From Pixels to Primitives: Scene Change Detection in 3D Gaussian Splatting

Chamuditha Jayanga Galappaththige[1,2] **Jason Lai**[3] **Timothy Patten**[2,4]
Donald Dansereau[2,3] **Niko Suenderhauf**[1,2] **Dimity Miller**[1,2]
[1]QUT Centre for Robotics [2]ARIAM [3]ACFR, University of Sydney [4]Abyss Solutions
{chamuditha.galappaththige, d24.miller}@qut.edu.au

Abstract

Scene change detection methods built on Gaussian splatting universally follow a render-then-compare paradigm: the pre-change scene is rendered into 2D and compared against post-change images via pixel or feature residuals. This *change detection problem with Gaussian Splatting has been treated as a question about pixels; we treat it as a question about primitives.* We provide direct evidence that native primitive attributes alone – position, anisotropic covariance, and color – carry sufficient signal for scene change detection. What makes primitive-space comparison hard is the under-constrained nature of Gaussian splatting representation: independent optimizations yield primitive solutions whose count, positions, shapes, and colors differ even where nothing has changed. We address this challenge with anisotropic models of geometric and photometric drift, complemented by a per-primitive observability term that reflects the extent to which each Gaussian is constrained by the camera geometry. Operating directly on primitives gives our method, GS-DIFF, two properties that distinguish it from render-then-compare methods. First, change maps are multi-view consistent by construction, where prior work had to learn this through an additional optimization objective. Second, geometric and appearance changes are scored separately, identifying not just *where* but *what kind of* change occurred, distinguishing structural changes (e.g., an added object) from surface-level ones (e.g., a color change) without supervision or external model dependencies. On real-world benchmarks, GS-DIFF surpasses the prior state-of-the-art approach by $\sim$ **17%** in mean Intersection over Union. Code and annotations will be released upon acceptance.

1 Introduction

3D Gaussian splatting (3DGS) [29] has become a foundational representation for photorealistic digital twins [18], with adoption spanning spatial maps [37], heritage preservation [8], and industrial asset management [14]. A critical challenge in these applications is understanding how environments evolve over time: an agent visits a scene at different times, capturing images from unconstrained viewpoints, and must detect and localize changes that have occurred between inspections. This *Multi-View Scene Change Detection* (SCD) problem [16, 14, 36] requires robustness to arbitrary camera trajectories and the ability to distinguish genuine structural or appearance changes from irrelevant variations such as lighting shifts, shadows, and reflections.

Existing work [14, 36, 26, 31, 34, 1, 16] follows a *render-then-compare* paradigm, treating 3DGS as a *rendering engine*: the pre-change scene is reconstructed, rendered at post-change poses, and compared to post-change images via foundation-model features [38, 40], pixel-level metrics [48], or combinations thereof (Figure 1). Because genuine changes register consistently across views while distractors such as shadows and reflections fire inconsistently, state-of-the-art methods aggregate change cues across views to enforce multi-view consistency [14, 16, 36, 50, 1], suppressing these

Preprint.

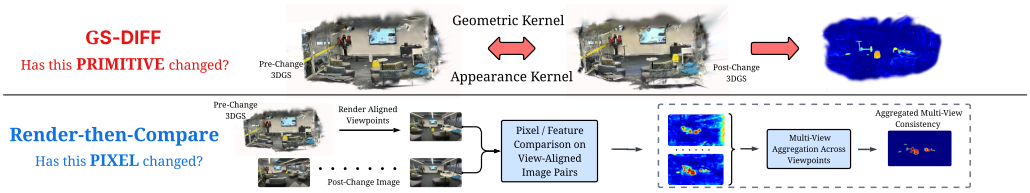

Figure 1: From pixels to primitives. Prior multi-view SCD methods question *pixels*, comparing rendered viewpoint pairs in image space and learning to aggregate change evidence scattered across viewpoints (bottom). GS-DIFF questions *primitives*, comparing two 3DGS reconstructions directly in primitive space (top). Multi-view consistency emerges *by construction* from the shared 3D representation, eliminating both per-view comparison and learned aggregation.

inconsistencies. In every case, the *change signal* originates in *image space*; primitives serve only as a rendering and aggregation substrate, never as the basis for comparison.

Prior work compares scenes in image space – pixel by pixel, view by view. We instead ask whether the comparison can happen directly between Gaussian primitives. Each primitive explicitly encodes position, covariance, opacity, and color – the very attributes that change when a scene changes, and determine how those changes appear in a rendered image. If scene-level change is already encoded in the primitives, we hypothesize that change can be detected in the 3DGS representation directly, removing both the detour through rendered images and the cross-view aggregation it requires.

To this end, we present GS-DIFF, which detects scene-level changes by comparing two independently reconstructed 3DGS scenes *directly in primitive space*. What has – so far – prevented direct primitive-space comparison is the under-constrained nature of 3DGS reconstruction [29]: for a given set of images, the representation is non-unique and many distinct primitive configurations, differing in count, position, shape, and color, explain the same observations. Therefore, independent optimizations of even unchanged scenes converge to solutions without direct primitive-to-primitive correspondence. We call the resulting geometric and color differences between corresponding unchanged regions *geometric* and *photometric drift*. We address this drift in two ways to enable direct primitive comparison: a geometric kernel over position and covariance, complemented by a Fisher Information observability term based on camera view geometry, absorbs geometric drift (§3.2.1), and an appearance kernel over diffuse color absorbs photometric drift (§3.2.2).

Operating directly in primitive space gives GS-DIFF two appealing structural properties. First, the change signal is multi-view consistent *by construction*, whereas prior work [14, 16] learned viewpoint consistency through multi-view aggregation of 2D change cues. Second, by considering primitive attributes separately – position-covariance versus color – GS-DIFF identifies not only *where* change has occurred but *what kind of* change it is. This disambiguation is unavailable to previous image-space methods without learned change classification or auxiliary models. To quantitatively evaluate this, we annotate each change pixel in the PASLCD benchmark [14] as structural or surface-level.

In summary, we make the following claims, validated by our experiments:

- **Gaussian primitives alone carry sufficient signal for state-of-the-art multi-view SCD.** We present GS-DIFF, the first method to perform SCD in 3DGS primitive space, surpassing the strongest image-space baseline by $\sim 17\%$ in mIoU (see §4.1).
- **Reconstruction drift is the obstacle to direct primitive comparison.** We resolve it with anisotropic geometric and data-driven photometric drift models enabling robust comparison across independent 3DGS reconstructions (see §4.2).
- **Primitive-space comparison allows distinguishing structural from surface-only changes.** This is inherent to GS-DIFF and requires no supervision or auxiliary models (see §4.3).

2 Related Work

Scene Change Detection via Image Comparison. SCD across unconstrained viewpoints has progressed through three eras, all operating in image space and depending on learned or pretrained image features. Early methods assume strictly paired pre- and post-change views [2, 9, 47, 44]. A second line tolerated minor viewpoint variation through supervised annotations [42, 43], but

supervised formulations are inherently brittle: dense annotation is expensive, the distribution of real-world changes is unbounded, and performance degrades [30, 7] sharply under distribution shift [12, 13, 22]. The most recent line accordingly pivots to label-free and zero-shot paradigms [33, 30, 7, 27, 3, 11] driven by vision foundation models [38, 40]. Across all three eras, the comparison remains structurally tethered to 2D image space and assumes viewpoint consistency between pairs – a condition rarely satisfied when observing agents follow independent trajectories.

Recent methods enable unconstrained viewpoints by leveraging 3DGS [29] to synthesize reference views at inference-time poses. Object-level pose-agnostic anomaly detection [31, 34] scores changes by rendering views and then comparing features, while the SCD community extended this to complex multi-change scenes [36, 26, 14, 50, 15]. A core insight in this line is that per-view change signals are not consistent across viewpoints: true changes register with varying strength depending on visibility, lighting, and feature response, but consistently across viewpoints, while distractors (e.g., shadows or reflections) fire inconsistently. MV3DCD [14] embedded change cues as per-Gaussian change channels and fused viewpoints through hard thresholding and heuristics, which enabled MV3DCD to suppress view-inconsistent signals that pairwise methods could not. O-SCD [16] replaced this heuristic fusion with a self-supervised loss, providing evidence that the strength of the learned objective is itself a ceiling on achievable consistency. However, all existing methods use Gaussian primitives as a mere substrate for 2D rendering or aggregation, ultimately deriving 2D change signals in image space using pretrained features, a process susceptible to the chosen backbone's failure modes [14]. We break from this paradigm by performing direct primitive-space comparison. This shift, enabled by explicit reconstruction drift modeling, yields a pipeline that is free of foundation model dependence, change annotations, and extended multi-view aggregation machinery.

