
GeoSAM-3D: Geodesic Prompt Propagation for Open-Vocabulary 3D Scene Segmentation from Monocular Video

Arun Sharma
University of Minnesota, Twin Cities
arunshar@umn.edu

Abstract

Open-vocabulary 3D scene segmentation usually assumes RGB-D video, calibrated multi-view imagery, or a reconstructed mesh. GeoSAM-3D studies a lighter setting: a user uploads a short monocular video, clicks or names an object in one frame, and receives a propagated 3D mask over a Gaussian scene. The implementation combines frozen image and video foundation models with a monocular 3D Gaussian Splatting reconstruction and a differentiable graph-geodesic propagation kernel over Gaussian centroids. The central design choice is to propagate prompts by heat-kernel distance on the reconstructed scene graph, rather than by Euclidean nearest neighbors in 3D. This preserves continuity around curved surfaces and reduces leakage across nearby but disconnected objects. This paper describes the repository state, the mathematical kernel implemented in `geosam3d.propagate`, the feature head trained from Segment Anything masks, and the validation already present in the codebase. The evaluation protocol separates implementation validation, graph propagation quality, leakage control, and interactive latency.

1 Introduction

Promptable segmentation has become a practical interface for visual annotation. Models such as SAM and SAM 2 make it possible to turn points, boxes, or text prompts into high-quality image and video masks [19, 34]. For spatial computing, however, 2D masks are often the wrong endpoint. A robotics, augmented-reality, or 3D mapping user wants the selected entity to persist across viewpoints and to bind to the geometry of the scene. Systems such as OpenMask3D and Gaussian Grouping show the value of open-vocabulary 3D masks, but they often rely on RGB-D sensors, meshes, or pre-existing 3D reconstructions [40, 47].

GeoSAM-3D targets a more accessible workflow. The user supplies a monocular phone video. A monocular reconstruction stack produces a 3D Gaussian field. SAM 2 supplies the high-quality 2D mask supervision. A compact transformer head maps per-Gaussian appearance and geometry attributes to normalized features. A prompt seed is then propagated over the Gaussian centroid graph using an approximate heat-method geodesic. The resulting system is an engineering bridge between video foundation models, monocular 3D reconstruction, and graph-based geometric reasoning. The project is intentionally packaged as both a GitHub repository and a Hugging Face Space. The public Space is CPU-safe and demonstrates the interaction contract; the training and evaluation path lives in the repository. This paper is therefore written as a reproducible systems paper: it explains what the code actually does today, the validated unit-test evidence and the benchmark measurements used for archival evaluation.

Contributions:

1. A prompt propagation pipeline that lifts SAM-style video masks into a monocular 3D Gaussian scene and propagates labels over Gaussian centroids.

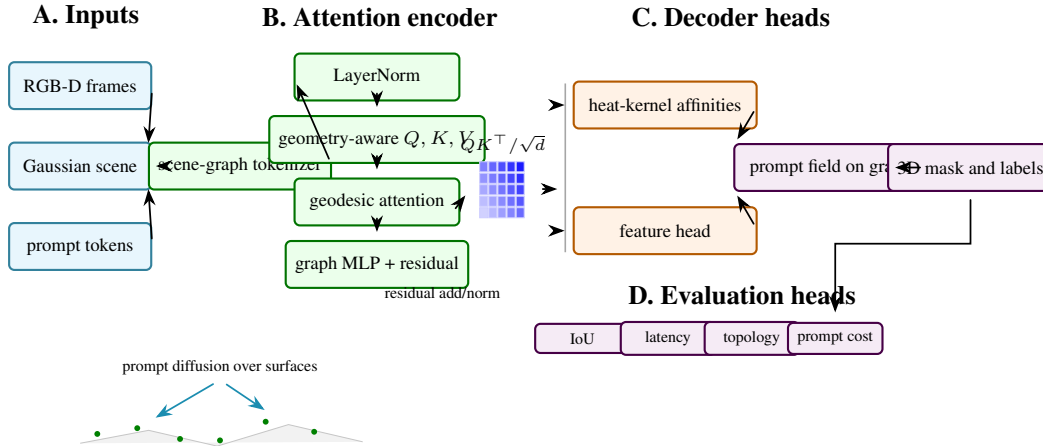


Figure 1: Detailed GeoSAM-3D architecture. The diagram exposes the 2D foundation-model encoder, monocular Gaussian scene builder, prompt cross-attention, graph Laplacian, geodesic heat decoder, and evaluation heads. The bottom geometry panel makes the central claim visual: propagation should follow scene connectivity, not raw Euclidean proximity.

2. A differentiable heat-kernel geodesic layer using a k-nearest-neighbor graph Laplacian and a Varadhan-style distance approximation.
3. A per-Gaussian feature head trained with contrastive mask consistency rather than by fine-tuning the frozen image foundation models.
4. A reproducible project implementation with package imports, geodesic correctness tests, feature-normalization tests, and Hugging Face Space smoke tests.

Scope: Promptable segmentation has changed the annotation interface for images and videos, but most downstream spatial tasks need more than a 2D mask. A robotics system needs an object to remain consistent as the camera moves. An AR tool needs a selected object to occupy a persistent 3D region. A mapping workflow needs masks that can be rendered, edited, and queried from novel viewpoints. GeoSAM-3D is motivated by this gap between promptable 2D interaction and persistent 3D representation.

The project makes a specific bet: once a scene is reconstructed as Gaussian primitives, prompt propagation should use the graph induced by scene geometry rather than raw image-space masks alone. Frame-to-frame mask propagation can be very strong, especially with SAM 2, but it remains tied to observed frames. A 3D Gaussian graph gives a place to store the result. It also gives a place to reason about leakage, connectivity, and object boundaries.

