# GAUSSLIFT: RENDERER-CONSISTENT DSM EXPORT FROM SATELLITE GAUSSIAN FIELDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

3D Gaussian Splatting (3DGS) can reproduce satellite imagery with high photometric fidelity, yet a learned Gaussian field does not uniquely determine a metrically accurate digital surface model (DSM). The ambiguity arises at the representation–measurement interface: primitive centers, projected footprints, and renderer-visible support can induce different elevation surfaces from the same field. We introduce **GaussLift**, which makes this interface explicit by reusing front-to-back alpha-compositing weights to define a normalized renderer-conditioned altitude observable, then localizing supported samples in world coordinates and realizing them on the DSM grid. The same observable is used for surface supervision and export, without introducing an auxiliary DSM decoder. Fixed-field controls on DFC2019 confirm that changing only the export operator alters both reconstruction error and coverage; under a common-mask comparison across frozen fields, renderer-conditioned altitude attains the lowest MAE and RMSE among the tested readouts. Across four scene-specific reconstructions, the final 12-member map-space ensemble achieves 0.787 m raw MAE, 2.344 m RMSE, and 95.91% coverage. The underlying Gaussian fields remain renderable from 68 held-out views without target-view fitting, reaching 22.562 dB mean PSNR. These results show that surface readout is a consequential component of metric 3D Gaussian reconstruction, complementary to representation learning.

## 1 INTRODUCTION

Satellite neural rendering reconstructs appearance-bearing 3D scenes from sparse orbital imagery (Xie et al., 2023; Marí et al., 2023; Billouard et al., 2024). Gaussian methods add explicit primitives and efficient rasterization (Aira et al., 2025; Bai et al., 2025; Bournez et al., 2025), extending reconstruction toward geospatial measurement. A central product is the digital surface model (DSM), which records world-referenced height on a regular map grid (Panagiotou et al., 2020; Mou & Zhu, 2018; Du et al., 2023). Yet Gaussian scenes and DSMs encode geometry differently: image formation depends on anisotropic support, opacity, and ordered visibility, whereas a DSM stores one height per valid cell.

A Gaussian field can therefore explain images without uniquely defining its DSM surface. Primitive centers may lie away from visible surfaces, projected footprints may overlap height layers, and occlusion changes their contributions. Multiple world-localized samples may also enter the same map cell. Writing the exported surface as $\hat{H}^{(k)} = \mathcal{E}_k(\mathcal{G}; \Pi, \Omega_H)$ makes this ambiguity explicit: $\mathcal{G}$ is the field, $\Pi$ is calibrated observation geometry, $\Omega_H$ is the target grid, and $\mathcal{E}_k$ is the export rule. Different rules can induce different DSMs from the same $\mathcal{G}$ (Figure 1).

We call this mismatch the *rendering-to-mapping gap*. Direct image-to-DSM methods learn the raster product (Chen et al., 2023b;a; 2024), whereas satellite neural-field and Gaussian methods optimize a scene through image formation. Their scene-to-DSM conversion is often embedded in depth rendering, primitive filtering, or rasterization (Bai et al., 2025; Bournez et al., 2025; Huang et al., 2026), obscuring whether measured geometry arises from the representation or its readout.

GaussLift makes this interface explicit through three stages: renderer-consistent surface attribution, calibrated world localization, and deterministic map-space realization. RGB, support, and altitude share front-to-back compositing weights, so geometry suppressed by the renderer cannot reappear



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

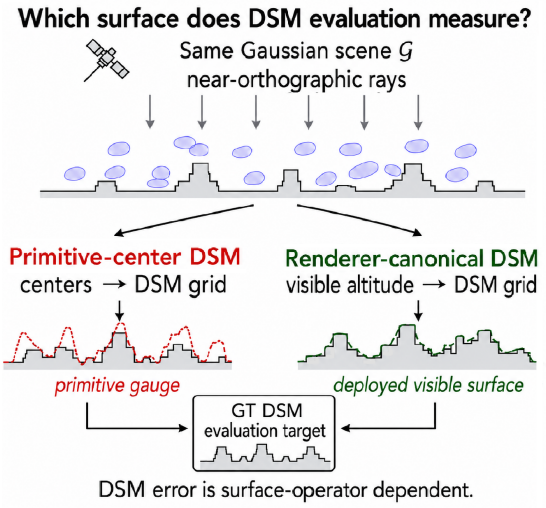


Figure 1: The rendering-to-mapping gap. A fixed satellite Gaussian scene can induce different DSMs under primitive-center and renderer-visible readouts. The export operator determines which surface is compared with the reference DSM.

through an unrelated DSM readout. We isolate export behavior by freezing the optimized Gaussian field and varying only the exporter. The protocol uses full-scene DSM supervision and measures scene-specific reconstruction, not unseen-site prediction; held-out rendering assesses the field's retained appearance capability.

Our contributions are threefold: (1) identifying Gaussian-to-DSM ambiguity under a fixed field; (2) introducing shared visibility attribution followed by calibrated localization and deterministic map realization; and (3) evaluating export semantics separately from Gaussian optimization, alongside DSM reconstruction and retained rendering capability.

## 2 Related Work

### 2.1 Single-Image DSM Estimation

Single-image DSM methods directly predict a height raster. Early approaches use dense regression or classification–regression (Panagiotou et al., 2020; Liu et al., 2020; Mou & Zhu, 2018); later work adds multiscale interaction, semantic guidance, and structure-aware learning (Du et al., 2023; Chen et al., 2023b;a). Transformer and feature-fusion models improve global context and local recovery (Chen et al., 2024; Chen & Zhu, 2025; Ma et al., 2024), with recent work exploring multimodal and object-aware cues (Gültekin et al., 2025; Peng et al., 2025; Memar et al., 2025). These methods define the surface directly through DSM supervision. GaussLift instead exports a georeferenced DSM from a renderable Gaussian scene; published image-to-DSM results provide numerical context, not protocol-equivalent comparisons.

### 2.2 Satellite Neural Rendering and Gaussian Reconstruction

Satellite neural rendering adapts radiance fields to rational polynomial coefficient (RPC) cameras, multi-date imagery, and illumination or transient variation (Marí et al., 2023; Xie et al., 2023; Billouard et al., 2024). Further work targets efficiency, sparse-view depth supervision, implicit surfaces, and physically based appearance (Zhou et al., 2024; Zhang & Rupnik, 2023; Qu & Deng, 2023; Zhang et al., 2025).

General 3D Gaussian Splatting provides an explicit, efficient representation (Kerbl et al., 2023), subsequently extended for antialiasing, accurate geometry, surface-aligned mesh extraction, and compactness (Yu et al., 2024; Huang et al., 2024; Guédon & Lepetit, 2024; Lee et al., 2024). Satellite Gaussian methods address photogrammetry, temporal appearance, shadows, and sparse views (Aira

et al., 2025; Bai et al., 2025; Huang et al., 2026; Luo et al., 2026; Bournez et al., 2025). Their elevation outputs depend on combinations of compositing, filtering, projection, and rasterization. Native-RPC rendering analyzes camera-induced error (Wagner et al., 2026), while Sat-DN extracts an SDF surface before DSM generation (Liu et al., 2025). GaussLift isolates the export procedure while holding the scene fixed.

### 2.3 Rendered Geometry and Surface Supervision

Opacity-weighted depth and feature moments are common renderer outputs, and surface-aligned Gaussian and satellite neural-field methods demonstrate the value of geometry-aware supervision (Huang et al., 2024; Guédon & Lepetit, 2024; Zhang & Rupnik, 2023). However, a ray-space statistic alone does not specify a georeferenced DSM. Weighted moments, terminal opacity, weighted medians, soft-argmax, SDF, and TSDF readouts encode different surface choices; support selection, localization, and map-cell conflict resolution further alter the raster. GaussLift makes this interface explicit rather than treating export as a neutral geometric readout.

## 3 Method

### 3.1 Problem Formulation and Operator Interface

A satellite Gaussian field is a visibility-bearing volumetric representation, whereas a DSM is a single-valued geospatial surface. Figure 2 summarizes the GaussLift pipeline. Let

$$\mathcal{G} = \{g_i = (\mu_i, \Sigma_i, \alpha_i, c_i)\}_{i=1}^{N_g} \tag{1}$$

denote a field of Gaussian centers, covariances, opacities, and colors. We define Gaussian-to-DSM export as

$$(\hat{H}, \hat{M}_H) = \mathcal{E}(\mathcal{G}, \Pi, T_w, \Omega_H, \eta, \Gamma), \tag{2}$$

where $\Pi$ is the observation geometry, $T_w$ maps the scene frame to world coordinates, $\Omega_H$ is the target grid, $\eta$ is a support threshold, and $\Gamma$ is a deterministic cell operator. GaussLift imposes an occlusion-attribution principle: a primitive contributes height only through the ordered weight assigned to it during image formation. Hence, a primitive suppressed by front-to-back compositing cannot re-enter the DSM through independent center or footprint rasterization. The resulting operator makes surface attribution, geospatial localization, and map realization explicit and auditable.