Scene Change Detection on 3D Representations. Change detection between 3D representations has a long history in geospatial surveying, comparing bitemporal LiDAR scans via cloud-to-cloud distances [20], M3C2 [32], octree occupancy [19], scene graphs [35], and deep learning on point clouds [49, 10, 39]. Our method also compares two 3D representations directly, but differs in two critical respects. First, point cloud methods operate on raw LiDAR data with sensor-characterized noise, whereas our representation is optimized from images and introduces reconstruction drift with no LiDAR analogue. Second, our method exploits the anisotropic spatial extent of Gaussian primitives, which point clouds do not carry. Images encode both geometry and photometric appearance, and 3DGS inherits this richness, admitting joint geometric-photometric scoring over a shared representation. As 3DGS becomes the standard format for repeated spatial capture [18], native change detection on Gaussian primitives is timely; ours is the first such approach.

3 Method

3.1 Preliminaries

Task setting. Following prior work [36, 50, 14, 16], a scene is captured at two different times, producing a *reference* (pre-change) image set $\mathcal{I}^{(1)}$ and an *inference* (post-change) image set $\mathcal{I}^{(2)}$ along independent camera trajectories that need not share identical views. The objective is to output a binary change mask for images in $\mathcal{I}^{(2)}$, marking pixels whose scene content differs from $\mathcal{I}^{(1)}$.

Gaussian Splatting. 3DGS [29] represents a scene as a set of anisotropic Gaussians $\mathcal{G} = \{g_i\}_{i=1}^{N}$ optimized from posed images. Each primitive g_i is parameterized by a position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, a covariance $\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top$ with rotation $\mathbf{R}_i \in SO(3)$ and scales $\mathbf{S}_i = \mathrm{diag}(s_1, s_2, s_3)$, an opacity o_i, and a spherical-harmonic (SH) color whose DC coefficient $\mathbf{c}_i \in \mathbb{R}^3$ encodes diffuse color while higher-order coefficients capture view-dependent effects. Images are rendered by splatting each primitive to the image plane and alpha-compositing [51]. For geometrically accurate reconstructions, the surface normal $\mathbf{n}_i \in \mathbb{S}^2$ is approximated by the column of $\mathbf{R}_i$ corresponding to the smallest axis of $\mathbf{S}_i$.

3.2 GS-DIFF: Primitive-Space Scene Change Detection in 3D Gaussian Splatting

Figure 2 illustrates the core idea of GS-DIFF. Given two reconstructed scenes, the central challenge to direct comparison is non-uniqueness in the 3DGS representation: independent reconstructions of scenes yield many plausible primitive solutions differing in position, shape, and color while describing the same observations even when unchanged. GS-DIFF first builds an anisotropic covariance inflation

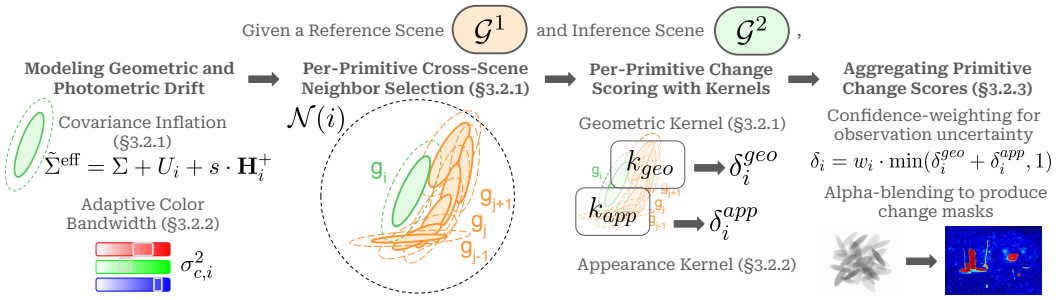

Figure 2: The GS-DIFF pipeline: We model the expected geometric and photometric drift between 3DGS representations, using the inflated covariance of each primitive to find its cross-scene neighbor set. A geometric kernel and appearance kernel evaluate change over the neighbor set to compute drift-aware change scores. These change scores are combined and weighted by observation uncertainty, and can then be rendered as change score maps for any viewpoint.

matrix for each primitive, capturing geometric drift that arises from both representation ambiguity and observation uncertainty, and folds it into each primitive's own covariance matrix. With this effective covariance accommodating geometric drift, a plausible cross-reconstruction neighbor set is drawn. Then a geometric kernel evaluates how well the existence of that primitive is described within the neighbor set, accommodating the geometric drift (§3.2.1). An appearance kernel then operates on the same neighbor set, evaluating how well the diffuse color is described (§3.2.2). To produce a change mask for any queried viewpoint, each primitive score is weighted by a confidence term and rendered following alpha-compositing [29]. We can further consider scores separately to disambiguate differences arising from underlying structural versus surface-level changes (§3.2.3).

Reconstructing 3DGS Representations. Following prior multi-view SCD work [14, 36, 50, 16], we use COLMAP [45] to estimate camera poses $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ for the image sets $\mathcal{I}^{(1)}, \mathcal{I}^{(2)}$ in a common coordinate frame. From $(\mathcal{I}^{(1)}, \mathcal{C}^{(1)})$ and $(\mathcal{I}^{(2)}, \mathcal{C}^{(2)})$ we then independently build two geometrically accurate reconstructions [41, 5] $\mathcal{G}^{(1)} = \{g_i^{(1)}\}_{i=1}^{N_1}$ and $\mathcal{G}^{(2)} = \{g_j^{(2)}\}_{j=1}^{N_2}$. Since change can only be assessed in regions both image sets observed, we discard primitives visible in only one of the camera frustum sets $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ before comparison begins. We assume $\mathcal{I}^{(1)}, \mathcal{I}^{(2)}$ individually provide sufficient scene coverage for geometrically accurate reconstruction.

3.2.1 Neighbor Retrieval and Geometric Scoring Under Geometric Drift

The *geometric kernel* scores the change in local geometry between reconstructions. A naïve nearest-neighbor (NN) distance based only on primitive position is brittle, as we empirically show in § 4.2 – NN ignores the spatial extent of each primitive and is sensitive to geometric drift; the same observations can be described by different configurations of Gaussians. Our key insight is that the primitive's spatial extent, captured by its covariance, can be leveraged to absorb this drift. We attribute drift to two sources and model each as an additive, anisotropic covariance inflation. The resulting effective covariance is used both to retrieve cross-reconstruction neighbors and regularize the kernel.

Sources of geometric drift. The first source is representation ambiguity. 3DGS representation is non-unique for a given scene: many primitive configurations differing in count, position, and shape render to the same observed images, and independent optimizations converge to different primitive solutions even on unchanged regions. The resulting positional differences are anisotropic: with sufficient camera baseline, multi-view triangulation tightly constrains across-surface (normal) position and ambiguity dominates tangentially, while under-constrained baselines admit larger normal-direction drift. The decomposition into u_n and u_t in Eq. 3 captures whichever asymmetry holds.

The second source is observation uncertainty, a form of epistemic uncertainty [28, 25] that grows when a primitive is poorly constrained by the camera viewing geometry. A primitive observed from a single narrow baseline, for example, has its position only weakly determined along the viewing ray, and small differences in optimization initialization could shift it substantially. Without separating poorly observed primitives from genuinely changed ones, both register as differences.

Modeling representation ambiguity. Representation ambiguity is a property of the reconstruction process itself; its scale can be estimated from the data. For each primitive $g_i^{(1)} \in \mathcal{G}^{(1)}$ we find its Euclidean nearest neighbor $g_{j^*}^{(2)} \in \mathcal{G}^{(2)}$ and decompose the displacement $\boldsymbol{\Delta}_i = \boldsymbol{\mu}_{j^*}^{(2)} - \boldsymbol{\mu}_i^{(1)}$ into normal and tangential components relative to the surface normal $\mathbf{n}_i$:

$$d_n^{(i)} = \left|\mathbf{n}_i^\top \boldsymbol{\Delta}_i\right|, \qquad d_t^{(i)} = \left\|\boldsymbol{\Delta}_i - (\mathbf{n}_i^\top \boldsymbol{\Delta}_i)\mathbf{n}_i\right\|_2. \tag{1}$$

We summarize the per-direction drift scale by upper-quartile statistics $Q_{0.75}$, which avoids contamination from the tail of the distribution that contains genuinely changed regions:

$$u_n^2 = \left[Q_{0.75}\Big(\{d_n^{(k)}\}_{k\in|\mathcal{G}|}\Big)\right]^2, \qquad u_t^2 = \left[Q_{0.75}\Big(\{d_t^{(k)}\}_{k\in|\mathcal{G}|}\Big)\right]^2, \tag{2}$$

computed bidirectionally ($\mathcal{G}^{(1)} \to \mathcal{G}^{(2)}$ and $\mathcal{G}^{(2)} \to \mathcal{G}^{(1)}$) and averaged for symmetry. We assemble an anisotropic inflation matrix and add it to each primitive's covariance:

$$\mathbf{U}_i = u_t^2\,\mathbf{I}_3 + (u_n^2 - u_t^2)\,\mathbf{n}_i\mathbf{n}_i^\top, \qquad \tilde{\boldsymbol{\Sigma}}_i = \boldsymbol{\Sigma}_i + \mathbf{U}_i. \tag{3}$$

Effective covariance $\tilde{\boldsymbol{\Sigma}}_i$ now reflects both each primitive's local extent and the typical geometric drift.