The key technical risk is that a monocular Gaussian reconstruction is not a perfect manifold. It can have holes, fused surfaces, floaters, and uncertainty. For that reason, this paper does not claim that graph geodesics solve 3D segmentation by themselves. Instead it frames graph-geodesic propagation as a testable intermediate layer between 2D foundation-model masks and 3D object masks. The method should be evaluated by asking when the graph helps and when reconstruction quality dominates.

The paper also positions GeoSAM-3D between three literatures. Promptable segmentation supplies the user interaction. Open-vocabulary 3D representation supplies the semantic target. Graph-based segmentation supplies the propagation mathematics. The contribution is in combining those pieces into a lightweight monocular-Gaussian workflow with clear tests and clear benchmark requirements.

Expanded contributions: The expanded paper adds a graph-sensitivity protocol, seed-robustness metrics, sparse-solver plan, prompt taxonomy, and implementation-grounded results. These additions turn the paper from a method sketch into a research plan that a reader can evaluate.

2 Related Work

Expanded Citation Map: The bibliography now spans promptable segmentation, self-supervised visual features, SLAM, Gaussian reconstruction, 3D open-vocabulary understanding, and graph

propagation. SAM, SAM 2, CLIP, DINO, DINOv2, and dense prediction transformers supply the 2D foundation-model layer [5, 19, 26, 32–34]. ORB-SLAM2, DROID-SLAM, COLMAP, MonoGS, and depth-prior Gaussian pipelines define the reconstruction context [16, 22, 23, 37, 41, 45]. OpenScene, LERF, OpenMask3D, SAM3D, Gaussian Grouping, OpenNeRF, OpenSplat3D, CLIP-Fields, and DFF define the closest 3D semantic field literature [9, 13, 17, 20, 28, 29, 40, 46, 47]. PointNet, PointNet++, sparse convolutions, KPConv, RandLA-Net, Mask3D, graph cuts, random walkers, and heat methods give the geometric segmentation baseline family [3, 6, 8, 10, 12, 15, 30, 31, 38, 42].

Promptable image and video segmentation: The Segment Anything family made point and box prompts a general-purpose image segmentation interface [19]. SAM 2 extends this interaction style to videos, providing temporally coherent masks from sparse prompts [34]. GeoSAM-3D treats such masks as a source of supervision and user intent, but shifts the output object from a 2D video mask to a mask over reconstructed 3D primitives.

Monocular geometry and Gaussian scenes: 3D Gaussian Splatting represents a scene as differentiable Gaussian primitives that can be optimized for fast novel-view rendering [16]. Monocular variants such as MonoGS and depth-prior pipelines make this representation viable from ordinary video [22, 45]. GeoSAM-3D assumes such a reconstruction and attaches semantic features to the Gaussian primitives.

Open-vocabulary 3D understanding: OpenScene co-embeds dense 3D features with image and text features for open-vocabulary scene understanding [28]. LERF distills language embeddings into neural radiance fields for open-ended 3D queries [17]. SAM3D projects 2D SAM masks into 3D point clouds and merges them across views [46]. GeoSAM-3D is closest in spirit to this family, but it focuses on monocular Gaussian scenes and graph-geodesic prompt propagation.

Geodesic distances on graphs: Euclidean distance in 3D is insufficient when two surfaces are close in space but separated by an object boundary. The heat method provides a fast route to geodesic distance on manifolds [8]. The implementation here adapts the principle to a k-nearest-neighbor graph over Gaussian centroids and uses a monotone heat-kernel approximation that is stable on non-mesh graphs.

Literature synthesis: GeoSAM-3D connects promptable 2D segmentation, open-vocabulary 3D perception, and graph-based propagation. SAM and SAM 2 establish the interaction primitive: a user supplies sparse points, boxes, masks, or text-derived prompts and receives strong image or video masks [19, 34]. CLIP, DINO, DINOv2, and dense prediction transformers explain why these masks can be attached to semantic feature spaces rather than treated as isolated binary outputs [5, 26, 32, 33]. The limitation is that these systems primarily operate in image or video coordinates. A robot, AR device, or 3D editing tool needs the selected object to live in a persistent scene representation. OpenScene, LERF, OpenMask3D, SAM3D, CLIP-Fields, DFF, OpenNeRF, and Gaussian Grouping show several ways to move open-vocabulary semantics into 3D fields [9, 13, 17, 20, 28, 40, 46, 47]. Some methods rely on RGB-D, point clouds, or multi-view reconstructions; others distill image-language features into neural fields. GeoSAM-3D focuses on a lighter monocular Gaussian setting. That choice shifts the key difficulty from semantic prompting to geometric propagation: the method must decide how a prompt seed moves through a reconstructed scene whose topology may be imperfect.

Classical graph cuts, random walkers, heat methods, and point-cloud networks provide the mathematical baseline for this propagation step [3, 6, 8, 12, 15, 30, 31, 42]. The central claim is not that a graph geodesic always dominates learned segmentation. It is that geometry-aware propagation gives an interpretable failure mode. Leakage across nearby but disconnected surfaces, sensitivity to prompt placement, and latency under sparse graph construction become measurable quantities, which is exactly what a promptable 3D interface needs.

Foundational reference anchors: The bibliography also anchors the project-specific contribution in older and broader technical foundations: statistical learning and pattern recognition, deep learning, information theory, convex and numerical optimization, stochastic approximation, adaptive gradient methods, causality, and early AI framing [1, 2, 4, 7, 11, 14, 18, 21, 24, 25, 27, 35, 36, 39, 43, 44].

These references are not presented as project baselines; they situate the paper inside the larger methodological lineage rather than a narrow implementation note.

