initialized with the pre-trained 3DGS$_{\text{ref}}$ so that Gaussians relating to structural changes in the inference scene are retained.

Critically, we model $\tilde{c}$ using a spherical harmonics coefficient degree of zero. Typically in 3DGS [12], a higher degree (degree 3) of spherical harmonics coefficients is used to model view-dependent color, effectively capturing color variations across different viewing directions. We hypothesize that changes in a scene are largely view-independent, and that most view-dependent variations in our change masks arise from false positive change predictions, such as reflections, shadows, or minor misalignment between the rendered and inference images. Under this hypothesis, it is then preferable to model change with a low degree of spherical harmonics coefficients so that we can effectively leverage individual change masks to collectively learn true regions of change in the scene while not overfit-

ting to view-dependent false positive changes – we confirm this in Sec. 5.4.

3.5. Rendering Multi-View Change Masks

Given Change-3DGS$_\text{inf}$ and any query 6D pose P_query, we can render a multi-view change mask. Given our problem setup, we render change masks for all poses from the inference scene, $\mathcal{M}_\text{ren} = \{M_\text{ren}^k\}_{k=1}^{n_\text{inf}}$. Notably, our approach allows us to also render change masks for viewpoints that are novel to both the reference and inference scene (See Sec. 5.3 for a further discussion).

As the reference and inference scene are collected with random trajectories, independently, it is possible that inference images capture scene regions that were absent in the reference image set. Previously unseen regions of the 3DGS do not contain Gaussians, and thus rendered images of such regions are represented with black pixels (the 3DGS background color). To avoid falsely calculating these unseen areas as changes, we exclude them from the rendered change mask as a final post-processing step.

We render the alpha channel $\mathcal{A}_\text{ren} = \{A_\text{ren}^k\}_{k=1}^{n_\text{inf}}$ alongside $\mathcal{I}_\text{ren}$ as it provides per-pixel opacity between the foreground pixel versus the background. For unseen regions, a 3DGS renders the unseen region as the background color, resulting in alpha channel values close to 0 for unseen areas, and values close to 1 for well-observed regions. We binarize the alpha channel and use this to filter out false changes produced from unseen areas. This produces our final multi-view change masks as follows:

$$M^k = M_\text{ren}^k \cdot \mathbf{1}(A_\text{ren}^k \geq 0.5). \qquad (4)$$

3.6. Data Augmentation for Learning Change Channels

In this section, we explain how the set of individual image change masks can be augmented by also considering the *reference* scene poses with a 3D representation of the *inference* scene – effectively reversing the change comparison between the scenes.

Following our pipeline, we obtain a change-aware 3DGS representing the inference scene, Change-3DGS$_\text{inf}$. This Change-3DGS$_\text{inf}$ can then be used to render inference scene (post-change) images for all reference scene (pre-change) viewpoints $\mathcal{P}_\text{ref}$. Following the process outlined in Sec. 3.3, we can then generate feature and structure-aware change masks by comparing the original $\mathcal{I}_\text{ref}$ with these newly rendered images. These change masks can be concatenated with those initially calculated from the inference scene viewpoints $\mathcal{P}_\text{inf}$ to create an augmented set of masks and once again re-optimize the change channels in Change-3DGS$_\text{inf}$ as described in

4. Experimental Setup

4.1. Datasets

We evaluate our approach on the released subset of the MAD-Real dataset [47], originally developed for pose-agnostic anomaly detection. The MAD-Real dataset contains both simulated and real-world scenes containing close-up captures of single LEGO objects, where a single anomaly is simulated per scene and ground-truth masks localizing the anomaly are provided. Anomalies are either structural (missing parts) or surface-level (stains). With our focus on real-world change detection, we use the 10 publicly available real-world LEGO toy scenes in the dataset. We use the "train" subset to represent the reference scenes (each scene containing approximately 50 images), and the "test" subsets with anomalies to represent the inference scenes (each scene containing between 17-32 images).