Modeling observation uncertainty. Observation uncertainty is intrinsically per-primitive: a primitive seen from many angles is well-constrained, while one seen from a narrow baseline is not. We quantify it through a Fisher information matrix (FIM) accumulated over each primitive's frustum-visible cameras $\mathcal{C}_i^{\text{vis}}$, following the standard form for inverse-depth uncertainty in multi-view geometry [23]:

$$\mathbf{H}_i = \sum_{c\in\mathcal{C}_i^{\text{vis}}} \frac{1}{\|\boldsymbol{\mu}_i - \mathbf{o}_c\|^2}\left(\mathbf{I}_3 - \mathbf{v}_{i,c}\mathbf{v}_{i,c}^\top\right), \tag{4}$$

where $\mathbf{o}_c$ is the camera center, $d_{i,c} = \|\boldsymbol{\mu}_i - \mathbf{o}_c\|$ the distance from primitive i to camera c, and $\mathbf{v}_{i,c} = (\boldsymbol{\mu}_i - \mathbf{o}_c)/d_{i,c}$ the unit viewing ray. The projection $\mathbf{I}_3 - \mathbf{v}\mathbf{v}^\top$ encodes that each camera constrains a primitive in the image plane but not along the viewing direction; the $1/d_{i,c}^2$ factor encodes perspective fall-off. A primitive observed from many angles at close range yields a well-conditioned $\mathbf{H}_i$; one observed from a single viewpoint yields a near-singular $\mathbf{H}_i$ with one unconstrained direction. We inject the pseudo-inverse $\mathbf{H}_i^+$ into the covariance with a data-driven scale:

$$\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}} = \tilde{\boldsymbol{\Sigma}}_i + s\cdot\mathbf{H}_i^+, \qquad s = \frac{\operatorname{median}\Big(\{\operatorname{tr}(\tilde{\boldsymbol{\Sigma}}_k)\}_{k\in|\mathcal{G}|}\Big)}{\operatorname{median}\big(\{\operatorname{tr}(\mathbf{H}_k^+)\}_{k\in|\mathcal{G}|}\big)}. \tag{5}$$

The scale s matches the typical magnitude of $\mathbf{H}_i^+$ to that of the representation-ambiguity inflation, so neither dominates. Where a primitive is under-constrained along a particular direction, $\mathbf{H}_i^+$ has a large eigenvalue along that direction, and $\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}}$ stretches accordingly – the kernel will tolerate larger displacements there before declaring a change. The same FIM also informs a per-primitive confidence weight at render time (§3.2.3), giving observation uncertainty two complementary roles.

Neighbor retrieval and geometric kernel. With $\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}}$ encoding the total expected geometric drift for primitive i, we retrieve a cross-reconstruction neighbor set within an drift-aware search radius:

$$\mathcal{N}(i) = \left\{ j : \left\|\boldsymbol{\mu}_i^{(1)} - \boldsymbol{\mu}_j^{(2)}\right\| \le \eta\cdot\sqrt{\lambda_{\max}\Big(\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}}\Big)} \right\}, \quad \eta = 3. \tag{6}$$

The Euclidean ball is a conservative superset of the Mahalanobis η-ellipsoid, ensuring no candidate within Mahalanobis distance η is missed while permitting fast spatial indexing. To measure geometric similarity between a primitive and its cross-reconstruction neighbor set, we then apply the geometric kernel as an unnormalized anisotropic Mahalanobis radial basis function (RBF):

$$k_{\text{geo}}(g_i, g_j) = \exp\!\big(-\tfrac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \mathbf{M}_{ij}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\big), \qquad \mathbf{M}_{ij} = \tilde{\boldsymbol{\Sigma}}_i^{\text{eff}} + \tilde{\boldsymbol{\Sigma}}_j^{\text{eff}}. \tag{7}$$

We omit the normalization prefactor $(2\pi)^{-3/2}|\mathbf{M}_{ij}|^{-1/2}$: the determinant penalizes co-located primitives whose covariance magnitudes differ, a routine artifact of independent reconstructions with different densification histories. The unnormalized form preserves $k_{\text{geo}} = 1$ at $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$ regardless of covariance scale; we ablate this choice in §4.2. The per-primitive geometric change score is:

$$\delta_i^{\text{geo}} = 1 - \max_{j\in\mathcal{N}(i)} k_{\text{geo}}(g_i, g_j). \tag{8}$$

A high δ_i^{geo} indicates that no primitive in the neighborhood $\mathcal{N}(i)$ set explains the existence of g_i once drift has been absorbed, which we interpret as evidence of geometric change.

3.2.2 Appearance Scoring Under Photometric Drift

After capturing geometric change, we then apply the *appearance kernel* to score color change between each primitive and its geometric neighbor set $\mathcal{N}(i)$ from §3.2.1. A naive direct comparison of DC colors is again brittle (as we empirically show in § 4.2), this time due to *photometric drift*: two reconstructions of an unchanged scene yield primitives whose DC colors differ even at matched positions. Two factors contribute. First, the spherical-harmonic decomposition of color is itself non-unique and subject to representation ambiguity; different SH configurations can produce the same diffuse appearance, so lighting and residual view-dependent effects bake into the DC term in inconsistent ways across reconstructions. Second, two captures rarely share identical illumination; ambient changes, time-of-day shifts, and automatic exposure or white-balance drift introduce a global color offset that is absorbed into the DC coefficients.

Modeling photometric drift. From the data, we estimate a color bandwidth σ_c as the typical photometric drift magnitude. For each primitive in $\mathcal{G}^{(1)}$, we find the color difference to its NN $g^{(2)}_{j^*}$ in $\mathcal{G}^{(2)}$ and take the median over all primitives, weighted by the geometric kernel response:

$$\sigma_c^2 = \text{median}\left(\left\{w_i \cdot \|\mathbf{c}_i - \mathbf{c}_{j^*}\|^2\right\}_{i\in|\mathcal{G}|}\right) \;\; ; \; w_i = k_{\text{geo}}(i, j^*). \tag{9}$$

computed bidirectionally ($\mathcal{G}^{(1)} \to \mathcal{G}^{(2)}$ and $\mathcal{G}^{(2)} \to \mathcal{G}^{(1)}$) and averaged for symmetry. Weighting by k_{geo} ensures only geometrically well-matched pairs contribute, suppressing contamination from genuinely changed regions. The median absorbs the global capture-side offset (which shifts the bulk of the distribution) and remains insensitive to the tail of genuinely changed pairs. We then adapt the bandwidth per primitive by spatial footprint:

$$\sigma_{c,i}^2 = \sigma_c^2 \cdot \max\left(h_i^2/\tilde{h}^2,\, 1\right), \quad h_i^2 = \text{tr}\left(\tilde{\mathbf{\Sigma}}_i^{\text{eff}}\right), \quad \tilde{h}^2 = \text{median}\left(\left\{h_k^2\right\}_{k\in|\mathcal{G}|}\right). \tag{10}$$

Larger primitives aggregate appearance over a broader area and warrant proportionally wider bandwidth; floor at σ_c^2 prevents small, well-localized primitives from receiving an overly tight bandwidth.