### 3.2 Renderer-Consistent Ray-Space Attribution

For pixel $p \in \Omega_I$, Gaussian $i$ projects to mean $\bar{p}_i = \Pi(\mu_i)$ and covariance

$$\Sigma_{i,2D} = J_i \Sigma_i J_i^\top + \sigma_{\min}^2 I, \qquad J_i = \left.\frac{\partial \Pi}{\partial X}\right|_{\mu_i}. \tag{3}$$

Its footprint response is

$$d_i(p) = \exp\left[-\tfrac{1}{2}(p - \bar{p}_i)^\top \Sigma_{i,2D}^{-1}(p - \bar{p}_i)\right]. \tag{4}$$

A local viewing direction $r_\Pi$ is obtained from the RPC at the scene center. The null-space direction $v$ of the local affine image Jacobian is signed by $v^\top r_\Pi > 0$, and $z_i^\Pi = v^\top \mu_i$ defines front-to-back ordering. With $a_i(p) = \alpha_i d_i(p)$, ordered compositing gives

$$T_i(p) = \prod_{j<i}(1 - a_j(p)), \qquad w_i(p) = T_i(p) a_i(p). \tag{5}$$

The same weights define appearance and surface evidence:

$$\hat{I}(p) = \sum_{i=1}^{N_g} w_i(p) c_i + T_{N_g+1}(p) c_{\mathrm{bg}}, \quad \hat{M}(p) = \sum_{i=1}^{N_g} w_i(p), \tag{6}$$

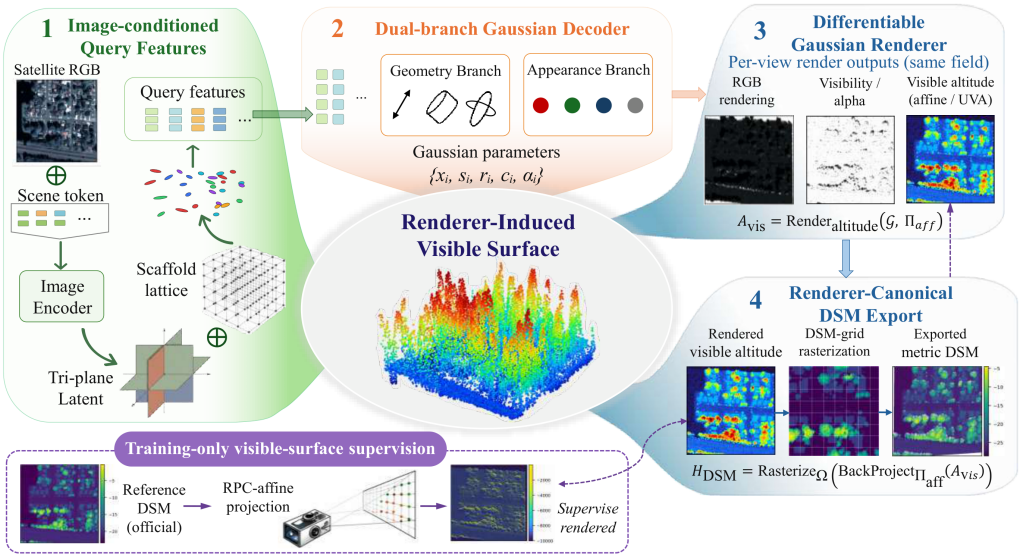


Figure 2: GaussLift pipeline. A dual-branch decoder constructs the Gaussian scene. Shared rendering weights define appearance and altitude, followed by world localization and DSM-grid realization. Reference DSM supervision is training-only.

where $T_{N_g+1}(p) = \prod_i (1 - a_i(p))$. Their normalized form defines the renderer-conditioned altitude moment

$$\pi_i(p) = \frac{w_i(p)}{\hat{M}(p) + \epsilon}, \qquad \hat{A}_\Pi(p) = \sum_{i=1}^{N_g} \pi_i(p) h_i^\Pi, \tag{7}$$

with $h_i^\Pi$ the altitude coordinate in the calibrated scene frame. Attribution consistency follows directly:

$$w_i(p) = 0 \quad \Longrightarrow \quad \pi_i(p) = 0. \tag{8}$$

Thus, RGB and altitude cannot select contradictory primitive support. A sample is retained only when

$$\hat{M}(p) > \eta, \tag{9}$$

which controls the accuracy–coverage trade-off. The weights also define ray-wise dispersion

$$\sigma_A^2(p) = \sum_i \pi_i(p) \big(h_i^\Pi - \hat{A}_\Pi(p)\big)^2, \tag{10}$$

which identifies separated contributing layers. The first moment is smooth and RGB-coupled, but may lie between partially visible layers; terminal opacity, weighted median, and soft-argmax therefore constitute alternative ray-space semantics.

### 3.3 Georeferenced Localization and Map Realization

Let $\Phi_\Pi(p, h) = X^w$ localize an image–altitude pair using the RPC-derived local affine camera and fixed scene-to-world transform. For each supported pixel,

$$X_p^w = \Phi_\Pi\big(p, \hat{A}_\Pi(p)\big), \qquad \hat{M}(p) > \eta. \tag{11}$$

Localization is therefore applied after renderer visibility resolves the ray-space height evidence. Across four scenes, the affine fit has 0.012-pixel mean reprojection error and 0.044-pixel worst-scene P95 error. For $X_p^w = (x_p^w, y_p^w, h_p^w)$, the target-grid index is

$$u_x = \left\lfloor \frac{x_p^w - x_0}{\Delta x} \right\rfloor, \qquad u_y = \left\lfloor \frac{y_0 - y_p^w}{\Delta y} \right\rfloor, \tag{12}$$

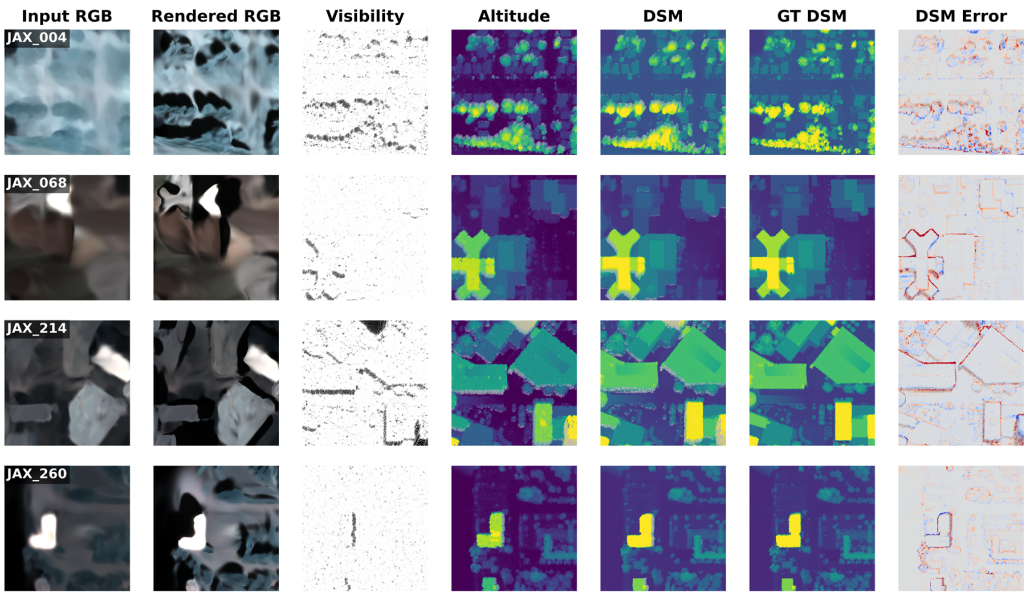


Figure 3: Qualitative GaussLift outputs on four JAX scenes. Each row shows the conditioning RGB image, rendered appearance, accumulated support, renderer-conditioned altitude, exported DSM, reference DSM, and signed error map. Mean and per-scene metrics are reported in Tables 1 and 2.

where $(x_0, y_0)$ and $(\Delta x, \Delta y)$ are the raster origin and spacing. The supported heights in cell $u$ are

$$\mathcal{S}(u) = \{h_p^w \mid \hat{M}(p) > \eta, \mathrm{Grid}(X_p^w) = u\}. \tag{13}$$

The DSM and validity mask are

$$\hat{H}(u) = \Gamma(\mathcal{S}(u)), \qquad \hat{M}_H(u) = \mathbf{1}[\mathcal{S}(u) \neq \emptyset]. \tag{14}$$

The base realization uses the top envelope

$$\Gamma_{\max}(\mathcal{S}(u)) = \max \mathcal{S}(u), \tag{15}$$

The complete interface is

$$\mathcal{E}_{\mathrm{GL}} = \Gamma_{\Omega_H} \circ \Phi_\Pi \circ \mathcal{R}_\Pi, \tag{16}$$

where $\mathcal{R}_\Pi$ is defined by Equations (3)–(9). This decomposition prevents surface selection, camera localization, and raster aggregation from being conflated into an unspecified readout.