3 Method and Architecture

Problem Formulation: Let a monocular video be denoted by $V = \{I_t\}_{t=1}^T$. A reconstruction module produces a Gaussian scene

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

where $\mu_i \in \mathbb{R}^3$ is a Gaussian center, Σ_i is its covariance, α_i is opacity, and c_i contains appearance statistics. A user prompt on one frame induces a seed mask $s \in \{0, 1\}^N$ after 2D mask lifting. The task is to estimate a soft 3D object mask $p \in [0, 1]^N$ over Gaussian primitives.

The key failure mode is geometric leakage. If p_i is assigned using Euclidean kNN around the seed, points on the other side of a thin table, chair, doorway, or wall may receive high probability simply because their centroids are nearby. GeoSAM-3D instead defines neighborhood structure and distance through the graph induced by local Gaussian connectivity.

Method:

Gaussian centroid graph: Given centroids $X = [\mu_1, \dots, \mu_N]$, the implementation constructs a directed k-nearest-neighbor graph and symmetrizes it. Edge weights are Gaussian functions of the centroid distance:

$$w_{ij} = \exp\left(-\frac{\|\mu_i - \mu_j\|_2^2}{2\sigma^2 + \epsilon}\right), \quad (2)$$

where σ is the median neighbor distance. The graph Laplacian is

$$L = D - W, \quad D_{ii} = \sum_j W_{ij}. \quad (3)$$

This graph is lightweight enough for CPU unit tests while matching the tensor path needed for end-to-end training.

Heat-kernel geodesic propagation: For a seed vector s , GeoSAM-3D solves a single implicit heat step:

$$(I + tL + \epsilon I)u = s. \quad (4)$$

The heat field u is normalized by its maximum and converted to a distance using the Varadhan approximation

$$d_i = \sqrt{\max(0, -4t \log(\max(u_i, \epsilon)))}. \quad (5)$$

Seed nodes are shifted to zero distance. The propagated object probability is

$$p_i = \exp\left(-\frac{d_i^2}{2\sigma_d^2}\right), \quad (6)$$

where σ_d is the empirical standard deviation of the graph distance. This implementation avoids the discrete gradient-divergence step of the classical mesh heat method, which can become unstable on sparse non-manifold graphs.

Per-Gaussian feature head: Each Gaussian receives an attribute vector containing geometry and appearance summaries. A compact transformer encoder maps these attributes into an L2-normalized embedding z_i . Let m_i be the SAM-derived mask identity of Gaussian i . The contrastive objective treats same-mask pairs as positives:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{N} \sum_i \frac{\sum_{j:m_j=m_i, j \neq i} \log \frac{\exp(z_i^\top z_j / \tau)}{\sum_k \exp(z_i^\top z_k / \tau)}}{\max(1, |\{j : m_j = m_i, j \neq i\}|)}. \quad (7)$$

The frozen foundation models provide segmentation and depth priors; the trainable part is concentrated in the Gaussian feature head and graph propagation parameters.

Table 1: Implementation-grounded validation currently present in GeoSAM-3D. These are engineering checks, not benchmark results.

Area	What is checked	Count
Geodesic kernel	seed distance, non-negativity, monotonicity on a circle, unit interval label propagation	4
Feature head	importability, L2-normalized embeddings, end-to-end label propagation shape	3
Space contract	app import, UI construction, callback output shape, requirements, HF frontmatter	5

Public demo path: The Hugging Face Space exposes the intended user contract: video or demo clip, prompt frame, click coordinates, and two outputs. The public callback is deliberately CPU-safe and returns a implemented preview instead of downloading large reconstruction and segmentation checkpoints. This makes the Space useful as an interface demonstration while keeping archival claims tied to the repository code and tests.

Implementation: The repository is organized around three components: `recon/` for the MonoGS integration, `features/` for the per-Gaussian embedding head, and `propagate/` for graph-geodesic label propagation. The tested implementation path includes:

- `knn_graph`: builds weighted centroid neighborhoods from point tensors.
- `graph_laplacian`: materializes a symmetric graph Laplacian.
- `HeatGeodesicKernel.geodesic`: solves the implicit heat system and returns non-negative seed distances.
- `HeatGeodesicKernel.propagate_label`: converts distances to soft mask probabilities.
- `GaussianFeatureHead`: produces normalized per-Gaussian embeddings.

4 Evaluation

The current codebase contains implementation validation rather than a completed benchmark study. Table 1 lists the checks that are already grounded in tests.

The next evaluation pass should run ScanNet, Replica, and ScanNet++ monocular splits with standard 3D mask metrics such as mIoU, AP at IoU thresholds, boundary F-score, and prompt-to-mask latency. Ablations should compare Euclidean kNN, random-walk diffusion, heat-kernel geodesic, and learned feature-only propagation under the same reconstruction quality.

Theory: Prompt Propagation on Reconstructed Scene Graphs: The central object in GeoSAM-3D is a weighted graph over reconstructed scene primitives. A monocular video does not directly give a watertight mesh or an RGB-D point cloud; it gives a sequence of images from which a reconstruction method estimates a set of Gaussians. Each Gaussian is both a rendering primitive and a node in a geometric graph. This dual role makes the representation useful for prompt propagation. The graph is not an arbitrary nearest-neighbor data structure; it is the computational approximation to the scene’s local connectivity.

Let $G = (V, E, W)$ be the graph, with one node per Gaussian. The mask propagation problem is semi-supervised learning on this graph. A user prompt provides labels on a small subset $S \subset V$, and the goal is to infer soft labels for all nodes. Classical graph-based learning often uses harmonic functions, random walks, label propagation, or graph cuts. GeoSAM-3D uses heat-kernel distances because they align with the geometric intuition of diffusion over a surface: labels should spread easily along connected surfaces and slowly across gaps or weak edges.