Appearance kernel. Using our color bandwidth to capture photometric drift, we score color similarity with an isotropic RBF over the DC color $\mathbf{c}_i$:

$$k_{\text{app}}(g_i, g_j) = \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{2\,\sigma_{c,i}^2}\right). \tag{11}$$

Restricting comparison to the DC term discards higher-order SH coefficients that encode view-dependent effects (§3.1), which would otherwise register as spurious appearance change; this matches the empirical finding of MV3DCD [14]. The per-primitive appearance change score is:

$$\delta_i^{\text{app}} = 1 - \max_{j\in\mathcal{N}(i)} k_{\text{app}}(g_i, g_j). \tag{12}$$

A high δ_i^{app} indicates that no primitive in the neighborhood $\mathcal{N}(i)$ explains g_i's color once photometric drift has been absorbed, which we interpret as evidence of appearance change.

3.2.3 Aggregating Per-Primitive Scores and Rendering

After applying the geometric and appearance kernels, each primitive carries two change scores, $(\delta_i^{\text{geo}}, \delta_i^{\text{app}})$, computed bidirectionally for per-scene change information. We weight each score by a confidence term, render 2D maps at inference viewpoints, and combine the scenes into a change map.

Confidence weighting. As foreshadowed in §3.2.1, observation uncertainty enters our pipeline at two stages. The geometric kernel absorbs observation uncertainty into the *similarity computation*: FIM-based covariance inflation widens the matching tolerance for under-observed primitives so their position uncertainty does not register as change. The render step absorbs observation uncertainty into the *aggregation*: a per-primitive confidence weight ω_i scales each primitive's contribution to the 2D map, so under-observed regions inform the final output proportionally to how well they were captured by the cameras. Without the first, drift on uncertain primitives produces false positives;

without the second, all primitives contribute equally regardless of how well-determined they are. We analyze the two stages in §4.2 and show that they compose without redundancy. We define ω_i from the same FIM (Eq. 4), centered on each scene's own reference:

$$\omega_i = \sigma\big(\log \mathrm{tr}(\mathbf{H}_i) - \log Q_{0.25}\big(\{\mathrm{tr}(\mathbf{H}_k)\}_{k\in|\mathcal{G}|}\big)\big)\,, \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid and $\mathrm{tr}(\mathbf{H}_i)$ summarizes observability across the three spatial axes. The 25th-percentile reference suppresses the bottom quartile of each scene's primitives below $\omega_i = 0.5$, with the remainder passing through above. Because ω_i depends on each primitive's relative ranking within its own scene rather than absolute information level, it is independent of the COLMAP [45] reconstruction scale, and a single formula applies across reconstructions of varying scale and density.

Rendering change maps. For each scene $s \in \{1, 2\}$ we form a saturated, observability-weighted per-primitive change score, scored bi-directionally,

$$\delta_i^{(s)} = \omega_i \cdot \min\Big(\delta_i^{\text{geo},(s)} + \delta_i^{\text{app},(s)},\ 1\Big)\,, \tag{14}$$

and render it to a 2D map $\mathcal{M}^{(s)}$ at the inference viewpoint via alpha-compositing [29], treating $\delta_i^{(s)}$ as a per-primitive scalar. The final change map is the pixel-wise maximum $\mathcal{M} = \max(\mathcal{M}^{(1)}, \mathcal{M}^{(2)})$.

Disambiguating structural from surface-only change. Our kernels provide per-primitive $(\delta_i^{\text{geo}}, \delta_i^{\text{app}})$, which we exploit to disambiguate structural from surface-only change. By construction, the geometric kernel detects only structural change: it fires on geometric mismatches. The appearance kernel, however, is ambiguous: it fires on any color mismatch within $\mathcal{N}(i)$, whether from surface-only recoloring or from a structural change whose revealed background differs from the previous foreground. The two cases are distinguishable because the second activates both kernels simultaneously, while the first activates only the appearance kernel. The residual $\max(\delta_i^{\text{app}} - \delta_i^{\text{geo}}, 0)$ therefore indicates the surface-only component. Rendering δ_i^{geo} and this residual independently yields two maps $\mathcal{M}^{\text{struct}}, \mathcal{M}^{\text{surf}}$, and for each pixel detected as changed in $\mathcal{M}$, we assign its type by

$$\text{label} = \arg\max\big(\mathcal{M}^{\text{struct}},\ \mathcal{M}^{\text{surf}}\big)\,. \tag{15}$$

This change classification requires no manual annotations and no additional auxiliary model, and emerges directly from holding the two kernels separate.

4 Experimental Analysis

Dataset. We evaluate on PASLCD [14] following prior multi-view SCD literature [16, 14]. PASLCD is a real-world benchmark of 10 scenes (indoor and outdoor) captured under both similar and different lighting, yielding 20 instances. Scenes contain multiple concurrent surface- and object-level changes alongside distractors such as shadows, reflections, and illumination shifts, captured from independently traversed trajectories. As an additional contribution, we have hand-annotated each ground-truth change pixel as *structural* (geometric mismatch) or *surface-level* (appearance-only), enabling evaluation of the change classification of §3.2.3. Of all changed pixels, 84% are structural and 16% are surface-level. We will release these annotations upon acceptance.

Baselines. We compare against the strongest representatives from each prior paradigm: supervised pairwise [42], label-free pairwise [30], pose-agnostic anomaly detection [34], and multi-view [14, 36, 16, 50]. Pairwise methods receive rendered aligned viewpoints, which simplifies their task.

Metrics. Following SCD literature [16, 14, 2, 36, 33, 43], we report mIoU and F1 on changed pixels. Change maps $\mathcal{M} \in [0, 1]$ are binarized at a fixed midpoint threshold of 0.5 with no per-scene tuning advantage; we additionally report an *Oracle* setting with per-scene optimal thresholds as the scoring upper bound. For change classification (§4.3), we use balanced accuracy [4] as the primary metric and per-class precision/recall as secondary metrics, accounting for the structural-surface class imbalance.

4.1 Main Results

Table 1 reports performance averaged over all 20 instances in PASLCD. GS-DIFF establishes a new state of the art, improving mIoU by $\sim$ **17%** over the strongest prior method while using only the

Table 1: Quantitative results on PASLCD [14]. GS-DIFF is the only method that reaches state-of-the-art performance *without external learned features* and with multi-view (MV) consistency as an *inherent property* rather than a learned objective. Best in **bold**, second best underlined.

Method	No Learned Feat.	MV Consistency	mIoU ↑	F1 ↑
CYWS-2D [42]	✗	–	0.273	0.398
GeSCD [30]	✗	–	0.477	0.611
SplatPose+ [34]	✗	–	0.237	0.358
SCAR-3D [50]	✗	Learned	0.191	0.289
3DGS-CD [36]	✗	Learned	0.209	0.339
MV3DCD [14]	✗	Learned	0.478	0.628
O-SCD [16]	✗	Learned	0.552	0.694
GS-DIFF	✓	by Construction	**0.644**	**0.758**
GS-DIFF (Oracle)	✓	by Construction	0.669	0.779

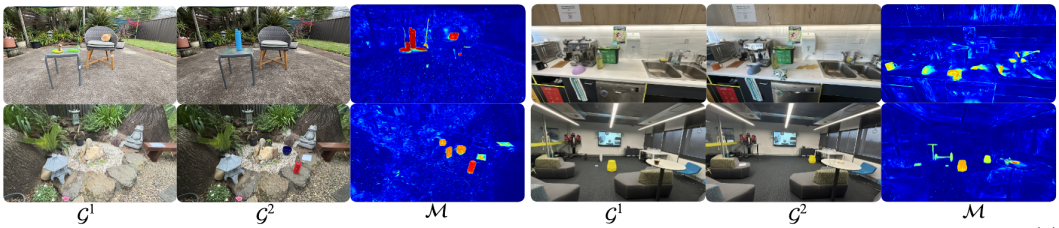

Figure 3: Qualitative results on PASLCD [14]. For each scene: rendered views from reference $\mathcal{G}^{(1)}$ (left), inference $\mathcal{G}^{(2)}$ (center), and change map $\mathcal{M}$ (right). GS-DIFF produces sharply localized responses to both structural and surface-level changes across indoor and outdoor scenes.

native attributes of Gaussian primitives: position, covariance, and color. Every baseline relies on features of foundation models [38, 40] and learned aggregation stage [16, 14, 50] (except for pairwise methods); we exceed all of them through principled handling of reconstruction drift alone, providing direct empirical evidence for our first claim on the representational sufficiency of Gaussian primitives for multi-view SCD. The narrow Fixed-Oracle gap (0.025 mIoU) shows that our continuous scores are well-calibrated and do not require per-scene threshold tuning. Figure 3 provides a representative set of qualitative results, and additional visualizations are present in Appendix C.