### 3.4 Scene Construction and Surface-Aligned Optimization

The reference DSM $H^* : \Omega_H \to \mathbb{R}$ is projected into the conditioning view:

$$(A_\Pi^*, M_\Pi^*) = \mathrm{ProjectAlt}_\Pi(H^*). \tag{17}$$

The renderer-conditioned moment is supervised in the same camera-aligned system:

$$\mathcal{L}_{\mathrm{alt}} = \frac{\sum_{p \in \Omega_I} M_\Pi^*(p) \rho_{\mathrm{SL1}}\big(\hat{A}_\Pi(p) - A_\Pi^*(p)\big)}{\sum_{p \in \Omega_I} M_\Pi^*(p) + \epsilon}. \tag{18}$$

This aligns the optimized statistic with the quantity exported by GaussLift. The full objective is

$$\mathcal{L} = \lambda_{\mathrm{rgb}}\mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{alt}}\mathcal{L}_{\mathrm{alt}} + \lambda_{\mathrm{grad}}\mathcal{L}_{\mathrm{grad}} + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{ssi}}\mathcal{L}_{\mathrm{ssi}}, \tag{19}$$

where the auxiliary terms constrain appearance, local variation, structural correlation, and scale-and-shift-invariant geometry. Each scene is represented by a Gaussian field initialized from 60,000 anchors and optimized for 1000 Adam epochs. A frozen image backbone and trainable tri-plane generator predict Gaussian attributes, while surface-near anchor replacement improves geometric support. Held-out rendering changes only the projection camera. We report both a single-field export and a 12-member cell-wise median ensemble to separate representation quality from aggregation gains.

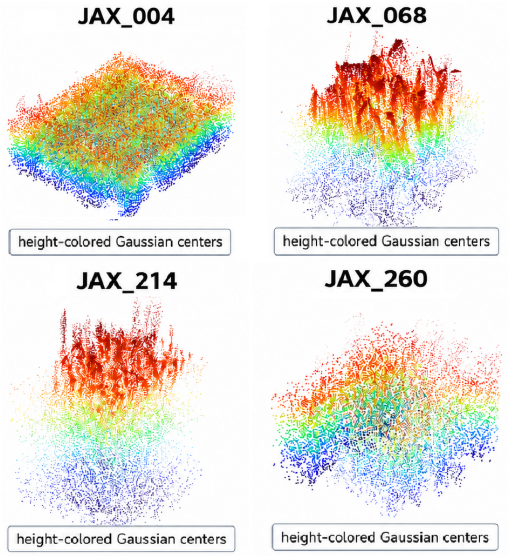


Figure 4: Height-colored Gaussian fields for the four JAX scenes. Centers visualize the learned 3D structure; DSM heights are obtained through renderer-conditioned surface attribution.

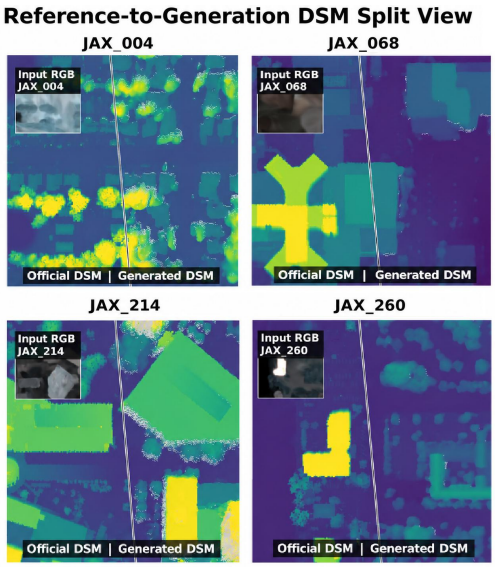


Figure 5: Qualitative DSM reconstruction on four DFC2019 JAX scenes. Each panel compares the official and GaussLift DSMs with the conditioning RGB image inset.

Table 1: Protocol-separated comparison on four DFC2019 JAX scenes. Published cross-scene results provide context; reimplementations and GaussLift use matched target-scene fitting with full DSM supervision. 3D: explicit scene; DSM: full-scene DSM optimization; TTO: target-view test-time optimization; G+R+D: Gaussian scene, RGB rendering, and DSM.

| Method | 3D | DSM | TTO | MAE↓ | RMSE↓ | PSNR↑ | SSIM↑ | Output |
|---|---|---|---|---|---|---|---|---|
| HTC-DC-Net | × | × | × | 6.34 | 9.19 | 15.31 | 0.03 | DSM |
| IM2ELEVATION (SENet) | × | × | × | 2.61 | 3.75 | 22.53 | 0.85 | DSM |
| IMG2HEIGHT (SM-EDNet) | × | × | × | 2.14 | 3.34 | 23.99 | 0.88 | DSM |
| ImageToDEM (pix2pix) | × | × | × | 1.18 | 2.11 | 27.60 | 0.96 | DSM |
| TSE-Net | × | × | × | 1.61 | 2.79 | 24.94 | 0.91 | DSM |
| CMT reimpl. | × | ✓ | × | 1.007 | 2.672 | 25.935 | 0.950 | DSM |
| HeightFormer reimpl. | × | ✓ | × | 1.004 | 2.545 | 26.074 | 0.953 | DSM |
| LUMNet reimpl. | × | ✓ | × | 0.955 | 2.545 | 26.238 | 0.953 | DSM |
| GaussLift (single field) | ✓ | ✓ | × | 0.907 | 2.585 | 26.141 | 0.951 | G+R+D |
| GaussLift (12-member ensemble) | ✓ | ✓ | × | **0.787** | **2.344** | **27.078** | **0.960** | G+R+D |

# 4 Experiments

## 4.1 Experimental Setup

We evaluate GaussLift on four DFC2019 (Le Saux et al., 2019) Jacksonville scenes: JAX_004, JAX_068, JAX_214, and JAX_260. A separate Gaussian field is optimized for each scene using one RGB image and the complete reference DSM, then frozen before export. The evaluation therefore targets scene-specific reconstruction rather than unseen-site prediction.

We report Raw MAE, Bias MAE, RMSE, coverage, DSM PSNR, and DSM SSIM on the official 0.5 m EPSG:32617 grid. The experiments examine protocol-matched reconstruction, fixed-field export rules, and held-out RGB rendering. Appendices A–E provide implementation details and quantitative and qualitative diagnostics.

Table 2: Scene-wise performance of the final 12-member ensemble. JAX_260 achieves the lowest reconstruction error, whereas JAX_214 shows the largest residual error and lowest coverage.

| Scene | Raw MAE↓ | Bias MAE↓ | RMSE↓ | PSNR↑ | SSIM↑ | Cov.↑ |
|---|---|---|---|---|---|---|
| JAX_004 | 0.772 | 0.764 | 1.711 | 22.596 | 0.924 | 95.60 |
| JAX_068 | 0.678 | 0.739 | 2.032 | 28.646 | 0.984 | 97.85 |
| JAX_214 | 1.184 | 1.300 | 4.274 | 27.422 | 0.956 | 91.81 |
| JAX_260 | 0.514 | 0.555 | 1.359 | 29.648 | 0.977 | 98.36 |
| Mean | **0.787** | **0.839** | **2.344** | **27.078** | **0.960** | **95.91** |

Table 3: Diagnostic analyses. **(a)** JAX_004 fixed-field export-rule comparison: all rows use the same learned Gaussian field and differ only in DSM readout. **(b)** Ensemble-size ablation using nested subsets of the same saved Gaussian exports with an identical map-space median fusion procedure.

**(a)**

| Export rule | MAE↓ | RMSE↓ | Cov. (%)↑ |
|---|---|---|---|
| Center projection | 0.953 | 1.645 | 20.39 |
| Footprint/opacity | 1.546 | 2.294 | 100.00 |
| Renderer-visible altitude | 1.167 | 1.936 | 94.04 |

**(b)**

| Members | Raw MAE↓ | Bias MAE↓ | RMSE↓ | Cov.↑ | MAE@95↓ |
|---|---|---|---|---|---|
| 1 | 0.907 | 0.904 | 2.585 | 95.30 | 4.009 |
| 3 | 0.847 | 0.846 | 2.486 | 95.46 | 3.759 |
| 6 | 0.819 | **0.817** | 2.384 | 95.84 | 3.621 |
| 12 | **0.787** | 0.839 | **2.344** | **95.91** | **3.475** |

### 4.2 Main DSM Reconstruction

Table 1 separates published cross-scene results from matched scene fitting. GaussLift remains competitive with direct raster predictors while preserving a renderable scene. The single-field and ensemble results distinguish representation quality from post-export aggregation gains. Figure 3 traces renderer outputs to the DSM; Figure 4 shows Gaussian centers as latent scene parameters rather than direct height samples. In Table 2, JAX_260 has the lowest Raw MAE and RMSE, whereas JAX_214 has the largest RMSE and lowest coverage, highlighting scene-dependent difficulty.

### 4.3 Fixed-Field Export Operator

Table 3(a) changes only the exporter of a fixed JAX_004 Gaussian field. Center projection has the lowest MAE and RMSE but covers only 20.39% of the grid. Footprint/opacity achieves full coverage at higher error; renderer-visible altitude covers 94.04% with lower error than footprint/opacity. Because each rule covers a different cell subset, this establishes operator dependence rather than an error ranking on identical support. Appendix D.2 provides a separate common-mask diagnostic on the final v52 fields.