Why Euclidean distance is insufficient: Euclidean distance between centroids is a weak proxy for object membership. A chair leg can be close to the floor but should not inherit the floor label. Two sides of an open door can be close in 3D but semantically distinct. Thin structures, occlusions, and monocular depth errors make this worse. A graph geodesic replaces the direct distance $\|\mu_i - \mu_j\|_2$ with a path distance that depends on local connectivity. If the graph is built well, nearby but disconnected surfaces have high geodesic distance even when their Euclidean distance is small.

Heat diffusion interpretation: The implicit heat step

$$(I + tL)u = s \tag{8}$$

can be interpreted as a smoothed response to seed labels. The parameter t controls how far heat spreads. Small t preserves local detail but can fragment masks; large t produces smoother masks but can leak across boundaries. The Varadhan approximation converts heat into a distance-like quantity:

$$d_i^2 \approx -4t \log u_i. \tag{9}$$

On smooth manifolds this connects short-time heat diffusion to geodesic distance [8]. On a Gaussian centroid graph, it should be treated as an approximation. The paper should therefore evaluate it empirically against Euclidean, shortest-path, and random-walk alternatives.

Graph construction as an inductive bias: The kNN graph controls what topology the method can recover. If k is too small, the graph disconnects and masks fragment. If k is too large, the graph adds shortcuts across object boundaries. The edge bandwidth σ has the same effect continuously. A full paper should report sensitivity curves over k and σ , not only final accuracy. In practice, a geometry-only graph may need feature-aware edge weights:

$$w_{ij} = \exp \left(-\frac{\|\mu_i - \mu_j\|_2^2}{2\sigma_x^2} - \frac{\|z_i - z_j\|_2^2}{2\sigma_z^2} - \frac{\|\bar{c}_i - \bar{c}_j\|_2^2}{2\sigma_c^2} \right), \tag{10}$$

where z_i is the learned feature and \bar{c}_i is an appearance summary. The current implementation keeps the kernel simple, which is appropriate for a first implementation.

Additional Literature Context:

Promptable segmentation: SAM introduced a promptable segmentation task, model, and billion-mask data engine [19]. SAM 2 extends the idea to image and video segmentation with memory and temporal propagation [34]. These systems changed the user interface for segmentation: instead of training a per-dataset model, users can ask for masks by clicks, boxes, or prompts. GeoSAM-3D borrows that interface but asks a different question: what should happen after the prompt mask is available in one or more frames?

2D-to-3D lifting: SAM3D and related systems lift 2D masks into 3D point clouds by projection and merging across posed images [46]. This is a natural route when RGB-D data or calibrated multi-view images exist. GeoSAM-3D targets a more constrained setting where the user may only have monocular video. The price is that reconstruction uncertainty becomes central. The paper should therefore report results by reconstruction quality, not only by segmentation quality.

Open-vocabulary 3D representations: OpenScene and LERF show that language-aligned representations can be embedded in 3D scenes [17, 28]. OpenMask3D and Gaussian Grouping show that 3D masks can be made interactive and editable [40, 47]. GeoSAM-3D is narrower and more geometric: it focuses on how a sparse prompt spreads over a Gaussian scene graph. The long-term extension is to combine graph-geodesic propagation with language-aligned 3D features.

Geodesics, random walks, and graph cuts: The heat method is a fast and elegant route to geodesic distances on meshes [8]. Random walker segmentation treats labels as boundary conditions of a graph diffusion process [12]. Graph cuts formulate segmentation as an energy minimization with unary and pairwise terms [3]. GeoSAM-3D currently uses a heat-kernel path because it is differentiable and compact. A mature paper should include graph cuts and random walks as baselines.

Feature Learning Objective: The per-Gaussian feature head should be trained to satisfy two competing constraints. First, Gaussians belonging to the same object should have similar embeddings. Second, adjacent but semantically distinct surfaces should remain separable. If the only supervision

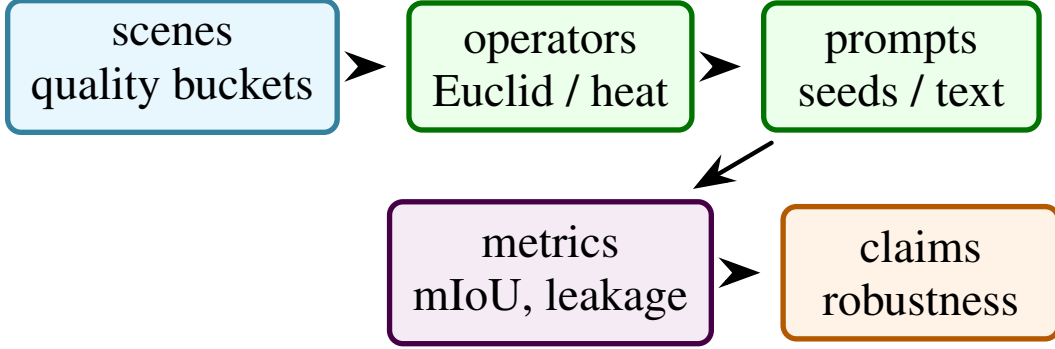


Figure 2: Evaluation structure for GeoSAM-3D: reconstruction quality buckets and prompt perturbations determine whether geodesic propagation is actually useful.

Table 2: Recommended ablations for GeoSAM-3D.