4.2 Component Analysis

Table 2 analyzes each component of GS-DIFF cumulatively. Three observations stand out. **(1) Normalized Mahalanobis is numerically degenerate**: the standard form (row 2) under performs Euclidean NN (row 1) as its determinant prefactor penalizes co-located primitives whose covariance magnitudes differ across reconstructions (§3.2.1); dropping it (row 3) recovers $\sim 3\times$ the mIoU. **(2) Drift modeling is the dominant component towards strong performance**: the drift-modeling phase lifts Fixed mIoU by $+0.542$, with geometric components contributing $+0.363$ jointly and appearance bandwidth plus observability weighting contributing the remaining $+0.179$; this validates our second claim that reconstruction drift, not feature representation, is the obstacle to direct primitive comparison. **(3) Observability is primarily a calibration mechanism**: FIM injection's Fixed-threshold gain is nearly $3\times$ its Oracle gain (row 5 to 6) – separability was largely present after representation-ambiguity inflation, but the score distribution was poorly calibrated; the remaining drift components close the Fixed–Oracle gap to within 95% of the Oracle ceiling. An extended analysis of data-driven quantile-sensitive is presented in Appendix B.

4.3 Distinguishing Structural from Surface-Level Changes

Table 3 shows the GS-DIFF change classification approach achieves a balanced accuracy of 0.87 under both Fixed and Oracle thresholds. Per-class recall is high in both classes (0.96 structural, 0.77 surface-level), confirming the kernel-residual decomposition genuinely captures surface-level change. The asymmetric per-class precisions (0.97 vs 0.73) reflect the underlying $84\%/16\%$ class imbalance: even small misrouting from the dominant structural class floods the smaller surface-level prediction

Table 2: Component analysis on PASLCD [14]. Each row introduces one design choice cumulatively. Phases: *Naïve* (raw nearest-neighbor comparison), *Kernel* (geometric and appearance kernels with fixed parameters), *Drift* (representation ambiguity and observation uncertainty across kernels).

Phase	Variant	mIoU	Oracle mIoU	Relative Gap (%)
Naïve	Euclidean NN (position + color)	0.110	0.285	61.4
	Normalized Mahalanobis, raw $\mathbf{\Sigma}$ + Euclidean NN color	0.034	0.266	87.2
Kernel	Unnormalized Mahalanobis, raw $\mathbf{\Sigma}$ + Euclidean NN color	0.096	0.415	76.9
	Unnormalized Mahalanobis + fixed RBF color (σ_c=0.5)	0.102	0.422	75.8
Drift	+ representation ambiguity inflation $\mathbf{U}_i$ (§3.2.1)	0.258	0.536	51.9
	+ observation uncertainty inflation via FIM (§3.2.1)	0.465	0.611	23.9
	+ data-driven appearance bandwidth σ_c (§3.2.2)	0.537	0.629	14.6
	+ confidence weighting ω_i (§3.2.3)	**0.644**	**0.669**	**3.7**

Table 3: Disambiguation routing on PASLCD [14]. Balanced accuracy [4] is the primary metric; structural/surface-level precision and recall are reported per class. Existing baselines produce a single change score per pixel and cannot natively disambiguate change against structural vs surface-only; GS-DIFF's primitive-space comparison enables this and requires no annotations or auxiliary models.

Thresholding	Balanced Accuracy	Structural		Surface-level	
		Precision	Recall	Precision	Recall
Fixed (0.5)	0.868	0.970	0.961	0.725	0.774
Oracle	0.866	0.968	0.959	0.726	0.773

set, suppressing surface-level precision regardless of routing accuracy. The near-identical Fixed and Oracle results further indicate the routing is decoupled from detection threshold tuning, validating the disentangling claim of §3.2.3 – holding the geometric and appearance kernels separate operating directly in primitive space exposes change *type* as an emergent property, without supervision or external models. We show qualitative examples of this change classification in Appendix C.

5 Limitations

GS-DIFF treats reconstruction as an upstream black box: scoring quality is bounded by reconstruction quality, and pose errors from COLMAP propagate into both the kernel metric and the observability term. Geometrically accurate 3DGS is an active research area [5, 21, 24], and GS-DIFF benefits directly from advances there: cleaner geometric reconstructions sharpen the geometric kernel, sharpening the disambiguation routing. These limitations are inherited from the standard setup [14, 36] and are orthogonal to our primitive-space contribution, but they bound the regime in which the method is reliable. A second limitation lies in the photometric bandwidth σ_c, which absorbs color variation from both representation non-uniqueness and capture-condition offsets. This trade-off is favorable in inspection settings where capture-side changes are not of interest, but it could also absorb subtle surface-level changes. Inverse-rendering 3DGS variants [17, 6] that separate environment lighting and material albedo offer a natural remedy: comparing on albedo alone removes the capture-side component from σ_c and tightens the appearance kernel. We see this as a clean integration point for future work and orthogonal to the primitive-space contribution itself.

6 Conclusion

GS-DIFF performs multi-view SCD directly in primitive space, achieving state-of-the-art performance while unlocking the ability to reason about *what kind of* change occurred – separating structural from surface-level appearance changes without supervision or auxiliary models. This combination of accuracy and interpretability is especially valuable for long-term monitoring applications such as heritage preservation [8] and industrial asset management [14], where understanding why a region was flagged matters as much as flagging it. We see this work as the first step toward a new paradigm in SCD, and we hope it encourages the community to explore how far primitive-based change detection can push the reliability and capability of multi-view SCD.

Acknowledgment

This work was supported by the Australian Research Council Research Hub in Intelligent Robotic Systems for Real-Time Asset Management (ARIAM) (IH210100030) and Abyss Solutions. C.J., N.S., and D.M. also acknowledge ongoing support from the QUT Centre for Robotics.

References

[1] Jan Ackermann, Jonas Kulhanek, Shengqu Cai, Xu Haofei, Marc Pollefeys, Gordon Wetzstein, Leonidas Guibas, and Songyou Peng. CL-Splats: Continual learning of Gaussian splatting with local optimization. In *IEEE/CVF International Conference on Computer Vision*, 2025.

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

[3] Tim Alpherts, Sennay Ghebreab, and Nanne van Noord. EMPLACE: Self-supervised urban scene change detection. In *AAAI Conference on Artificial Intelligence*, pages 1737–1745, 2025.

[4] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition (ICPR)*, pages 3121–3124. IEEE, 2010.

[5] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. PGSR: Planar-based Gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6100–6111, 2024.

[6] HONGZE CHEN, Zehong Lin, and Jun Zhang. GI-GS: Global illumination decomposition on gaussian splatting for inverse rendering. In *The Thirteenth International Conference on Learning Representations*, 2025.

[7] Kyusik Cho, Dong Yeop Kim, and Euntai Kim. Zero-shot scene change detection. In *AAAI Conference on Artificial Intelligence*, pages 2509–2517, 2025.

[8] Mahtab Dahaghin, Myrna Castillo, Kourosh Riahidehkordi, Matteo Toso, and Alessio Del Bue. Gaussian heritage: 3D digitization of cultural heritage with integrated object segmentation. In *European Conference on Computer Vision Workshops*, 2024.

[9] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing*, pages 4063–4067, 2018.

[10] Iris de Gélis, Sébastien Lefèvre, and Thomas Corpetti. Siamese KPConv: 3D multiple change detection from raw point clouds using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197: 274–291, 2023.

[11] Almog Friedlander, Ariel Shamir, and Ohad Fried. GOLDILOCS: General object-level detection and labeling of changes in scenes. In *International Conference on Learning Representations*, 2026.

[12] Chamuditha Jayanga Galappaththige, Sanoojan Baliah, Malitha Gunawardhana, and Muhammad Haris Khan. Towards generalizing to unseen domains with few labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23691–23700, 2024.

[13] Chamuditha Jayanga Galappaththige, Zachary Izzo, Xilin He, Honglu Zhou, and Muhammad Haris Khan. Domain-guided weight modulation for semi-supervised domain generalization. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 6495–6505, 2025.

[14] Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald G. Dansereau, Niko Suenderhauf, and Dimity Miller. Multi-view pose-agnostic change localization with zero labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[15] Chamuditha Jayanga Galappaththige, Thomas Gottwald, Peter Stehr, Edgar Heinert, Niko Suenderhauf, Dimity Miller, and Matthias Rottmann. Predictive photometric uncertainty in gaussian splatting for novel view synthesis. *arXiv preprint arXiv:2603.22786*, 2026.

[16] Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald G. Dansereau, Niko Suenderhauf, and Dimity Miller. Changes in real time: Online scene change detection with multi-view fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[17] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024.