The underlying motivation of our work is to enable autonomous systems operating in real-world scenes to detect change, where simultaneous distinct changes can occur in the scene and there may be "distractor" visual changes such as varying lighting, shadows, or reflections. To this end, we have created the novel **Pose-Agnostic Scene-Level Change Detection Dataset (PASLCD)**, comprising data collected from ten complex, real-world scenes, including five indoor environments and five outdoor environments (see Fig. 3 for a visualization of the scenes). Among the indoor and outdoor scenes, two are 360° scenes, while the remaining three are front-facing (FF) scenes.

For all ten scenes, there are two available instances to test change detection: (1) an instance for change detection under consistent lighting conditions, and (2) an instance for change detection under varied lighting conditions. Images were captured using an iPhone following a random and independent trajectory for each scene instance. Every inference scene contains multiple changes compared to the reference scene (ranging between 5 to 16), encompassing both surface-level and structural modifications to objects or surfaces in the scene. For evaluation purposes, we provide 50 human-annotated change segmentation masks per scene (25 for consistent lighting conditions and 25 for changed lighting conditions), totaling 500 annotated masks for the dataset. Reference scenes contain between 50-109 (68 on average) images. For a full description of the data structure, changes introduced in each scene, data collection protocol, and additional visualizations, we kindly refer readers to the GitHub repository. Our dataset will be made publicly available upon paper acceptance.

4.2. Baselines and Metrics

We evaluate against the two state-of-the-art approaches in pose-agnostic, self-supervised anomaly detection: Omni-PoseAD [47] and SplatPose [15]. We additionally pro-

Table 1. Performance on the MAD-Real dataset, with results averaged over all ten LEGO object scenes.

Method	mIoU ↑	F1 ↑	AUROC ↑
OmniPoseAD [47]	0.064	0.115	0.937
SplatPose [15]	0.077	0.123	0.898
Feature Difference	0.052	0.089	**0.967**
Ours	**0.132**	**0.210**	0.953

vide the baseline "Feature Difference" (Feature Diff.) using our feature-difference change masks $D^k_{\text{normalized}}$ calculated in Sec. 3.3 – this represents our method's performance before the inclusion of our key contributions with only per-view feature difference from a pre-trained model.

Following standard practices in the scene-level change detection literature [2, 17, 20, 32–34], we report mean Intersection over Union (mIoU) and F1 score as our primary evaluation metrics. For the MAD-Real dataset [47], we follow the initial evaluation and additionally report the Area Under the Receiver Operating Characteristic Curve (AUROC). While the AUROC additionally considers the magnitude of change values in the mask (whereas mIoU and F1 do not), it can be misleading when there is large class imbalance (e.g. many more negative samples than positive samples, as is the case for change detection) [31]. We additionally report the Area Under the Precision-Recall Curve (AUPR), which is a more robust metric for imbalanced binary classification scenarios [31].

When calculating mIoU and F1, all methods are required to produce a score-less binary mask (change vs. no change). Given that we are operating in a self-supervised setting without labels or a validation set, it is not possible to optimize for a threshold to convert continuous change masks into a binary mask. For all methods, we therefore threshold change masks with a value of 0.5 to provide a binarized change mask. We select 0.5 as it is the midpoint of possible change values, ranging from 0 to 1.

5. Experimental Results

5.1. Performance on Single-Object Scenes

As shown in Tab. 1, our method surpasses the state-of-the-art on the MAD-Real single-object LEGO scenes across all three metrics. In particular, our mIoU achieves approximately a **1.7×** improvement over SplatPose [15], our closest competitor.

While our multi-view method is superior to the Feature Diff. baseline for mIoU and F1, the AUROC metric shows our Feature Diff. baseline to achieve higher performance by 1.4%. There are a number of potential contributors to this result: (1) the susceptibility of the AUROC metric to class imbalance (see Sec. 4.2), (2) the in-built design of our method towards binary change masks rather than continu-

ous change scores, and (3) diminishing advantages of our multi-view change masks in very simple scenes.

5.2. Performance on Multi-Object Multi-Change Scenes under Different Lighting

In Tab. 2, we present results for each scene in our PASLCD dataset averaged across the two instances with varying lighting conditions. Our method consistently outperforms all baselines, validating our claim that we achieve state-of-the-art performance for multi-object scene change detection – we achieve approximately **1.7×** improvement in mIoU and **1.6×** in F1 score over the best-performing competitor.