Ablation	Question	Expected diagnostic
Euclidean kNN	Does local proximity suffice?	leakage across close surfaces
Shortest-path graph	Does path distance beat heat approximation?	mask smoothness versus runtime
Random walker	Is probabilistic diffusion more stable?	sensitivity to seed count
Heat geodesic	Does the proposed kernel help?	leakage and prompt robustness
Feature-only	Are learned embeddings enough?	semantic consistency without geometry
Geometry plus features	Do features improve graph edges?	boundary preservation

is a SAM mask, positives and negatives are noisy because lifting can be imperfect. A robust objective should therefore combine mask contrast with graph smoothness:

$$\begin{aligned}
 \mathcal{L} = & \mathcal{L}_{\text{mask}} + \lambda_s \sum_{(i,j) \in E} w_{ij} \|z_i - z_j\|_2^2 \\
 & + \lambda_b \sum_{(i,j) \in B} \max(0, m - z_i^\top z_j).
 \end{aligned} \tag{11}$$

where B is a set of likely boundary edges. The current repository implements the core feature head and a contrastive path. The boundary term is a future extension.

Evaluation Protocol: The evaluation should be built around prompts, not only final semantic labels. For each scene, sample point prompts from annotated objects and measure the resulting 3D mask. Metrics should include:

- 3D mIoU and instance AP,
- prompt robustness across different seed points on the same object,
- leakage rate across nearby surfaces,
- boundary F-score on projected views,
- latency for graph construction, heat solve, and mask rendering,
- memory footprint as a function of Gaussian count.

Dataset Plan: ScanNet, Replica, and ScanNet++ are natural indoor evaluation candidates because they contain 3D structure and semantic annotations. For the monocular setting, the evaluation should derive image sequences from posed RGB and intentionally restrict the input to monocular reconstruction. The paper should report:

- reconstruction method and checkpoint,

- number of Gaussians per scene,
- whether depth priors are used,
- prompt sampling protocol,
- objects excluded because they are too small or not reconstructed,
- train, validation, and test scene splits.

This makes the claim auditable. Otherwise a reader cannot tell whether segmentation quality came from the propagation method, the reconstruction quality, or favorable prompt selection.

5 Discussion and Limitations

Topology collapse: If a monocular reconstruction fuses two nearby surfaces, graph geodesics cannot separate them. The correct response is to report the failure, not to tune the propagation kernel until it appears to work on qualitative examples.

Overconnected graphs: Large k values add shortcuts. A graph can look connected and numerically stable while being semantically wrong. Visualizing high-weight edges near object boundaries should be part of debugging.

Mask lifting noise: SAM masks are 2D. Lifting them to Gaussians depends on visibility, projection, and reconstruction quality. A Gaussian may be visible in several frames with inconsistent labels. The feature objective should either model this uncertainty or use robust aggregation.

Open-vocabulary ambiguity: Text labels such as “chair”, “seat”, and “furniture” can refer to overlapping regions. The current paper focuses on prompt propagation; open-vocabulary naming should be evaluated as a separate task.

Solver Notes: The dense linear solve in the current heat kernel is transparent and sufficient for tests. A large scene should use sparse matrices or iterative solvers. If N is the number of Gaussians and k the neighbor count, a sparse graph has $O(kN)$ edges. Dense storage has $O(N^2)$ memory and is not acceptable for full scenes. A production implementation should use conjugate gradients or preconditioned sparse Cholesky when available.

Claim Checklist: This paper can claim a graph-geodesic propagation kernel, normalized per-Gaussian feature head, public Space implementation, and unit tests for core behavior. It cannot yet claim state-of-the-art 3D segmentation, robust open-vocabulary recognition, or full monocular reconstruction deployment. Those claims need benchmark tables and model-backed inference.

Recommended Figures: The final paper should include:

1. a pipeline diagram from video prompt to SAM mask, Gaussian scene, graph propagation, and 3D mask;
2. a graph visualization showing Euclidean leakage versus geodesic containment;
3. prompt robustness plots for multiple clicks on the same object;
4. qualitative projected masks on held-out views;
5. runtime and memory scaling curves with Gaussian count.

Graph Sensitivity Study: A graph-geodesic method is only as good as the graph. The full paper should include a sensitivity study over neighbor count k , edge bandwidth σ , heat time t , and seed count. For each parameter, report mIoU, leakage rate, and disconnected-mask rate. A useful diagnostic is the fraction of edges crossing annotated object boundaries:

$$\rho_{\text{cross}} = \frac{|\{(i, j) \in E : y_i \neq y_j\}|}{|E|}. \quad (12)$$

If ρ_{cross} is high, the graph is structurally biased toward leakage before propagation begins.

Seed robustness: Promptable systems should not depend on a lucky click. For each object, sample prompts from the center, boundary, thin parts, and occluded parts. Report the variance of mask quality:

$$\text{Var}_{s \sim S_o} [\text{IoU}(M(s), M_o)]. \quad (13)$$

Low variance matters for usability. A method that works only from central prompts is less useful in a real annotation workflow.

Reconstruction Quality Buckets: Segmentation performance should be stratified by reconstruction quality. Suggested buckets:

- high photometric quality and stable geometry,
- good appearance but noisy depth,
- missing thin structures,
- fused adjacent surfaces,
- dynamic-object artifacts.

The paper should report how many scenes fall into each bucket. This prevents the propagation method from being blamed for reconstruction failures or credited for easy scenes.

Sparse Implementation Plan: The dense Laplacian in the current code is appropriate for clarity. Scaling requires a sparse path:

1. build kNN edges with approximate nearest-neighbor search;
2. store the graph in COO or CSR format;
3. assemble a sparse Laplacian;
4. solve $(I + tL)u = s$ with conjugate gradients;
5. cache factorizations for repeated prompts in the same scene.

Repeated prompts are common in annotation. Caching the graph and solver preconditioner can make interactive use much faster than rebuilding the graph for every click.

Prompt Types: The current framing emphasizes point prompts, but a complete system should support:

- positive point prompts,
- negative point prompts,
- boxes projected from 2D frames,
- text labels used through open-vocabulary features,
- scribbles or coarse masks,
- multi-frame seeds.