### 4.4 Post-Export Ensemble Aggregation

The final result combines independently optimized cross-resolution exports by cell-wise median fusion. Table 3(b) varies only the number of saved member outputs. Raw MAE, RMSE, coverage, and MAE@95 improve with ensemble size, whereas Bias MAE is lowest with six members. The gain is therefore attributable to post-export aggregation, but not every metric improves monotonically. Appendix D.3 reports the same member-scaling results and further analyzes the gains from map-space fusion.

### 4.5 Training and Ensemble Configuration

Table 4 compares training and ensemble configurations under the same four-scene protocol. Cross-resolution robust aggregation performs best among these configurations, motivating scene-wise optimization followed by robust map-space fusion.

Table 4: Training and ensemble ablation under the same four-scene protocol and DSM evaluator. Variants change initialization, supervision, member composition, or resolution. Bold marks the best result per metric.

| Configuration | Raw MAE↓ | Bias MAE↓ | RMSE↓ | PSNR↑ | SSIM↑ | Cov.↑ |
|---|---|---|---|---|---|---|
| Base ensemble | 0.943 | 1.007 | 2.545 | 26.406 | 0.955 | 95.75 |
| From-scratch training | 0.847 | 0.911 | 2.479 | 26.581 | 0.956 | 95.46 |
| CNN pretraining | 0.881 | 0.950 | 2.495 | 26.566 | 0.955 | 95.62 |
| Edge-aware training | 0.883 | 0.946 | 2.562 | 26.405 | 0.954 | 95.62 |
| Heterogeneous ensemble | 0.808 | 0.866 | 2.383 | 26.961 | 0.959 | 95.77 |
| High-resolution configuration | 0.876 | 0.926 | 2.493 | 26.335 | 0.953 | 95.54 |
| Cross-resolution robust | **0.787** | **0.839** | **2.344** | **27.078** | **0.960** | **95.91** |
| Mid-resolution balanced | 0.834 | 0.888 | 2.400 | 26.677 | 0.959 | 95.78 |

### 4.6 Retained Appearance Capability

The optimized Gaussian fields remain renderable from held-out viewpoints without target-view fitting or test-time refinement (Appendix D.4). This capability is important because the reconstructed scene is not reduced to a height raster during optimization: it retains an explicit appearance-bearing 3D representation whose geometry participates in the same visibility-aware rendering process used for DSM extraction. In particular, the rendered RGB image and the renderer-conditioned altitude share the front-to-back compositing weights defined in Section 3, so the appearance and geometric outputs are supported by the same set of visible Gaussian contributions rather than by independent prediction heads.

We emphasize that rendering and final DSM reconstruction are evaluated at different levels of the pipeline. Held-out rendering is a capability of an individual optimized Gaussian field, whereas the reported final DSM is obtained after map-space aggregation of multiple independently optimized exports. The ensemble therefore improves the rasterized surface without altering the rendering mechanism of any constituent field. Figure 5 further illustrates that the final DSM is produced from renderer-supported altitude samples after calibrated world localization and deterministic grid realization, rather than by directly projecting Gaussian centers. Together, these results show that GaussLift preserves the dual role of the Gaussian scene as both a renderable representation and a source of georeferenced surface measurements.

### 4.7 Qualitative Error Analysis

Figures 6 and 7 provide a closer view of the residual geometry on JAX_260. Residual errors are concentrated around sharp height discontinuities, thin structures, vegetation, occlusion boundaries, and cells with weak renderer support. These regions are precisely where multiple Gaussian layers may contribute to the same image-space footprint or where the visible support changes abruptly across neighboring rays. As a result, the normalized compositing-weighted altitude moment can smooth competing height hypotheses and place the recovered surface between partially visible layers.

The scanline profile in Figure 6 shows that the reconstructed DSM generally follows the large-scale height structure of the reference surface, while the largest deviations occur near abrupt transitions. Figure 7 complements this spatial view with signed and absolute residual statistics, separating systematic vertical bias from the magnitude and tail of the reconstruction error. These observations are consistent with the design of the proposed readout: shared visibility attribution suppresses geometry that is not supported by the renderer, but the first-moment altitude observable does not by itself resolve multimodal ray-wise height distributions. Consequently, the residual error pattern is dominated by locally ambiguous surface attribution near complex or weakly observed structures rather than being distributed uniformly over the scene. This analysis motivates alternative robust ray-space statistics, such as weighted-median or terminal-opacity readouts, as potential extensions when sharper layer selection is required.

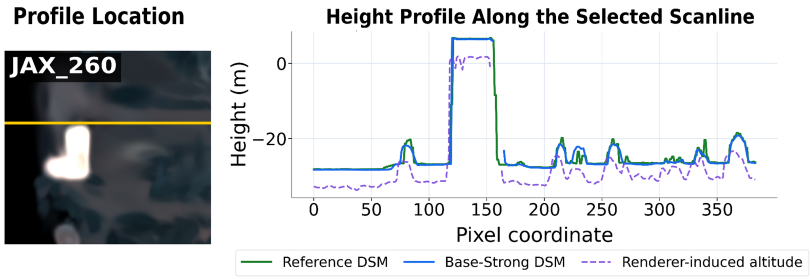


Figure 6: Height profile on JAX_260 along the yellow scanline. Curves compare the reference DSM, the GaussLift reconstruction, and renderer-induced altitude, with residuals concentrated near sharp transitions.

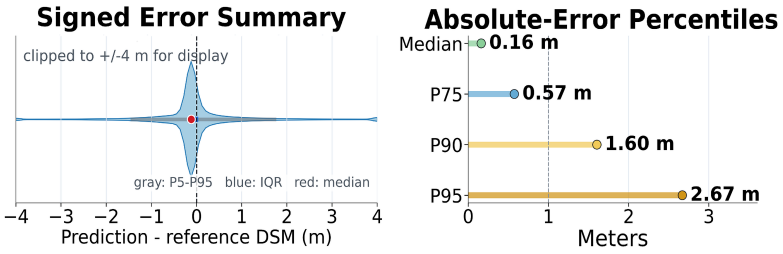


Figure 7: Error statistics on JAX_260. Signed errors characterize vertical bias; absolute-error percentiles summarize the magnitude and tail of the residuals.

## 5 Conclusion

We presented GaussLift, a renderer-consistent framework for converting satellite Gaussian fields into georeferenced digital surface models. Our central observation is that a learned Gaussian field does not uniquely determine a DSM: even with the representation fixed, primitive centers, projected footprints, and renderer-visible support can induce substantially different elevation surfaces. GaussLift makes this representation–measurement interface explicit by coupling altitude estimation to the same front-to-back compositing weights used for image formation, followed by calibrated world localization and deterministic map-space realization. Fixed-field experiments show that changing only the export operator can markedly affect both reconstruction error and coverage, while the full system preserves a renderable 3D representation and produces a metrically evaluated geospatial raster under the stated scene-specific, full-DSM-supervised protocol. These results indicate that geometric accuracy in satellite 3DGS depends not only on the learned representation, but also on how that representation is converted into a surface, and therefore motivate treating surface readout as an explicit component of the reconstruction pipeline rather than as a neutral post-processing step.

## AI Use Statement

ChatGPT was used only for literature and reference lookup during the preparation of this submission. It was not used to generate experimental results or perform data analysis. All retrieved information was independently verified by the authors.

## Reproducibility Statement

The main paper and appendix provide the methodological and experimental details necessary to understand and evaluate our approach. We also provide anonymized supplementary code covering the core components of the proposed method for technical inspection. The release focuses on method verification and does not include the full end-to-end training and experiment orchestration pipeline.

## REFERENCES


Luca Savant Aira, Gabriele Facciolo, and Thibaud Ehret. Gaussian splatting for efficient satellite image photogrammetry. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5959–5969, 2025.

Nan Bai, Anran Yang, Hao Chen, and Chun Du. SatGS: Remote sensing novel view synthesis using Multi-temporal satellite images with appearance-adaptive 3DGS. *Remote Sensing*, 17(9):1609, 2025.

Camille Billouard, Dawa Derksen, Emmanuelle Sarrazin, and Bruno Vallet. Sat-NGP: Unleashing neural graphics primitives for fast relightable transient-free 3D reconstruction from satellite imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 8749–8753. IEEE, 2024.

Pierrick Bournez, Luca Savant Aira, Thibaud Ehret, and Gabriele Facciolo. EOGS++: Earth Observation Gaussian Splatting with Internal Camera Refinement and Direct Panchromatic Rendering. *arXiv preprint arXiv:2511.16542*, 2025.

Sining Chen and Xiao Xiang Zhu. TSE-Net: Semi-supervised Monocular Height Estimation from Single Remote Sensing Images. *arXiv preprint arXiv:2511.13552*, 2025.

Sining Chen, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Adaptive bins for monocular height estimation from single remote sensing images. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7015–7018. IEEE, 2023a.

Sining Chen, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. HTC-DC Net: Monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023b.