Robustness to Distractor Visual Changes: In Tab. 3, we report the relative *loss in performance* of each method when evaluating under different lighting conditions versus consistent lighting conditions. For both the mIoU and F1 metrics, our multi-view change masks exhibit the least performance drop under different lighting conditions, demonstrating our robustness to distractor visual changes.

Qualitative Results: Fig. 3 presents a randomly sampled example change detection from each scene for all methods. Prior state-of-the-art methods OmniPoseAD [47] and SplatPose [15] scale poorly to multi-object scenes, with the optimization-based pose estimation often failing to converge to a global minimum (see the Cantina, Printing Area, and Pots scenes in Fig. 3). Convergence frequently fails when inference images lack sufficient overlap with the images in the reference set and the methods cannot obtain a reasonable coarse pose estimation for the optimization.

We also observe some consistent failure cases of our multi-view change masks in Fig. 3: (1) identifying color-based surface-level changes (spill on the bench in Cantina scene and T block color change in Meeting Room scene). Upon investigation, this is due to the failure of the pre-trained foundation model to produce feature changes in these conditions; (2) difficulty identifying very small changes in large-scale scenes (see Playground and Lunch Room scenes); (3) overestimating change masks for true changes, due to the patch-to-pixel interpolation of our feature masks. This is observed to a greater degree in the Feature Difference baseline.

5.3. Performance with Limited Inference Views

In Fig. 4, we explore how the number of images observed in the inference scene (n_{inf}) influences the performance of our multi-view change masks on seen and unseen views. For all indoor scenes in our PASLCD dataset, we randomly sample 5, 10, and 15 images for seen views from the total 25 available images in the scene. We hold-out poses of 10 images from the remaining 25 images as unseen views. We report the mean and standard deviation across 3 random trials.

Robustness to Limited Inference Scene Views: As shown in the left-hand of Fig. 4, our method's performance

Table 2. Quantitative results for our dataset, averaged across both similar and different lighting condition instances. Our method consistently improves over the baselines in all instances.