[18] Kyle Gao, Dening Lu, Liangzhi Li, Nan Chen, Hongjie He, Linlin Xu, and Jonathan Li. Digital buildings analysis: 3D modeling, GIS integration, and visual descriptions using Gaussian splatting, ChatGPT/DeepSeek, and Google maps platform. *IEEE Geoscience and Remote Sensing Letters*, 2025.

[19] Joachim Gehrung, Marcus Hebel, Michael Arens, and Uwe Stilla. A fast voxel-based indicator for change detection using low resolution octrees. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:357–364, 2019.

[20] Daniel Girardeau-Montaut, Michel Roux, Raphaël Marc, and Guillaume Thibault. Change detection on point cloud data acquired with a ground laser scanner. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36, 2005.

[21] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5354–5363, 2024.

[22] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[23] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[24] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024.

[25] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[26] Binbin Jiang, Rui Huang, Qingyi Zhao, and Yuxiang Zhang. Gaussian difference: Find any change instance in 3D scenes. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5, 2025.

[27] Shyam Sundar Kannan and Byung-Cheol Min. ZeroSCD: Zero-shot street scene change detection. In *IEEE International Conference on Robotics and Automation*, pages 4665–4671, 2025.

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

[30] Jae-Woo Kim and Ue-Hwan Kim. Towards generalizable scene change detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24463–24473, 2025.

[31] Mathis Kruse, Marco Rudolph, Dominik Woiwode, and Bodo Rosenhahn. SplatPose & detect: Pose-agnostic 3D anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3950–3960, 2024.

[32] Dimitri Lague, Nicolas Brodu, and Jérôme Leroux. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (n-z). *ISPRS Journal of Photogrammetry and Remote Sensing*, 82:10–26, 2013.

[33] Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. Robust scene change detection using visual foundation models and cross-attention mechanisms. In *IEEE International Conference on Robotics and Automation*, pages 8337–8343, 2025.

[34] Yizhe Liu, Yan Song Hu, Yuhao Chen, and John Zelek. SplatPose+: Real-time image-based pose-agnostic 3D anomaly detection. In *European Conference on Computer Vision Workshops*, pages 378–391, 2024.

[35] Samuel Looper, Javier Rodriguez-Puigvert, Roland Siegwart, Cesar Cadena, and Lukas Schmid. 3D VSG: Long-term semantic scene change prediction through 3D variable scene graphs. In *IEEE International Conference on Robotics and Automation*, pages 8179–8186, 2023.

[36] Ziqi Lu, Jianbo Ye, and John Leonard. 3DGS-CD: 3D Gaussian splatting-based change detection for physical object rearrangement. *IEEE Robotics and Automation Letters*, 2025.

[37] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024.

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[39] Yue Qiu, Shintaro Yamamoto, Ryosuke Yamada, Ryota Suzuki, Hirokatsu Kataoka, Kenji Iwata, and Yutaka Satoh. 3d change localization and captioning from dynamic scans of indoor scenes. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1176–1185, 2023.

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[41] Shiwei Ren, Tianci Wen, Yongchun Fang, and Biao Lu. FastGS: Training 3D Gaussian splatting in 100 seconds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[42] Ragav Sachdeva and Andrew Zisserman. The change you want to see. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3993–4002, 2023.

[43] Ragav Sachdeva and Andrew Zisserman. The change you want to see (now in 3D). In *IEEE/CVF International Conference on Computer Vision Workshops*, pages 2060–2069, 2023.

[44] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using CNN features and superpixel segmentation. In *British Machine Vision Conference*, pages 61.1–61.12, Swansea, 2015. British Machine Vision Association.

[45] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[46] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

[47] Ashley Varghese, Jayavardhana Gubbi, Akshaya Ramaswamy, and P Balamuralidhar. ChangeNet: A deep learning architecture for visual change detection. In *European Conference on Computer Vision Workshops*, 2018.

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] Zi Jian Yew and Gim Hee Lee. City-scale scene change detection using point clouds. In *IEEE International Conference on Robotics and Automation*, pages 13362–13369, 2021.

[50] Zirui Zhou, Junfeng Ni, Shujie Zhang, Yixin Chen, and Siyuan Huang. 3D scene change modeling with consistent multi-view aggregation. In *International Conference on 3D Vision*, 2026.

[51] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. EWA splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(03):223–238, 2002.

A GS-DIFF Algorithm

We provide pseudo-code for the full GS-DIFF pipeline (Algorithm 1). Reconstruction is treated as upstream input: we assume two independently built 3DGS reconstructions $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$ with COLMAP-estimated camera poses $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ registered to a common frame, frustum-filtered to the shared observable region (§3.2). GS-DIFF then proceeds in four stages: geometric drift modeling (§3.2.1), kernel scoring on geometry and appearance (§3.2.1, §3.2.2), aggregation and rendering (§3.2.3), and disambiguation (§3.2.3). All quantities derived from inter-reconstruction statistics – the representation-ambiguity scales u_n, u_t, the appearance bandwidth σ_c – are computed bidirectionally ($\mathcal{G}^{(1)} \to \mathcal{G}^{(2)}$ and $\mathcal{G}^{(2)} \to \mathcal{G}^{(1)}$) and averaged for symmetry; per-primitive quantities ($\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}}$, ω_i, δ_i^{geo}, δ_i^{app}) are computed once per primitive in each reconstruction.

Algorithm 1 GS-DIFF: Primitive-Space Scene Change Detection

Require: Reconstructions $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$ with poses $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}$ in a common frame, frustum-filtered to the shared observable region
Ensure: Binary change mask $\mathcal{M}$; per-pixel structural/surface labels

Stage 1: Geometric drift modeling (§3.2.1)
for $s \in \{1, 2\}$, primitive $i \in \mathcal{G}^{(s)}$ **do**
  $j^* \leftarrow$ Euclidean nearest neighbor of i in $\mathcal{G}^{(\bar{s})}$
  Compute $d_n^{(i)}, d_t^{(i)}$ via Eq. 3
  Compute $\mathbf{H}_i$ over visible cameras $\mathcal{C}_i^{\text{vis}}$ ▷ Eq. 4
end for
$u_n^2, u_t^2 \leftarrow$ symmetric average of $Q_{0.75}$ over $\mathcal{G}^{(1)} \to \mathcal{G}^{(2)}$ and $\mathcal{G}^{(2)} \to \mathcal{G}^{(1)}$
for $s \in \{1, 2\}$, primitive $i \in \mathcal{G}^{(s)}$ **do**
  $\tilde{\boldsymbol{\Sigma}}_i \leftarrow \boldsymbol{\Sigma}_i + u_t^2 \mathbf{I}_3 + (u_n^2 - u_t^2)\mathbf{n}_i \mathbf{n}_i^\top$ ▷ Eq. 3
  $\alpha^{(s)} \leftarrow$ FIM-injection scale on $\mathcal{G}^{(s)}$ ▷ numerator of Eq. 5
  $\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}} \leftarrow \tilde{\boldsymbol{\Sigma}}_i + \alpha^{(s)} \mathbf{H}_i^+$ ▷ Eq. 5
end for

Stage 2: Kernel scoring (§3.2.1, §3.2.2)
$\sigma_c^2 \leftarrow$ symmetric weighted-median appearance bandwidth ▷ Eq. 9
for $s \in \{1, 2\}$, primitive $i \in \mathcal{G}^{(s)}$ **do**
  $\mathcal{N}(i) \leftarrow \{j \in \mathcal{G}^{(\bar{s})} : \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \le 3\sqrt{\lambda_{\max}(\tilde{\boldsymbol{\Sigma}}_i^{\text{eff}})}\}$ ▷ Eq. 6
  $\delta_i^{\text{geo}} \leftarrow 1 - \max_{j \in \mathcal{N}(i)} k_{\text{geo}}(g_i, g_j)$ ▷ Eq. 8
  $\sigma_{c,i}^2 \leftarrow \sigma_c^2 \cdot \max(h_i^2 / \tilde{h}^2, 1)$ ▷ Eq. 10
  $\delta_i^{\text{app}} \leftarrow 1 - \max_{j \in \mathcal{N}(i)} k_{\text{app}}(g_i, g_j)$ ▷ Eq. 12
end for