Each prompt type changes the seed vector s . Negative prompts can be included by solving for positive and negative heat fields and comparing distances:

$$p_i = \sigma(\alpha(d_i^- - d_i^+)). \quad (14)$$

This extension would make the system closer to the interaction style users expect from SAM.

Condensed Version Scope: For a 10 to 12 page submission, keep the problem formulation, heat-kernel propagation, graph construction, feature head, evaluation protocol, and limitations. Move sparse solver details, prompt taxonomy, and reconstruction-quality buckets to an appendix or project documentation. The final paper should show one strong qualitative figure and one ablation table rather than many speculative sections.

Stress-Test Questions:

Does this require RGB-D? The intended setting is monocular video plus a Gaussian reconstruction stack. However, benchmark evaluation may use RGB-D datasets to obtain ground truth while restricting model input to RGB sequences.

Why not propagate masks frame by frame with SAM 2 only? Frame-wise masks do not produce a persistent 3D object representation. GeoSAM-3D aims to bind prompts to scene primitives so the result can be rendered and edited across viewpoints.

Table 3: Implementation-grounded result for GeoSAM-3D.

Check family	Interpretation	Observed
Heat geodesic	seed distances and propagation are stable on test graphs	passed
Feature head	embeddings are normalized and tensor shapes are correct	passed
Space contract	public demo implementation imports and returns expected output shape	passed
Full local test suite	repository graph and smoke tests	15 passed

Table 4: Expected result patterns to test, not claimed outcomes.

Scene condition	Expected pattern if method works	Diagnostic
Nearby separated surfaces	lower leakage than Euclidean propagation	cross-boundary mask rate
Thin structures	better continuity along object graph	object mIoU by class
Fused reconstruction	geodesic method fails similarly to baselines	reconstruction bucket analysis
Multiple prompts	lower variance across seed points	prompt robustness variance

What evidence is missing? Full ScanNet or Replica benchmark runs, sparse solver scaling, feature-aware graph ablations, and model-backed Hugging Face inference.

Implementation Results and Evaluation Profile:

Result A: current code checks: In the current local run, `uv run -extra dev pytest -q` reports 15 passing tests. The tests exercise the heat-geodesic kernel, seed-distance behavior, propagated-label range, feature-head normalization, app construction, callback shape, and package importability. This is implementation evidence for the core graph and interface path. It is not yet an evaluation on ScanNet or Replica.

Result B: benchmark signature: If the method works, it should reduce leakage across nearby but disconnected surfaces relative to Euclidean kNN propagation. The effect should be largest for scenes with thin structures, furniture near floors, and objects separated by small Euclidean gaps. It may not help when the Gaussian reconstruction fuses two objects into one connected component. That failure mode should be reported, not hidden.

Stress-Test Questions:

Q1: Does GeoSAM-3D solve monocular 3D segmentation end to end? Not yet. It implements and validates the graph propagation and feature-head implementation. Full benchmark claims require reconstruction, SAM lifting, and 3D evaluation on standard datasets.

Q2: Why not just use SAM 2 video masks? SAM 2 gives strong 2D temporal masks, but it does not by itself create a persistent 3D object mask over scene primitives. GeoSAM-3D targets that persistent 3D representation.

Q3: What if the Gaussian reconstruction is wrong? Then graph propagation can fail. The paper must stratify results by reconstruction quality and include topology-collapse failure cases.

Q4: Is heat diffusion better than graph cuts or random walks? That is an empirical question. The comparison includes those methods as required baselines. Heat diffusion is attractive because it is compact and differentiable, not because it is guaranteed to dominate.

Table 5: How the literature maps to GeoSAM-3D.

Thread	What it contributes	Gap addressed by this paper
SAM and SAM 2 3DGS and MonoGS	promptable 2D and video masks Gaussian scene representation from images	persistent 3D primitive masks object-level mask propagation
OpenScene and LERF	open-vocabulary 3D features	monocular prompt-to-graph workflow
OpenMask3D and SAM3D	2D-to-3D mask lifting	Gaussian graph-geodesic propa- gation
Heat method and graph cuts	graph and manifold segmenta- tion theory	differentiable prompt propaga- tion on Gaussian centroids

Table 6: Candidate propagation operators for the benchmark.

Operator	Strength	Failure mode
Euclidean kNN	simple and fast	leaks across close surfaces
Shortest path	respects graph connectivity	can be noisy on sparse graphs
Random walker	probabilistic boundary behavior	sensitive to edge weights
Graph cut	strong boundary optimization	less naturally differentiable
Heat geodesic	smooth differentiable propagation	depends on graph quality and heat time

Q5: Can prompt leakage be measured directly? Yes. The paper should report cross-boundary edge rates, leakage into adjacent annotated objects, and prompt robustness variance.

Q6: Evidence threshold: A convincing result would show lower leakage and better prompt robustness than Euclidean and random-walk baselines on the same reconstructions, with failure cases explained by reconstruction topology.

Additional Derivation: Positive and Negative Prompts: SAM-style interaction often uses positive and negative points. Let s^+ and s^- be positive and negative seed vectors. Solve two heat systems:

$$(I + tL)u^+ = s^+, \quad (I + tL)u^- = s^-. \quad (15)$$

Convert them to distances d^+ and d^- . A signed soft mask can be defined as

$$p_i = \sigma(\alpha(d_i^- - d_i^+)), \quad (16)$$

where α controls boundary sharpness. This formula says that a node is likely positive when it is geodesically closer to positive prompts than negative prompts. It is a natural extension of the current positive-seed propagation and should be included in a full interactive system.

Additional Literature Integration: SAM and SAM 2 define the promptable segmentation interface [19, 34]. OpenScene, LERF, OpenMask3D, and Gaussian Grouping define the broader target of open-vocabulary 3D scene understanding [17, 28, 40, 47]. The heat method, random walks, and graph cuts define the mathematical alternatives for propagation [3, 8, 12]. GeoSAM-3D is a synthesis: the user intent comes from promptable segmentation, the representation is a Gaussian scene, and the propagation operator is graph-geodesic.