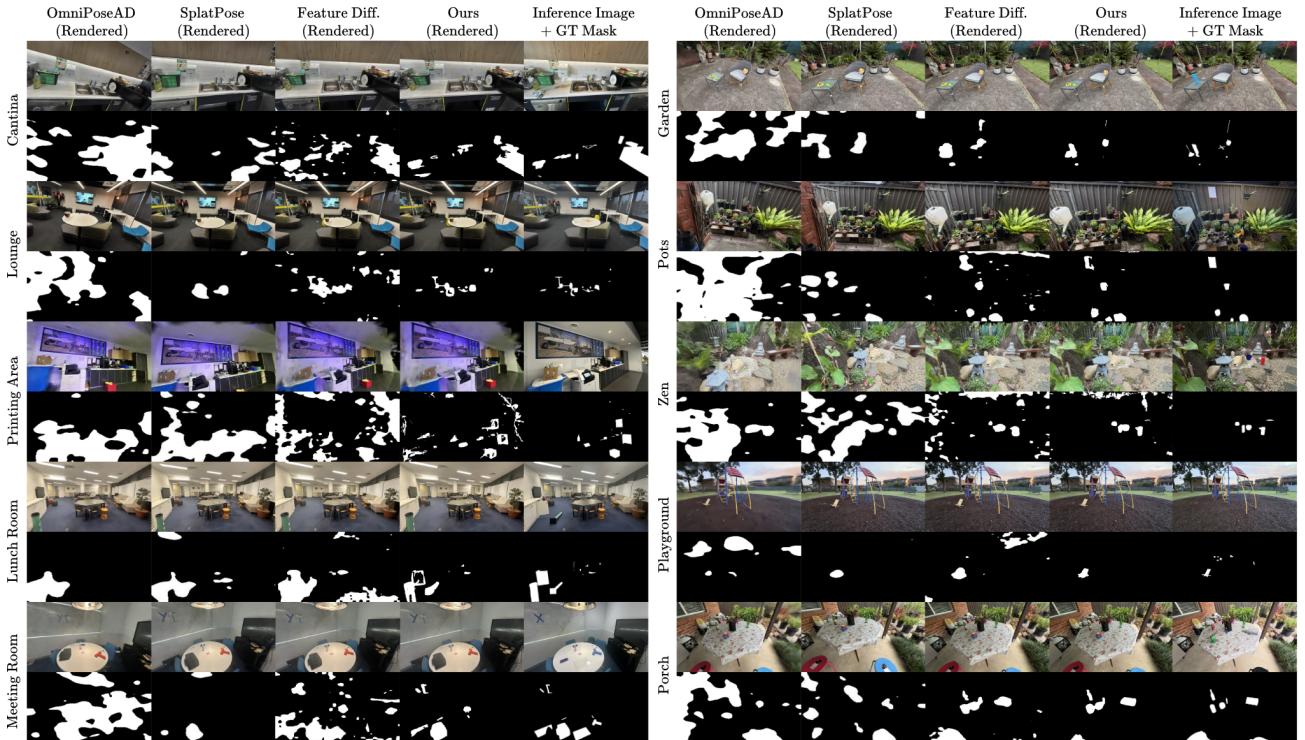| Scene	In/Outdoor	FF/360	OmniPoseAD [47]		SplatPose [15]		Feature Diff.		Ours	
			mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑
Cantina		FF	0.138	0.231	0.188	0.304	0.251	0.382	**0.580**	**0.729**
Lounge		FF	0.149	0.241	0.262	0.410	0.177	0.296	**0.463**	**0.626**
Printing Area	Indoor	FF	0.157	0.242	0.183	0.288	0.432	0.584	**0.588**	**0.734**
Lunch Room		360	0.161	0.247	0.133	0.215	0.101	0.174	**0.389**	**0.546**
Meeting Room		360	0.107	0.182	0.130	0.211	0.122	0.211	**0.350**	**0.507**
Garden		FF	0.273	0.411	0.185	0.300	0.292	0.445	**0.436**	**0.601**
Pots		FF	0.143	0.230	0.140	0.290	0.397	0.566	**0.540**	**0.693**
Zen	Outdoor	FF	0.205	0.317	0.186	0.296	0.444	0.577	**0.500**	**0.633**
Playground		360	0.076	0.125	0.081	0.133	0.047	0.089	**0.249**	**0.378**
Porch		360	0.274	0.396	0.239	0.363	0.379	0.538	**0.518**	**0.676**
Average	–	–	0.168	0.262	0.173	0.281	0.264	0.386	**0.461**	**0.612**

Figure 3. Qualitative results of each approach on our PASLCD dataset. Our generated change masks consistently agree more closely with the ground truth compared to the baselines.

increases as more views from the inference scene can be leveraged for the multi-view change masks. Notably, even with only 5 images from the inference scene, our method is still able to outperform the Feature Diff. baseline by an impressive margin (approximately $1.8\times$ the mIoU) averaged over all trials. As expected, the Feature Diff. baseline maintains consistent performance regardless of the number of images in the inference scene, as it treats images individu-

Table 3. Relative performance loss (Δ) of each method when detecting changes in scenes with different lighting conditions.

Method	ΔmIoU (%) ↓	ΔF1 (%) ↓
OmniPoseAD [47]	8.87	7.25
SplatPose [15]	19.3	19.7
Feature Diff.	17.2	12.6
Ours	**7.2**	**4.5**

Figure 4. Performance with varying numbers of inference scene views.

Table 4. Quantitative results for varying SH degree. Lower SH degrees yield better change detection performance.

SH Degree	mIoU ↑	F1 ↑	# FP_{im} ↓	# FN_{im} ↓
0	**0.4741**	**0.628**	**879**	534
1	0.460	0.614	1030	478
2	0.453	0.607	1142	442
3	0.442	0.596	1257	**416**

Table 5. Ablation of our method reported on PASLCD.

Component	mIoU ↑	F1 ↑
Feature Difference	0.264	0.382
Change-3DGS (feature-aware masks)	0.311	0.448
Change-3DGS (structure-aware masks)	0.324	0.461
Change-3DGS (combined)	0.449	0.598
Learned Mask + Aug.	0.457	0.605
Learned Mask + Aug. + Alpha Channel	**0.461**	**0.612**

ally when generating change masks.