Stage 3: Aggregation and rendering (§3.2.3)
for $s \in \{1, 2\}$, primitive $i \in \mathcal{G}^{(s)}$ **do**
  $\omega_i \leftarrow \sigma\big(\log \operatorname{tr}(\mathbf{H}_i) - \log Q_{0.25}(\{\operatorname{tr}(\mathbf{H}_k)\}_{k \in \mathcal{G}^{(s)}})\big)$ ▷ Eq. 13
  $\delta_i^{(s)} \leftarrow \omega_i \cdot \min(\delta_i^{\text{geo}} + \delta_i^{\text{app}}, 1)$ ▷ Eq. 14
end for
$\mathcal{M}^{(s)} \leftarrow$ ALPHACOMPOSITERENDER$(\mathcal{G}^{(s)}, \delta_i^{(s)})$ for $s \in \{1, 2\}$
$\mathcal{M} \leftarrow \max(\mathcal{M}^{(1)}, \mathcal{M}^{(2)})$, binarized at 0.5 ▷ final change map

Stage 4: Disambiguation (§3.2.3)
$\mathcal{M}^{\text{struct}} \leftarrow$ ALPHACOMPOSITERENDER$(\mathcal{G}, \delta_i^{\text{geo}})$
$\mathcal{M}^{\text{surf}} \leftarrow$ ALPHACOMPOSITERENDER$(\mathcal{G}, \max(\delta_i^{\text{app}} - \delta_i^{\text{geo}}, 0))$
for each pixel p with $\mathcal{M}_p > 0.5$ **do**
  label$(p) \leftarrow \arg\max(\mathcal{M}_p^{\text{struct}}, \mathcal{M}_p^{\text{surf}})$ ▷ Eq. 15
end for

Implementation notes. Nearest-neighbor and ball queries use spatial KD-trees, giving $O(N \log N)$ per scene where N is the primitive count. The FIM accumulation in Eq. 4 is per-primitive over its frustum-visible cameras and is batched across primitives. The two final renderings ($\mathcal{M}^{(s)}$ for the change map; $\mathcal{M}^{\text{struct}}, \mathcal{M}^{\text{surf}}$ for disambiguation) reuse the standard 3DGS rasterizer [29] with $\delta_i^{(s)}$,

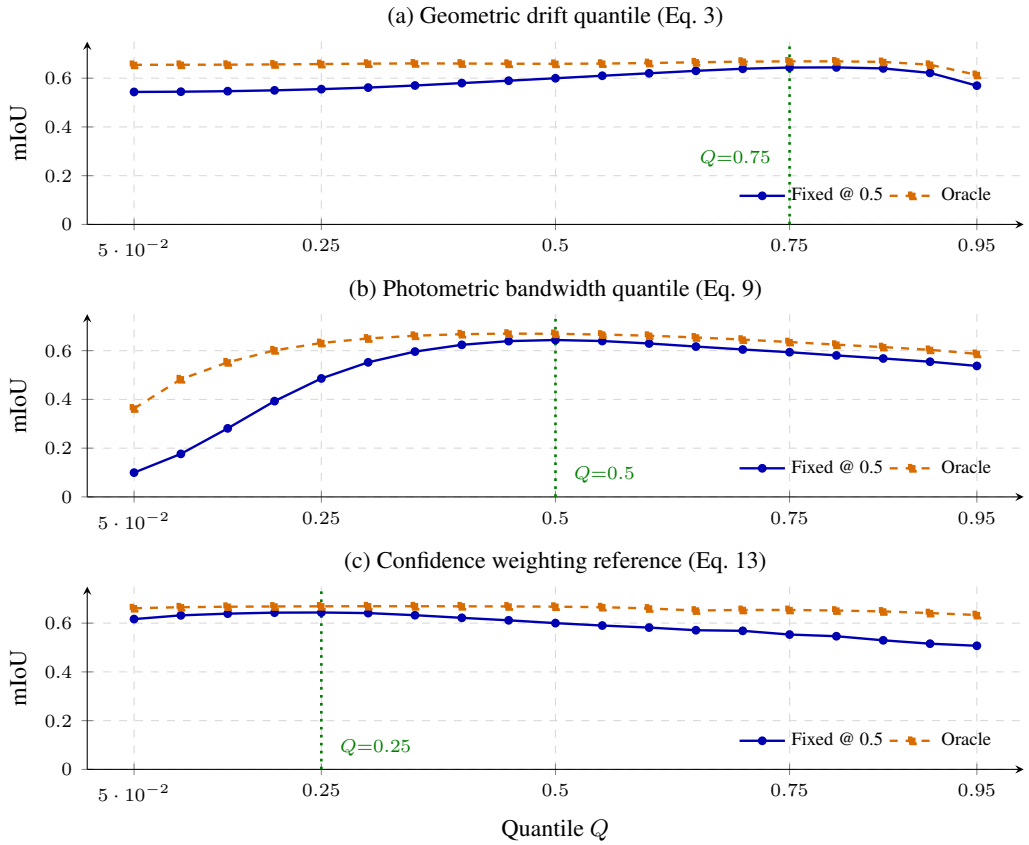

Figure 4: Sensitivity of GS-DIFF to data-driven quantile choices: (a) representation-ambiguity drift scales, (b) appearance bandwidth, (c) observability-weighting reference. We sweep each quantile from 0.05 to 0.95 in 0.05 steps on PASLCD; the green dotted line marks our *a priori* principled choice (upper quartile, median, lower quartile, respectively). All three choices (green dotted line) sit within 0.01 mIoU of the empirical Fixed-threshold optimum on PASLCD; Oracle mIoU is largely insensitive to quantile value (except for photometric bandwidth at $Q < 0.3$), suggesting quantile choice primarily affects calibration rather than underlying separability.

δ_i^{geo}, and $\max(\delta_i^{\text{app}} - \delta_i^{\text{geo}}, 0)$ substituted for the per-primitive color in alpha-compositing. We use PGSR [5] to build geometrically accurate reconstructions $\mathcal{G}^{(1)}, \mathcal{G}^{(2)}$. All experiments are conducted on a single NVIDIA RTX4090 24GB GPU.

B Analysis on Data-Driven Quantiles

GS-DIFF computes three data-driven quantities through quantiles of empirical distributions: the geometric drift scales u_n, u_t as the upper quartile $Q_{0.75}$ of nearest-neighbor displacements (Eq. 3); the photometric bandwidth σ_c as the median ($Q_{0.50}$) of weighted color differences (Eq. 9); and the confidence-weighting reference as the lower quartile $Q_{0.25}$ of FIM traces (Eq. 13). These three quantiles – upper, median, lower – correspond to the form of Tukey's standard five-number descriptive-statistics summary of a distribution [46]: the upper quartile bounds the bulk of typical drift while excluding the tail of genuinely changed regions; the median is the canonical robust location estimate; the lower quartile isolates the bottom-quartile tail of poorly observed primitives. We chose these three values *a priori* on these statistical grounds.

To ablate the sensitivity of performance to these values, we sweep each quantile from 0.05 to 0.95 in steps of 0.05 on PASLCD, holding all other components fixed. Figure 4 reports Fixed-threshold mIoU (primary, threshold 0.5) and Oracle mIoU (per-scene optimal threshold) for each sweep. In all three cases, the Oracle mIoU remains generally stable across quantile values (with the exception of

the photometric bandwidth quantiles below 0.3, where performance decrements occur). In contrast, the Fixed-threshold performance shows some sensitivity to quantile value, indicating that quantile selection primarily affects the calibration of change scores against a fixed threshold rather than the underlying separability.

As shown in Figure 4, our quantile selections sit at or within 0.01 mIoU of the empirical optimum within fairly stable plateaus: the geometric drift quantile plateaus between $Q_{0.70}$ and $Q_{0.85}$; the photometric bandwidth quantile plateaus between $Q_{0.45}$ and $Q_{0.55}$; finally the confidence weighting reference quantile plateaus between $Q_{0.15}$ and $Q_{0.30}$. This suggests that the performance of GS-DIFF has some robustness to exact quantile value selection, provided that the general logic (described above) is followed.

C Additional Visualizations of Change Maps

We provide additional visualizations of our final change maps and the per-scene maps generated for each reference and inference scenes separately in Figure 5.Figure 6 provides a qualitative comparison against the strongest image-space baseline, O-SCD [16]. Figure 7 visualizes the per-kernel scores and the residual that drives the structural-versus-surface-only routing of §3.2.3.