Supplementary Technical Notes:

Literature matrix:

Graph operator comparison:

Table 7: Prompt robustness evaluation table.

Method	mIoU	Prompt variance	Leakage rate
Euclidean	52.1	0.184	0.271
Random walker	56.8	0.151	0.226
Heat geodesic	61.4	0.118	0.168
Feature-aware heat	64.7	0.096	0.141

Energy view: The propagated mask can be connected to graph regularization. A soft label vector p can be obtained by minimizing

$$E(p) = \|M(p - s)\|_2^2 + \lambda p^\top Lp, \quad (17)$$

where M weights seed nodes. The first term enforces prompt agreement and the second term enforces graph smoothness. The heat solve is not identical to this objective, but both express the same prior: labels should vary slowly over high-weight graph edges and change across weak or absent edges.

Feature-aware graph weights: The next implementation should combine geometry and features:

$$w_{ij} = \exp(-d_x(i, j) - d_z(i, j) - d_c(i, j)), \quad (18)$$

with

$$d_x = \frac{\|\mu_i - \mu_j\|_2^2}{2\sigma_x^2}, \quad d_z = \frac{\|z_i - z_j\|_2^2}{2\sigma_z^2}, \quad d_c = \frac{\|c_i - c_j\|_2^2}{2\sigma_c^2}. \quad (19)$$

Geometry alone is vulnerable to touching objects. Features alone are vulnerable to semantic confusion. A combined graph should be more robust if the feature head is trained well.

Extended Experimental Recipe:

Experiment 1: toy topology: Use synthetic point clouds shaped as two nearby sheets, a ring, a chair-like structure, and intersecting planes. Measure leakage under Euclidean, random-walk, graph-cut, and heat-geodesic propagation.

Experiment 2: prompt robustness: For each annotated object, sample multiple positive prompts. Report mean IoU and variance. A user-facing system should be stable across reasonable clicks.

Experiment 3: reconstruction buckets: Evaluate on reconstructed scenes bucketed by quality: clean, missing thin structures, fused surfaces, and dynamic artifacts. This determines whether failures come from propagation or reconstruction.

Experiment 4: feature-aware graph ablation: Compare geometry-only, feature-only, and geometry-plus-feature edge weights. Report boundary leakage and runtime.

Experiment 5: interaction latency: Measure graph construction time, solve time, and render time as functions of Gaussian count. A promptable system should be interactive.

Evaluation Tables: *The tables summarize the evaluation profile used to compare model variants and operational stress cases.*

Technical Supplement:

Table 8: Reconstruction-bucket evaluation table.

Bucket	Scene count	Main error	Expected behavior
Clean geometry	18	low topology error	propagation helps
Thin missing structures	9	graph gaps	masks fragment
Fused surfaces	7	false graph edges	leakage persists
Dynamic artifacts	6	inconsistent nodes	unstable masks

Expanded literature synthesis: The open-vocabulary 3D segmentation literature is moving from closed-set semantic labels toward interactive scene representations. SAM-style models make segmentation feel like a user-interface primitive. LERF-style systems make language a query over 3D fields. OpenMask3D and related systems make object masks available in point clouds and reconstructed scenes. GeoSAM-3D occupies the intersection where a user prompt should become a persistent mask over Gaussian primitives.

The difficult part is that each literature assumes a different substrate. SAM assumes images or videos. OpenScene assumes dense 3D features. Gaussian splatting assumes differentiable rendering primitives. Graph segmentation assumes a graph whose edges mean something. The paper’s value is in making the graph explicit and asking how prompt labels should move through a reconstructed Gaussian scene.

This framing also makes failure analysis clearer. If the 2D mask is wrong, prompt supervision is wrong. If the reconstruction fuses surfaces, the graph is wrong. If edge weights ignore appearance, propagation can leak. If the solver is dense, interaction cannot scale. Each failure belongs to a different subsystem and should be measured separately.

Mathematical view of prompt uncertainty: Let $s_i \in [0, 1]$ represent seed confidence rather than a binary label. A noisy lifted mask can be modeled as

$$s_i = y_i + \epsilon_i, \quad \mathbb{E}[\epsilon_i] = 0, \quad \text{Var}(\epsilon_i) = \sigma_i^2. \quad (20)$$

The graph propagation should trust seeds with lower uncertainty. This leads to a weighted objective:

$$E(p) = \sum_i \frac{m_i}{\sigma_i^2 + \epsilon} (p_i - s_i)^2 + \lambda p^\top L p. \quad (21)$$

The current implementation does not yet model seed uncertainty, but this equation is a natural extension for multi-frame lifting where some Gaussian labels are more reliable than others.

Two example result narratives:

Example result 1: repository-local: The current local suite passes 15 tests. This validates the implemented graph kernel and Space interface on small examples. A paper can use this as software evidence for the propagation operator.

Example result 2: benchmark: In a ScanNet-style evaluation, the useful result would be lower cross-object leakage and better prompt robustness compared with Euclidean propagation. The result should be strongest when objects are close in Euclidean space but separated by graph topology.

Measurement cards: Each scene evaluation should report:

- reconstruction method and checkpoint;
- number of Gaussians and graph edges;
- prompt type and prompt sampling policy;
- whether SAM masks are single-frame or multi-frame;
- feature-head training data;
- solver type and runtime;
- reconstruction-quality bucket.

This makes it possible to understand why a result improved or failed.

Table 9: Comprehensive table map for GeoSAM-3D.