Generating Change Masks for Unseen Views: We also validate our claim that our method can generalize to unseen views by generating change masks for query poses that *have not been observed* in the inference scene. This is a new capability unlocked by our change detection method that has not been previously explored – only by embedding change information in a 3D representation can we render change masks for entirely unseen views.

For each trial, we render change masks for the 10 unseen query poses (there are only 25 images per scene so there are no unseen views when using 25 inference views). The right-hand of Fig. 4 shows that our approach generates change masks for *unseen* views that outperform the Feature Diff. baseline on *seen* data, with mIoU ranging between 0.36-0.45 on average depending on the number of views in the inference scene used to learn the multi-view change masks.

5.4. Ablations

Spherical Harmonics Degree: Tab. 4 validates our hypothesis that lower degrees of spherical harmonics (SH) coefficient allow the 3DGS to remove view-dependent false positive change predictions. Results are averaged over the 5 indoor scenes in our PASLCD dataset and show that the lowest SH degree provides the best mIoU and F1 for our multi-view change masks. We also report the average number of false positive (FP) and false negative (FN) pixels per image, showing an approximate 70% reduction in false change predictions (FPs) between the highest and lowest SH Degree. As expected, inhibiting view-dependent change mod-

elling with lower SH degrees also introduces a slight trade-off with increased missed changes (FNs), although not outweighing the gains from reduced FPs.

Ablation on Different Modules: Tab. 5 shows the performance contributed by the individual modules in our proposed method: (1) the Feature Difference baseline, (2) learning a Change-3DGS with only feature-aware masks, (3) a Change-3DGS with only structure-aware masks, (4) our proposed Change-3DGS, (4) when including data augmentation, and (5) when accounting for unseen regions with the alpha channel. In particular, we validate our claim that the feature-aware mask and structure-aware masks contain complementary information that can be combined for best performance – their combined mIoU performance improves the performance of either alone by a factor of approximately **1.4×**.

6. Conclusion

We presented a new state-of-the-art approach to label-free, pose-agnostic change detection. We integrate multi-view change information into a 3DGS representation, enabling robust change localization even for unseen viewpoints. We additionally introduced a new change detection dataset featuring multi-object real-world scenes, which we hope will drive further advancements in the change detection community. Future work should focus on addressing the limitations observed in the feature masks from the pre-trained foundation model, namely difficulty identifying color-based surface-level changes and difficulty producing refined change masks.

References

[1] Edward H. Adelson and James R. Bergen. The Plenoptic Function and the Elements of Early Vision. In *Computational Models of Visual Processing*. The MIT Press, 1991. 3

[2] Pablo F. Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7): 1301–1322, 2018. 1, 2, 6

[3] Wele Gedara Chaminda Bandara and Vishal M. Patel. A Transformer-Based Siamese Network for Change Detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210, Kuala Lumpur, Malaysia, 2022. IEEE. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469, New Orleans, LA, USA, 2022. IEEE. 3

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019. ISSN: 2575-7075. 2

[6] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully Convolutional Siamese Networks for Change Detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, Athens, 2018. IEEE. 1, 2

[7] Hao Chen and Zhenwei Shi. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10):1662, 2020. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. 1, 2

[8] 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. In *International Conference on Learning Representations*, 2021. 2

[9] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature Interaction is What You Need for Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023. 2

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500, New Orleans, LA, USA, 2022. IEEE. 3

[11] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, Brussels, Belgium, 2018. Association for Computational Linguistics. 1, 2

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3, 4

[13] Salman Khan, Xuming He, Fatih Porikli, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Learning deep structured network for weakly supervised change detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2008–2015, Melbourne, Australia, 2017. International Joint Conferences on Artificial Intelligence Organization. 2

[14] Tomáš Krajník, Jaime P. Fentanes, Oscar M. Mozos, Tom Duckett, Johan Ekekrantz, and Marc Hanheide. Long-term topological localisation for service robots in dynamic environments using spectral maps. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4537–4542, 2014. ISSN: 2153-0866. 2

[15] Mathis Kruse, Marco Rudolph, Dominik Woiwode, and Bodo Rosenhahn. Splatpose & detect: Pose-agnostic 3d anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3950–3960, 2024. 2, 3, 5, 6, 7, 8