Zhan Chen, Yidan Zhang, Xiyu Qi, Yongqiang Mao, Xin Zhou, Lei Wang, and Yunping Ge. HeightFormer: A multilevel interaction and image-adaptive classification–regression network for monocular height estimation with aerial images. *Remote Sensing*, 16(2):295, 2024.

Shouhang Du, Jianghe Xing, Shihong Du, Ximin Cui, Xiongwu Xiao, Wei Li, and Shaoyu Wang. IMG2HEIGHT: height estimation from single remote sensing image using a deep convolutional encoder-decoder network. *International Journal of Remote Sensing*, 44(18):5686–5712, 2023.

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*, pp. 5354–5363, 2024.

Furkan Gültekin, Alper Koz, Reza Bahmanyar, Seyed M Azimi, and Mehmet Lütfi Süzen. Fusing Convolution and Vision Transformer Encoders for Object Height Estimation from Monocular Satellite and Aerial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3709–3718, 2025.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Xuejun Huang, Xinyi Liu, Yi Wan, Zhi Zheng, Bin Zhang, Mingtao Xiong, Yingying Pei, and Yongjun Zhang. SkySplat: Generalizable 3D Gaussian splatting from multi-temporal sparse satellite images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5158–5166, 2026.

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.

Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown. 2019 IEEE GRSS Data Fusion Contest: Large-Scale Semantic 3D Reconstruction. *IEEE Geoscience and Remote Sensing Magazine*, 7(4):33–36, 2019.

Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3D gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21719–21728, 2024.

Chao-Jung Liu, Vladimir A Krylov, Paul Kane, Geraldine Kavanagh, and Rozenn Dahyot. IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sensing*, 12(17):2719, 2020.

Tianle Liu, Shuangming Zhao, Wanshou Jiang, and Bingxuan Guo. Sat-DN: Implicit Surface Reconstruction from Multi-View Satellite Images with Depth and Normal Supervision. *arXiv preprint arXiv:2502.08352*, 2025.

Feng Luo, Hongbo Pan, Xiang Yang, Baoyu Jiang, Fengqing Liu, and Tao Huang. ShadowGS: Shadow-Aware 3D Gaussian Splatting for Satellite Imagery. *arXiv preprint arXiv:2601.00939*, 2026.

Chao Ma, Yueting Zhang, Jiayi Guo, Guangyao Zhou, and Xiurui Geng. FusionHeightNet: A multi-level cross-fusion method from multi-source remote sensing images for urban building height estimation. *Remote Sensing*, 16(6):958, 2024.

Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Multi-date earth observation NeRF: The detail is in the shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2035–2045, 2023.

Babak Memar, Luigi Russo, Silvia Liberata Ullo, and Paolo Gamba. An Object-Based Deep Learning Approach for Building Height Estimation from Single SAR Images. *Remote Sensing*, 17(17): 2922, 2025.

Lichao Mou and Xiao Xiang Zhu. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv preprint arXiv:1802.10249*, 2018.

Emmanouil Panagiotou, Georgios Chochlakis, Lazaros Grammatikopoulos, and Eleni Charou. Generating elevation surface from a single RGB remotely sensed image using deep learning. *Remote Sensing*, 12(12):2002, 2020.

Xuanang Peng, Shixin Wang, Futao Wang, Jinfeng Zhu, Suju Li, Longfei Liu, and Zhenqing Wang. Synergistic semantic segmentation and height estimation for monocular remote sensing images via cross-task interaction. *Remote Sensing*, 17(9):1637, 2025.

Yingjie Qu and Fei Deng. Sat-Mesh: Learning neural implicit surfaces for multi-view satellite reconstruction. *Remote Sensing*, 15(17):4297, 2023.

Valentin Wagner, Sebastian Bullinger, Christoph Bodensteiner, and Michael Arens. RPC-GS: Gaussian Splatting with Native RPC Rendering for Satellite Imagery. *arXiv preprint arXiv:2606.06690*, 2026.

Songlin Xie, Lei Zhang, Gwanggil Jeon, and Xiaomin Yang. Remote sensing neural radiance fields for multi-view satellite photogrammetry. *Remote Sensing*, 15(15):3808, 2023.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-Splatting: Alias-free 3D gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19447–19456, 2024.

Lulin Zhang and Ewelina Rupnik. Sparsesat-NeRF: Dense depth supervised neural radiance fields for sparse satellite images. *arXiv preprint arXiv:2309.00277*, 2023.

Lulin Zhang, Ewelina Rupnik, Tri Dung Nguyen, Stéphane Jacquemoud, and Yann Klinger. BRDF-NeRF: Neural radiance fields with optical satellite images and BRDF modelling. *International Journal of Applied Earth Observation and Geoinformation*, 143:104747, 2025.

Xin Zhou, Yang Wang, Daoyu Lin, Zehao Cao, Biqing Li, and Junyi Liu. SatelliteRF: Accelerating 3D reconstruction in multi-view satellite images with efficient neural radiance fields. *Applied Sciences*, 14(7):2729, 2024.

# APPENDIX

**Appendix overview.** This appendix provides implementation details and additional diagnostic analyses supporting the main paper. Its organization follows the representation-to-measurement pipeline of GaussLift. We first describe the camera and scene-coordinate transformations used to localize renderer-conditioned altitude in geospatial coordinates. We then detail the Gaussian-field parameterization and the training-only surface-biased query allocation strategy, followed by the product-aligned supervision used to constrain the same altitude observable consumed during DSM export. The remaining sections provide complementary analyses of protocol-matched baselines, fixed-field export operators, ensemble scaling, held-out rendering, localization sensitivity, spatial error patterns, cross-scene transfer, and qualitative reconstructions. Together, these materials clarify the implementation, evaluation protocol, and empirical behavior of the proposed Gaussian-to-DSM reconstruction framework.

## A CAMERA AND GEOREFERENCING DETAILS

### A.1 RPC-DERIVED LOCAL AFFINE MODEL

Let $X_k^{\mathrm{aff}} = (x_k, y_k, h_k)^\top \in \mathbb{R}^3$ be the $k$-th control sample in the normalized scene frame, and let $p_k^{\mathrm{rpc}} \in \mathbb{R}^2$ be its image coordinate under the native rational polynomial camera (RPC). GaussLift uses a local affine approximation only as a coordinate interface between the renderer and the georeferenced DSM:

$$\Pi_{\mathrm{aff}}(X_k^{\mathrm{aff}}) = B \begin{bmatrix} x_k \\ y_k \end{bmatrix} + a_h h_k + b, \tag{S1}$$

where $B \in \mathbb{R}^{2\times 2}$ maps horizontal scene coordinates to the image, $a_h \in \mathbb{R}^2$ models altitude-induced displacement, and $b \in \mathbb{R}^2$ is the image offset. The parameters are estimated offline by least squares:

$$(B, a_h, b) = \underset{B,a_h,b}{\arg\min} \sum_{k=1}^{N_c} \left\| \Pi_{\mathrm{aff}}(X_k^{\mathrm{aff}}) - p_k^{\mathrm{rpc}} \right\|_2^2. \tag{S2}$$

The control footprint, altitude interval, and fitting density are fixed during preprocessing and reused unchanged during optimization and export. The affine approximation therefore does not introduce a learned or test-time-adjusted camera.

### A.2 DETERMINISTIC VIEW ORDERING AND BACK-PROJECTION

The affine projection matrix is

$$P_{\mathrm{aff}} = [B\ a_h] \in \mathbb{R}^{2\times 3}. \tag{S3}$$

Let $v$ be a unit vector in the null space of $P_{\mathrm{aff}}$. Its sign is selected once so that $v^\top r_{\mathrm{rpc}} > 0$, where $r_{\mathrm{rpc}}$ is the local RPC viewing direction at the scene center. The scalar

$$\kappa_{\Pi_{\mathrm{aff}}}(X) = v^\top X \tag{S4}$$

then provides a deterministic front-to-back ordering coordinate shared by Gaussian compositing and reference-surface projection.

Given image coordinate $p \in \mathbb{R}^2$ and normalized altitude $h$, altitude-conditioned back-projection is

$$\mathcal{B}_{\mathrm{aff}}(p, h) = \begin{bmatrix} B^\dagger(p - a_h h - b) \\ h \end{bmatrix}, \tag{S5}$$

where $B^\dagger$ is the inverse of $B$ when it is full rank and otherwise its Moore–Penrose pseudoinverse. A normalized point is converted to world coordinates by

$$\mathcal{W}_{\mathrm{scene}}(X^{\mathrm{aff}}) = S_{\mathrm{scene}} X^{\mathrm{aff}} + c_{\mathrm{scene}}, \tag{S6}$$

where $S_{\mathrm{scene}} = \mathrm{diag}(s_x, s_y, s_h)$ contains the horizontal and vertical denormalization scales and $c_{\mathrm{scene}} \in \mathbb{R}^3$ is the scene center. The inverse mapping is

$$(\mathcal{W}_{\mathrm{scene}})^{-1}(X^w) = S_{\mathrm{scene}}^{-1}(X^w - c_{\mathrm{scene}}). \tag{S7}$$

These equations separate two roles that are easy to conflate. The renderer determines which altitude is visible along a ray; the camera interface only localizes that already selected altitude in world coordinates. This distinction is central to the operator analysis in the main paper.