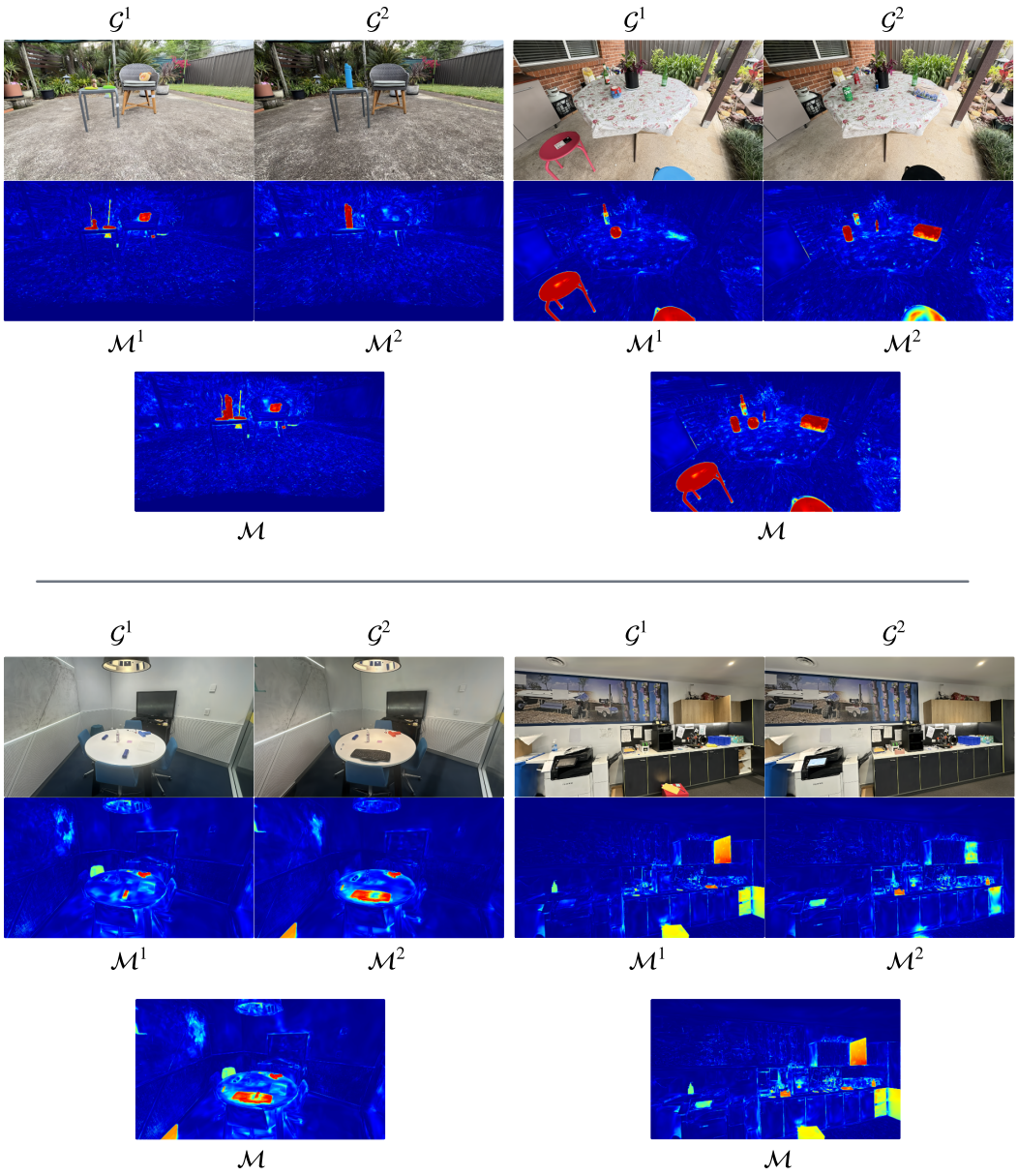

Figure 5: Qualitative results on PASLCD [14]. For each scene: rendered views from reference $\mathcal{G}^{(1)}$ (top-left), inference $\mathcal{G}^{(2)}$ (top-right), and change map for each individual scene $\mathcal{M}^1$ (bottom - left), $\mathcal{M}^2$ (bottom - right), and final change map $\mathcal{M}$ (bottom - center). GS-DIFF produces sharply localized responses to both structural and surface-level changes across indoor and outdoor scenes. Our bidirectional scoring of change (§3.2.3) allows us to generate change maps for each reference and inference scenes individually.

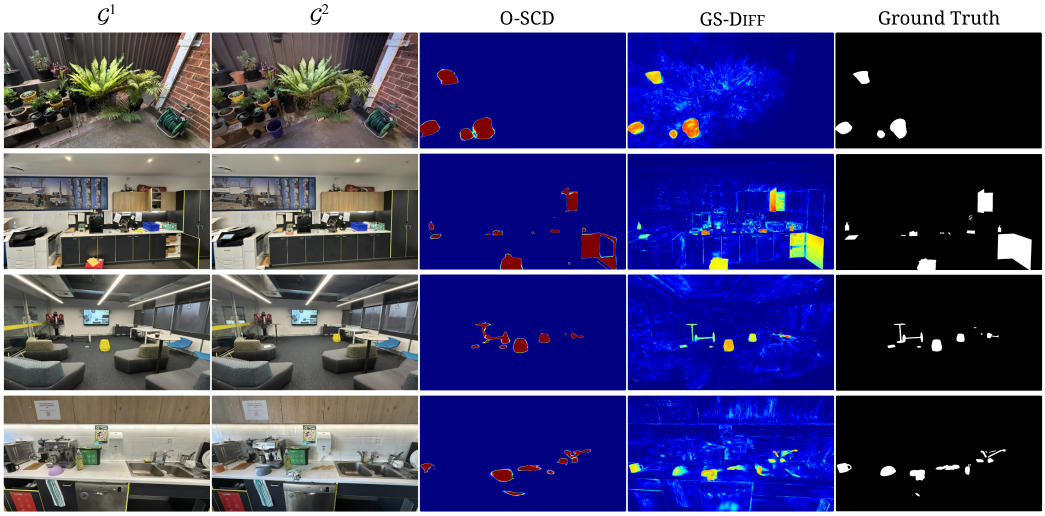

Figure 6: Qualitative comparison with O-SCD [16], the strongest image-space baseline. GS-DIFF produces more geometrically accurate change masks for two reasons. (1) *Patch granularity:* O-SCD and other image-space approaches compare features from an external foundation model [40] at its patch size (14×14 or 16×16 in pixels), which sets a floor on spatial resolution; primitive-space comparison is limited only by the fidelity of the representation itself. (2) *Multi-view aggregation:* O-SCD's learned multi-view objective converges to a consensus across views, smoothing the boundary in any single view; GS-DIFF's per-primitive scores are multi-view consistent by construction and preserve view-specific sharpness. The bottom row additionally shows our kernels detecting a surface-level change between semantically similar objects (bowl recolored pink-to-blue), which O-SCD misses, as foundation-model features are largely invariant to such fine appearance changes in semantically similar objects [16, 14]. O-SCD's multi-view learning objective drives its outputs toward binary scores 0 or 1, producing the bimodal red/blue maps shown above; GS-DIFF's scores arise from kernel evaluations and remain continuous, exposing graded confidence and enabling the threshold-stable behavior reported in §4.1.

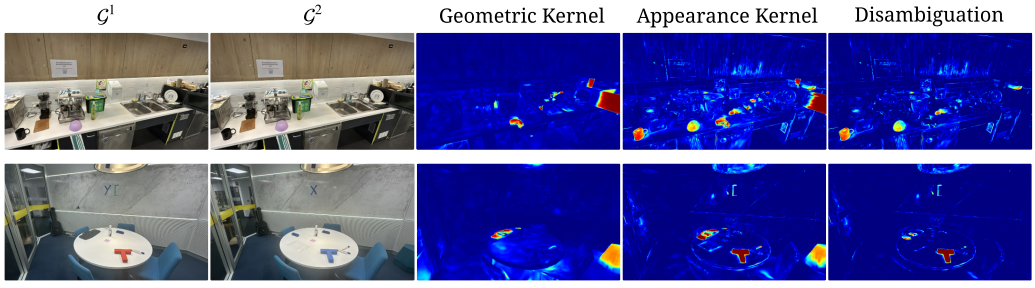

Figure 7: Qualitative results on kernel scores and disambiguation routing (§3.2.3). Left to right: rendered views from reference $\mathcal{G}^1$ and inference $\mathcal{G}^2$ scenes: geometric kernel change score δ^{geo}; appearance kernel change score δ^{app}; and the residual $\max(\delta_i^{\text{app}} - \delta_i^{\text{geo}}, 0)$ used in structural versus sufrace-only disambiguation§3.2.3. Note that residual has a high score in surface only changes (Row 1: color change in mug, bowl, tea pack on the oven and coffee spill, Row 2: Color change in T shape object and drawing on the whiteboard), whereas geometric change scores are minimal in these regions.