[16] Seonhoon Lee and Jong-Hwan Kim. Semi-Supervised Scene Change Detection by Distillation from Feature-metric Alignment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1224, Waikoloa, HI, USA, 2024. IEEE. 1

[17] Yinjie Lei, Duo Peng, Pingping Zhang, Qiuhong Ke, and Haifeng Li. Hierarchical Paired Channel Fusion Network for Street Scene Change Detection. *IEEE Transactions on Image Processing*, 30:55–67, 2021. 6

[18] Jie Li, Xing Xu, Lianli Gao, Zheng Wang, and Jie Shao. Cognitive visual anomaly detection with constrained latent representations for industrial inspection robot. *Applied Soft Computing*, 95:106539, 2020. 2

[19] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-Frequency Channel-Selection Representations for Unsupervised Anomaly Detection. *IEEE Transactions on Image Processing*, 32:4327–4340, 2023. 2

[20] Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. Robust Scene Change Detection Using Visual Foundation Models and Cross-Attention Mechanisms, 2024. arXiv:2409.16850 [cs]. 1, 2, 6

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[22] Eric Martinson and Paula Lauren. Meaningful Change Detection in Indoor Environments Using CLIP Models and NeRF-Based Image Synthesis. In *2024 21st International Conference on Ubiquitous Robots (UR)*, pages 603–610, 2024. 2

[23] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Net-

works: Learning 3D Reconstruction in Function Space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465. IEEE Computer Society, 2019. 3

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41 (4):102:1–102:15, 2022. 3

[27] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174. IEEE Computer Society, 2019. 3

[29] Ragav Sachdeva and Andrew Zisserman. The Change You Want to See (Now in 3D). In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2052–2061, Paris, France, 2023. IEEE. 1, 2

[30] Ragav Sachdeva and Andrew Zisserman. The Change You Want To See. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3993–4002, 2023. 2

[31] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3): e0118432, 2015. 6

[32] Ken Sakurada and Takayuki Okatani. Change Detection from a Street Image Pair using CNN Features and Superpixel Segmentation. In *Procedings of the British Machine Vision Conference 2015*, pages 61.1–61.12, Swansea, 2015. British Machine Vision Association. 2, 6

[33] Ken Sakurada, Weimin Wang, Nobuo Kawaguchi, and Ryosuke Nakamura. Dense Optical Flow based Change Detection Network Robust to Difference of Camera Viewpoints, 2017. arXiv:1712.02941 [cs].

[34] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly Supervised Silhouette-based Semantic Scene Change Detection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6861–6867, Paris, France, 2020. IEEE. 2, 6

[35] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[36] Nian Shi, Keming Chen, and Guangyao Zhou. A Divided Spatial and Temporal Context Network for Remote Sensing Change Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4897–4908, 2022. 2

[37] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural Geometric Level of Detail: Real-time Rendering with Implicit 3D Shapes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11353–11362, Nashville, TN, USA, 2021. IEEE. 3

[38] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P. Balamuralidhar. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In *Computer Vision – ECCV 2018 Workshops*, pages 129–145, Cham, 2019. Springer International Publishing. 1, 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[40] Guo-Hua Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation. *Pattern Recognition*, 2023. 1, 2

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2, 4

[42] Zhixue Wang, Yu Zhang, Lin Luo, and Nan Wang. Transcd: scene change detection via transformer-based architecture. *Opt. Express*, 29(25):41409–41427, 2021. 2

[43] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 2

[44] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8310–8319, Montreal, QC, Canada, 2021. IEEE. 2

[45] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DSR – A Dual Subspace Re-Projection Network for Surface Anomaly Detection. In *Computer Vision – ECCV 2022*, pages 539–554, Cham, 2022. Springer Nature Switzerland.

[46] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A Feature Selection Network with Realistic Synthetic Anomaly for Anomaly Detection, 2024. arXiv:2403.05897 [cs]. 2

[47] Qiang Zhou, Weize Li, Lihan Jiang, Guoliang Wang, Guyue Zhou, Shanghang Zhang, and Hao Zhao. Pad: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5, 6, 7, 8