# B GAUSSIAN-FIELD INSTANTIATION

## B.1 DECODER PARAMETERIZATION

GaussLift is defined at the export-operator level and is not tied to a specific Gaussian generator. In our implementation, a scene-specific tri-plane model produces a query descriptor $\phi_i$, after which separate geometry and appearance–opacity branches predict

$$\begin{aligned}(\Delta\mu_i, s_i, r_i) &= D_{\mathrm{geo}}(\phi_i),\\ (\widetilde{c}_i, \widetilde{\alpha}_i) &= D_{\mathrm{app}}(\phi_i).\end{aligned} \tag{S8}$$

Here, $\Delta\mu_i \in \mathbb{R}^3$ is the center displacement, $s_i \in \mathbb{R}^3$ contains log-scales, $r_i$ is mapped to a rotation $R(r_i) \in \mathrm{SO}(3)$, $\widetilde{c}_i \in \mathbb{R}^3$ is the RGB logit, and $\widetilde{\alpha}_i \in \mathbb{R}$ is the opacity logit. The final Gaussian parameters are

$$\begin{aligned}\mu_i &= q_i + \delta_\mu \tanh(\Delta\mu_i),\\ \Sigma_i &= R(r_i)\,\mathrm{diag}\big(e^{2s_i}\big)\,R(r_i)^\top,\\ c_i &= \mathrm{sigmoid}(\widetilde{c}_i),\\ \alpha_i &= \mathrm{sigmoid}(\widetilde{\alpha}_i),\end{aligned} \tag{S9}$$

where $q_i \in \mathbb{R}^3$ is the lattice anchor and $\delta_\mu > 0$ bounds the center displacement. The covariance construction guarantees $\Sigma_i \succeq 0$, while color and opacity remain in $[0, 1]$.

The parameterization supplies an explicit, renderable Gaussian field; it does not define the DSM surface by itself. Surface semantics are introduced only by the renderer-conditioned export operator analyzed in the main paper.

## B.2 SURFACE-BIASED QUERY ALLOCATION

Let $\mathcal{Q} = \{q_i\}_{i=1}^{N_g}$ be the deterministic inference lattice. During training, a fraction $r_{\mathrm{surf}} \in [0, 1]$ of its anchors is replaced:

$$\begin{aligned}\mathcal{Q}_{\mathrm{train}} &= \mathcal{Q}_{\mathrm{keep}} \cup \mathcal{Q}_{\mathrm{surf}},\\ |\mathcal{Q}_{\mathrm{surf}}| &= r_{\mathrm{surf}} N_g,\\ |\mathcal{Q}_{\mathrm{keep}}| &= (1 - r_{\mathrm{surf}}) N_g.\end{aligned} \tag{S10}$$

Here, $\mathcal{Q}_{\mathrm{keep}} \subseteq \mathcal{Q}$ contains retained lattice anchors and $\mathcal{Q}_{\mathrm{surf}}$ contains points sampled near valid reference-DSM surfaces. A surface query is represented as

$$q^{\mathrm{surf}} = (\mathcal{W}_{\mathrm{scene}})^{-1}(X_u^w + \xi), \qquad u \sim P_{\mathrm{valid}}, \quad \xi \sim \mathcal{D}_{\mathrm{surf}}, \tag{S11}$$

where $P_{\mathrm{valid}}$ is the sampling distribution over valid DSM cells and $\mathcal{D}_{\mathrm{surf}}$ is the spatial perturbation distribution. The query count remains $N_g$, and inference always returns to the deterministic lattice $\mathcal{Q}_{\mathrm{test}} = \mathcal{Q}$. Thus, reference-derived queries affect optimization capacity but are not consumed by the final DSM exporter.

# C PRODUCT-ALIGNED SURFACE SUPERVISION

## C.1 REFERENCE-ALTITUDE PROJECTION

Let $H^*(u)$ be the reference height at valid target-grid cell $u \in \Omega_H$, with horizontal world coordinate $(x_u^w, y_u^w)$. The corresponding world point is

$$X_u^w = \begin{bmatrix} x_u^w \\ y_u^w \\ H^*(u) \end{bmatrix}. \tag{S12}$$

It is converted to the normalized scene frame using Equation (S7) and projected into the conditioning image with $\Pi_{\mathrm{aff}}$.

Let $\mathcal{U}(p)$ denote the valid DSM cells whose projected locations fall into pixel $p$. The retained frontmost cell is

$$u^*(p) = \operatorname*{arg\,min}_{u \in \mathcal{U}(p)} \kappa_{\Pi_{\mathrm{aff}}}(X_u^{\mathrm{aff}}), \tag{S13}$$

using the same ordering convention as Gaussian compositing. The reference altitude and validity mask are

$$A^*_{\text{aff}}(p) = h^{\text{aff}}_{u^*(p)}, \qquad M^*_A(p) = \mathbf{1}[\mathcal{U}(p) \neq \emptyset]. \tag{S14}$$

Unsupported pixels receive $M^*_A(p) = 0$ and do not contribute to surface supervision.

This construction is product-aligned: the hard reference surface and the soft renderer-conditioned altitude are compared in the same camera-aligned coordinate system and under the same front-to-back ordering convention. It avoids supervising one geometric quantity while exporting another.

### C.2 Training Losses

Let $I(p)$ and $\hat{I}(p)$ be the observed and rendered RGB values. The Charbonnier reconstruction term is

$$\mathcal{L}_{\text{charb}} = \frac{1}{|\Omega_I|} \sum_{p \in \Omega_I} \sqrt{\left\| \hat{I}(p) - I(p) \right\|_2^2 + \epsilon^2_{\text{charb}}}, \tag{S15}$$

and the RGB objective is

$$\mathcal{L}_{\text{rgb}} = (1 - \lambda_{\text{dssim}})\mathcal{L}_{\text{charb}} + \lambda_{\text{dssim}}\mathcal{L}_{\text{DSSIM}}. \tag{S16}$$

Let $A(p) = \hat{A}_{\text{aff}}(p)$, $A^*(p) = A^*_{\text{aff}}(p)$, and $\Omega_A = \{p \mid M^*_A(p) = 1\}$. The primary altitude objective is

$$\mathcal{L}_{\text{alt}} = \frac{\sum_{p \in \Omega_A} \rho_{\text{SL1}}\big(A(p) - A^*(p)\big)}{|\Omega_A| + \epsilon}. \tag{S17}$$

The gradient term preserves local height transitions:

$$\mathcal{L}_{\text{grad}} = \frac{1}{|\Omega_A|} \sum_{p \in \Omega_A} \Big( |\nabla_x A(p) - \nabla_x A^*(p)| + |\nabla_y A(p) - \nabla_y A^*(p)| \Big). \tag{S18}$$

For scale-and-shift-invariant comparison, define

$$(s^*, t^*) = \arg\min_{s,t} \sum_{p \in \Omega_A} (sA(p) + t - A^*(p))^2, \tag{S19}$$

with

$$\mathcal{L}_{\text{ssi}} = \frac{1}{|\Omega_A|} \sum_{p \in \Omega_A} |s^* A(p) + t^* - A^*(p)|. \tag{S20}$$

The complete objective is

$$\mathcal{L} = \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{alt}}\mathcal{L}_{\text{alt}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{ssi}}\mathcal{L}_{\text{ssi}}, \tag{S21}$$

where all $\lambda_\bullet \geq 0$. The loss terms constrain appearance, absolute altitude, local transitions, global structural correlation, and relative shape, but only $\mathcal{L}_{\text{alt}}$ directly supervises the altitude statistic consumed by the export operator.

## D Additional Quantitative Diagnostics

The supplementary experiments separate three factors that are coupled in the final system: the representation used to fit each scene, the operator used to convert a fixed Gaussian field into a DSM, and the post-export aggregation applied across independently optimized maps. Unless stated otherwise, all results use the same four Jacksonville scenes and are averaged uniformly across scenes. Raw MAE and RMSE are treated as the primary geometric criteria; bias-aligned error, coverage, raster similarity, and held-out rendering provide complementary diagnostics.

Table S1: Protocol-matched direct DSM baselines. All methods use the same four target scenes and full DSM supervision; only GaussLift retains an explicit renderable Gaussian scene.

| Method | 3D | Raw↓ | Bias↓ | RMSE↓ | Cov.↑ |
|---|---|---|---|---|---|
| CMT reimpl. | no | 1.007 | 1.073 | 2.672 | 95.56 |
| HeightFormer reimpl. | no | 1.004 | 1.025 | 2.545 | 96.12 |
| LUMNet reimpl. | no | 0.955 | 1.008 | 2.545 | 95.90 |
| GaussLift v52 | yes | 0.787 | 0.839 | 2.344 | 95.91 |

### D.1 Does the Explicit Gaussian Representation Help?

Table S1 compares GaussLift with direct DSM decoders under the same target-scene fitting protocol. GaussLift reduces Raw MAE from 0.955 m for the strongest matched raster baseline to 0.787 m and reduces RMSE from 2.545 m to 2.344 m, while preserving similar coverage. The comparison isolates the choice of intermediate representation: direct baselines predict only a height raster, whereas GaussLift must encode geometry in an explicit Gaussian field that also supports image rendering. These results therefore show that retaining a renderable representation does not require sacrificing metric DSM accuracy in the evaluated scene-specific setting. They do not imply superiority under unseen-scene prediction.

### D.2 How Much Does the Export Operator Matter?

Table S2: Fixed-field Gaussian-to-DSM export comparison for the final v52 fields. The Gaussian fields, camera, grid, validity definition, and evaluator are fixed; only the ray-space surface functional or map realization changes.

| Export rule | CM-MAE↓ | RMSE↓ | Cov.↑ | MAE@95↓ |
|---|---|---|---|---|
| Center projection | 0.794 | 2.403 | 95.91 | 3.480 |
| Footprint / opacity | 0.794 | 2.403 | 98.78 | 4.402 |
| Renderer moment | 0.785 | 2.349 | 95.91 | 3.475 |
| Renderer + robust grid | 0.785 | 2.349 | 100.00 | 4.867 |

Table S2 holds the optimized Gaussian fields, camera, target grid, validity definition, and evaluator fixed. The remaining differences are therefore attributable to the Gaussian-to-DSM operator rather than to representation learning. Renderer-conditioned altitude gives the lowest common-mask MAE and RMSE, although the margin over center projection is modest. Footprint aggregation expands coverage but performs worse at matched support, showing that denser output alone does not imply a more accurate surface. The robust-grid variant reaches complete coverage yet further degrades MAE@95. Together, these results support the central claim that surface attribution and map realization are measurable design choices, while also showing that their benefits must be evaluated jointly with coverage.

### D.3 What Produces the Final System Gain?

Table S3: Member-scaling diagnostic for v52 using the same results as Table 3. Cell-wise median fusion combines independently optimized Gaussian exports. Bold marks the best result per metric.

| Members | Raw↓ | Bias↓ | RMSE↓ | Cov.↑ | MAE@95↓ |
|---|---|---|---|---|---|
| 1 | 0.907 | 0.904 | 2.585 | 95.30 | 4.009 |
| 3 | 0.847 | 0.846 | 2.486 | 95.46 | 3.759 |
| 6 | 0.819 | **0.817** | 2.384 | 95.84 | 3.621 |
| 12 | **0.787** | 0.839 | **2.344** | **95.91** | **3.475** |

Table S3 isolates the effect of post-export ensemble size. Increasing the number of independently optimized members from one to twelve lowers Raw MAE from 0.907 m to 0.787 m and RMSE from 2.585 m to 2.344 m, with a small coverage gain. Raw MAE, RMSE, coverage, and MAE@95 improve monotonically, whereas Bias MAE reaches its minimum of 0.817 m at six members and rises to 0.839 m at twelve. Because fusion is performed after each member has already produced a DSM, the gains reflect map-space robustness rather than a stronger individual Gaussian field.

Table S4: Additional surface-oriented training strategies under the same four-scene protocol. Individual variants improve selected aspects but do not explain the full gain of the final cross-resolution ensemble.

| Variant | Raw↓ | Bias↓ | RMSE↓ | PSNR↑ | SSIM↑ | Cov.↑ |
|---|---|---|---|---|---|---|
| Curriculum | 1.017 | 1.091 | 2.720 | 26.127 | 0.842 | 95.71 |
| Hard sampling | 0.989 | 1.063 | 2.611 | 26.085 | 0.839 | 95.94 |
| Multiscale altitude | 0.986 | 1.063 | 2.700 | 25.905 | 0.838 | 96.02 |
| Combined strategy | 0.995 | 1.069 | 2.647 | 26.131 | 0.840 | 95.84 |

Table S4 examines additional surface-oriented training strategies. Hard sampling gives the best RMSE, while multiscale altitude supervision gives the best Raw MAE and coverage among these variants. Their combination does not dominate the individual components, indicating that the mechanisms are not simply additive. These experiments support surface-aware optimization but do not explain the full gain of the final cross-resolution ensemble.

### D.4 Does the Field Remain Renderable, and How Sensitive Is Localization?

Table S5: Held-out RGB rendering by the optimized Gaussian scene models. The final 12-member DSM is a map-space product and is not itself rendered.

| Scene | Pairs | PSNR↑ | SSIM↑ | LPIPS↓ | MAE↓ |
|---|---|---|---|---|---|
| JAX_004 | 11 | $25.986 \pm 1.582$ | $0.722 \pm 0.050$ | $0.482 \pm 0.024$ | $0.034 \pm 0.008$ |
| JAX_068 | 18 | $20.332 \pm 1.457$ | $0.596 \pm 0.038$ | $0.544 \pm 0.021$ | $0.062 \pm 0.009$ |
| JAX_214 | 23 | $20.756 \pm 1.273$ | $0.611 \pm 0.038$ | $0.526 \pm 0.031$ | $0.057 \pm 0.008$ |
| JAX_260 | 16 | $23.173 \pm 1.710$ | $0.717 \pm 0.053$ | $0.446 \pm 0.043$ | $0.037 \pm 0.006$ |
| Mean / total | 68 | 22.562 | 0.662 | 0.500 | 0.047 |

The held-out results in Table S5 test whether the scene-specific Gaussian field retains appearance after the projection camera changes. Across 68 views, the models obtain 22.562 dB PSNR, 0.662 SSIM, and 0.500 LPIPS. JAX_004 and JAX_260 are rendered more reliably than JAX_068 and JAX_214, revealing substantial scene dependence. Since no target RGB content, exposure fitting, or test-time photometric optimization is used, the experiment verifies appearance persistence under reprojection. It does not measure cross-scene or illumination-invariant generalization.

Table S6: RPC/affine localization sensitivity. Horizontal perturbations increase RMSE and MAE@95 while leaving coverage nearly unchanged.

| Localization | Bias↓ | RMSE↓ | Cov.↑ | MAE@95↓ |
|---|---|---|---|---|
| Nominal | 0.839 | 2.344 | 95.91 | 3.475 |
| $x + 0.5\,\mathrm{m}$ | 0.803 | 2.402 | 95.82 | 3.554 |
| $y + 0.5\,\mathrm{m}$ | 0.861 | 2.514 | 95.94 | 3.869 |
| $xy + 0.5\,\mathrm{m}$ | 0.875 | 2.556 | 95.85 | 3.919 |
| $x + 1.0\,\mathrm{m}$ | 0.890 | 2.599 | 95.73 | 3.959 |
| $y + 1.0\,\mathrm{m}$ | 0.979 | 2.752 | 95.95 | 4.475 |

Table S6 perturbs the world-coordinate localization while leaving the rendered altitude unchanged. Every tested displacement increases RMSE and MAE@95. Bias MAE does not uniformly increase: the $x + 0.5$ m perturbation gives 0.803 m, compared with 0.839 m nominally. Equal perturbations along $y$ produce higher errors than those along $x$ for these scenes. Coverage remains nearly constant, indicating sensitivity to spatial misregistration rather than missing predictions. This diagnostic should not be interpreted as a direct native-RPC versus affine camera comparison.

### D.5 Where Does the Method Fail, and Does It Transfer Across Scenes?

Table S7: Error-by-region diagnostic for v52. Residual error is concentrated near high-gradient regions such as roof edges, vegetation boundaries, and sharp height discontinuities.

| Region | Bin | Ratio | Valid | Bias↓ | RMSE↓ | MAE@95↓ |
|---|---|---|---|---|---|---|
| Low grad. | 0–50% | 50.00 | 96.14 | 0.214 | 1.121 | 0.579 |
| Mid grad. | 50–90% | 40.00 | 96.82 | 0.850 | 1.894 | 2.888 |
| High grad. | 90–100% | 10.00 | 91.08 | 3.546 | 5.967 | 11.393 |

Table S7 stratifies the final errors by reference-surface gradient. Low-gradient regions constitute half of the evaluated cells and are reconstructed with 0.214 m Bias MAE. In contrast, the highest-gradient 10% reaches 3.546 m Bias MAE and 11.393 m MAE@95, with lower valid coverage. The remaining error is therefore concentrated around discontinuities rather than distributed uniformly across planar terrain. This pattern is consistent with partial visibility, broad Gaussian support, and map-cell collisions at roof edges, vegetation boundaries, and thin structures.

Table S8: Leave-one-scene-out calibration diagnostic. A global $H_0 \rightarrow$altitude linear calibration is fitted on three scenes and evaluated on the fourth; this is not the main scene-specific protocol.

| Held-out | Img. MAE | Img. RMSE | DSM MAE | DSM RMSE | Cov. | MAE@95 |
|---|---|---|---|---|---|---|
| JAX_004 | 1.886 | 3.198 | 2.050 | 3.450 | 97.26 | 8.428 |
| JAX_068 | 3.692 | 6.061 | 10.124 | 14.426 | 98.45 | 34.764 |
| JAX_214 | 5.133 | 10.558 | 12.739 | 19.523 | 95.11 | 50.470 |
| JAX_260 | 2.186 | 3.909 | 1.602 | 2.879 | 98.89 | 6.281 |
| Mean | 3.225 | 5.931 | 6.629 | 10.069 | 97.43 | 24.986 |

The leave-one-scene-out calibration results in Table S8 define the limit of the current evidence. Transfer remains relatively accurate for JAX_004 and JAX_260 but deteriorates sharply on JAX_068 and JAX_214, even though coverage stays high. The failure is therefore not explained by raster incompleteness alone; scene-dependent altitude calibration and representation fitting remain major obstacles. These results reinforce that the main experiments evaluate scene-specific reconstruction, not unseen-site prediction.

### D.6 Reference-Assisted Capacity Diagnostic

Table S9: Reference-assisted refinement. This variant accesses reference DSM information during refinement and is reported as a capacity diagnostic, not as part of fixed post-optimization export.

| Scene | Raw↓ | RMSE↓ | PSNR↑ | SSIM↑ | Cov.↑ |
|---|---|---|---|---|---|
| JAX_004 | 0.56 | 1.25 | 25.35 | 0.960 | 96.41 |
| JAX_068 | 0.94 | 2.37 | 27.31 | 0.977 | 98.79 |
| JAX_214 | 1.16 | 3.33 | 29.60 | 0.975 | 93.80 |
| JAX_260 | 0.65 | 1.46 | 29.03 | 0.972 | 98.65 |
| Mean | 0.83 | 2.10 | 27.82 | 0.971 | 96.91 |

The reference-assisted variant in Table S9 achieves a lower RMSE than the final ensemble but a slightly higher mean Raw MAE. Reference access changes the distribution of errors rather than uniformly improving every metric, so the result is not a strict upper bound. Because the reference DSM is consumed during refinement, this diagnostic is reported separately from the fixed post-optimization export protocol. Its role is to indicate remaining representation and refinement capacity, not deployable performance.

## E Additional Qualitative Analysis

The qualitative analysis examines four complementary aspects of the representation-to-product pathway: ambiguity among Gaussian-to-DSM export rules, the intermediate quantities produced by the renderer, the relation between renderer-visible surface evidence and the final DSM, and the explicit three-dimensional organization retained by the Gaussian field. Each visualization is therefore reported as an independent figure with its own caption, label, and interpretation.

Figure S1 directly visualizes the Gaussian-to-DSM ambiguity studied in the main paper. Center projection produces a sparse and fragmented product, whereas footprint-based and renderer-conditioned export generate denser but structurally non-equivalent surfaces. Their error maps show that changing only the export rule produces spatially systematic differences rather than random perturbations.

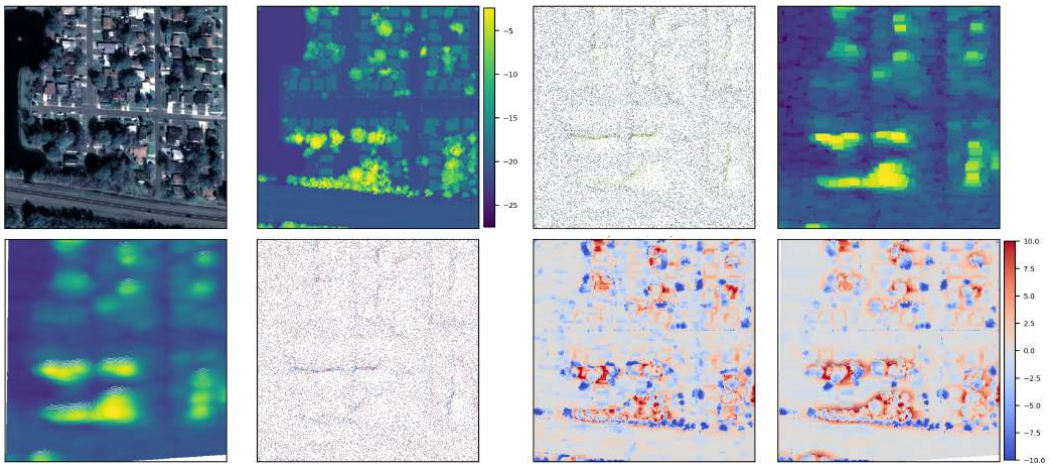


Figure S1: **Gaussian-to-DSM ambiguity under a fixed Gaussian scene.** Center projection, footprint/opacity export, and renderer-induced altitude produce different height fields from the same optimized Gaussian representation.

Figure S2 connects renderer evidence to the final geospatial product. Only sufficiently supported altitude samples are promoted to DSM realization. The dominant scene structures are recovered, whereas residual errors remain concentrated near boundaries, vegetation, and thin objects.

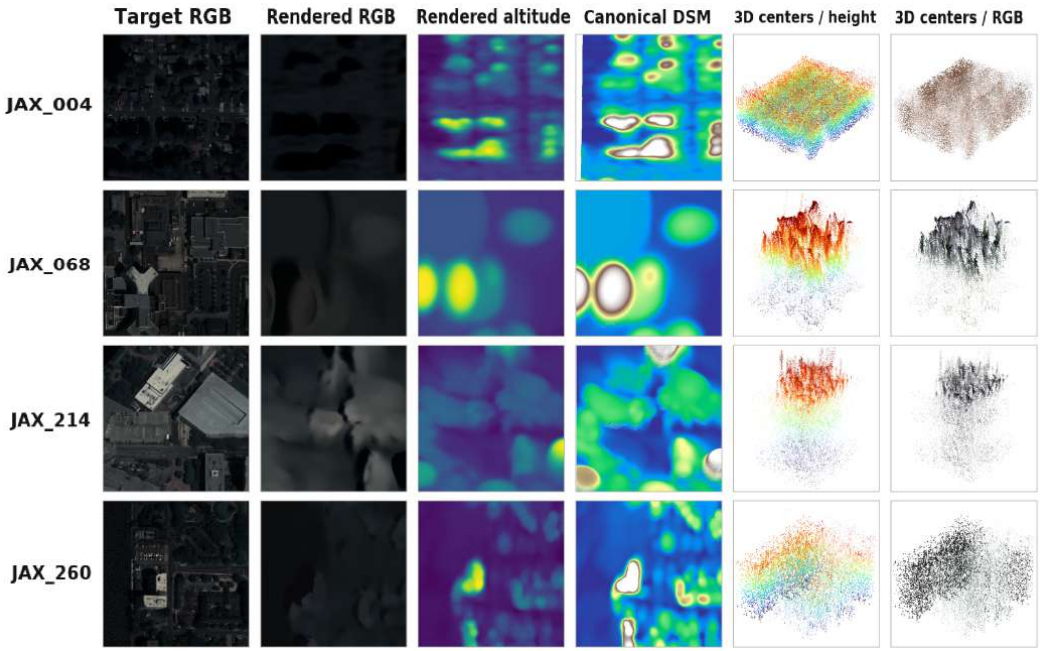


Figure S2: **Renderer-visible surface evidence across the four JAX scenes.** Rows are labeled JAX_004, JAX_068, JAX_214, and JAX_260. Columns show target RGB, rendered RGB, rendered altitude, canonical DSM, height-colored centers, and RGB-colored centers. The panel labels, including all four scene rows, are retained from the original visualization.

Figure S3 demonstrates that the method retains an explicit Gaussian scene instead of predicting only a height raster. The center cloud visualizes the learned three-dimensional organization, but the evaluated DSM is obtained only after renderer-conditioned surface attribution, calibrated world localization, and deterministic map-space realization.

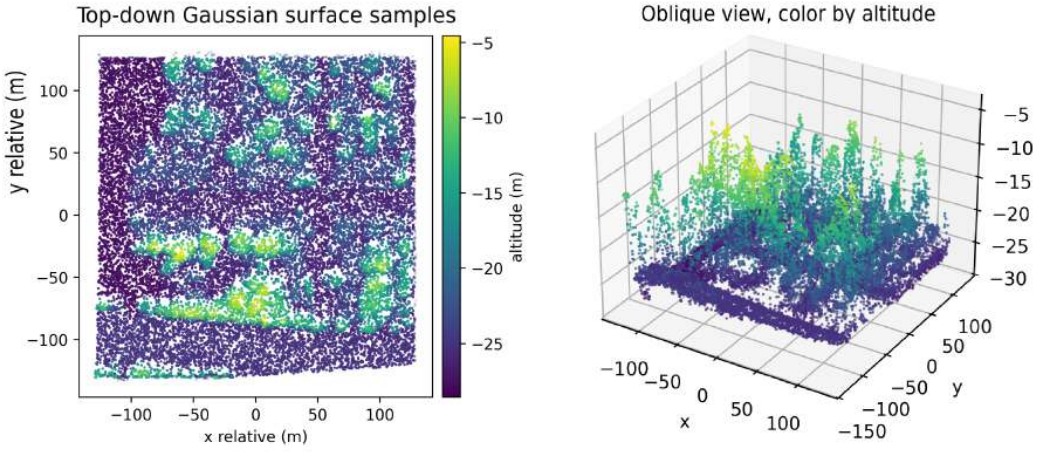


Figure S3: **Explicit three-dimensional Gaussian scene organization.** Top-down and oblique views show the scene-specific Gaussian field with samples colored by altitude.