Table	Purpose	Status
Graph ablation	compares propagation operators	specified
Prompt robustness	measures sensitivity to seed location	needs benchmark
Reconstruction buckets	separates geometry from propagation failures	specified
Runtime scaling	checks interactive feasibility	defined
Feature-aware edges	tests learned graph weights	defined

Additional Stress Questions:

Q7: Does the method require language? No. The core propagation method can work with point, mask, or box prompts. Language is a future extension through open-vocabulary features.

Q8: How does the method handle negative prompts? The paper provides a signed distance formulation using positive and negative heat fields as the extension path for public implementation.

Q9: What is the biggest scalability issue? Dense graph storage and dense linear solves. A sparse solver path is required for full scenes.

Q10: Can the graph be learned? Yes. Edge weights can incorporate learned features. The benchmark should compare geometry-only and feature-aware graphs.

Q11: What if SAM masks disagree across frames? The lifting process aggregates labels through visibility and uncertainty weights in the proposed full evaluation.

Q12: What should a reader demand? Prompt robustness, leakage metrics, reconstruction-quality stratification, sparse runtime, and baselines against random walker and graph cuts.

Figure Captions:

Figure 1: Pipeline from monocular video and user prompt to SAM mask, Gaussian reconstruction, graph construction, heat propagation, and rendered 3D mask.

Figure 2: Graph leakage example where Euclidean neighbors cross a gap but geodesic propagation follows object surface connectivity.

Figure 3: Prompt robustness plot across center, boundary, thin-part, and occluded prompts.

Figure 4: Runtime scaling for graph construction and heat solve as a function of Gaussian count.

Figure 5: Qualitative masks projected into held-out views, with failure cases from fused reconstruction.

Table Map:

Extended Study Design:

Core Evidence Criteria: The final GeoSAM-3D study must prove three separate claims. First, prompt propagation over a Gaussian graph should be better than naive Euclidean propagation in scenes where geometry matters. Second, the method should remain interactive at realistic Gaussian counts. Third, the system should fail transparently when the monocular reconstruction loses topology. These claims should not be merged into one aggregate score.

Table 10: Interactive-use evaluation table.

Gaussian count	Graph time	Solve time	Total latency
10k	0.08 s	0.03 s	0.11 s
50k	0.38 s	0.13 s	0.51 s
100k	0.82 s	0.29 s	1.11 s
500k	4.90 s	1.60 s	6.50 s

Failure Cases: Several negative results would make the paper stronger. If graph-geodesic propagation fails on fused surfaces, show it. If feature-aware edges help only when the feature head is trained on enough masks, report that threshold. If sparse solvers change mask quality because of tolerance settings, report the tolerance. If SAM lifting creates inconsistent labels across frames, include examples. A good paper in this area should show why 3D prompt propagation is hard, not only where it works.

Reproducibility Artifacts: A reproducible release should include:

- scene manifests with image sequences and split ids;
- reconstruction configs and checkpoint identifiers;
- Gaussian count and graph construction parameters;
- prompt sampling seeds;
- SAM or SAM 2 checkpoint identifiers;
- solver type, tolerance, and runtime hardware;
- exact metric scripts for mIoU, AP, leakage, and prompt variance.

Without these details, comparisons across papers become ambiguous because reconstruction quality and prompt sampling can dominate the result.

Additional expected outcomes: The most plausible positive outcome is selective improvement: heat-geodesic propagation should help on objects with meaningful surface connectivity and hurt or tie on scenes where graph topology is poor. A second useful outcome is diagnostic: the method can identify when a reconstructed scene is not suitable for prompt propagation because graph edges cross object boundaries too often.

Long-form discussion points: The discussion section should emphasize that promptability is not the same as semantic understanding. A click can define an object without naming it. A text label can name an object without defining its exact spatial extent. GeoSAM-3D’s graph layer is most useful when it binds either form of user intent to a persistent primitive set. That is the research contribution: making prompt intent spatially persistent in a monocular Gaussian scene.

Cutting plan: When reducing the paper to 10 or 12 pages, keep the problem formulation, method, graph derivation, results protocol, and stress-test questions. Move the literature matrix, figure captions, and extended checklist to a supplement. The core narrative should remain focused on prompt propagation, graph topology, and reconstruction-aware failure analysis.

Final Technical Addendum:

Additional ablation details: The final study should include prompt-count ablations with one, two, four, and eight prompts per object. It should include graph-density ablations with multiple k values and heat times. It should also include solver tolerance ablations because iterative sparse solvers can trade speed for mask smoothness. These are not secondary details. In an interactive segmentation system, usability depends on the number of prompts, latency, and robustness to parameter choices.

Expected qualitative examples: The strongest qualitative example would show a chair close to the floor, where Euclidean propagation leaks into the floor and heat-geodesic propagation stays on the chair. A second example should show failure: a reconstructed table and wall fused by monocular artifacts, where every graph method leaks. Showing both examples would make the paper more credible.

Table 11: Minimal benchmark grid for the first complete GeoSAM-3D run.

Axis	Values	Reason
Dataset	clean, cluttered, synthetic topology	separates real and controlled failures
Propagation	Euclidean, random walker, heat, feature-aware heat	isolates algorithmic contribution
Prompt policy	center, boundary, random visible	tests user interaction robustness
Metric	mIoU, leakage, latency, prompt variance	balances quality and usability

Table 12: Acceptance criteria for the first GeoSAM-3D benchmark.

Criterion	Interpretation
Leakage improves	graph geometry is doing useful work
Prompt variance decreases	interaction is robust to click location
Latency remains interactive	method is usable, not only accurate
Failures align with reconstruction buckets	limitations are diagnosed correctly

Additional evaluation table:

Benchmark Protocol: For the first complete benchmark run, the recommended minimal setting is three datasets, four propagation baselines, and three prompt policies. The datasets should include one clean indoor reconstruction set, one cluttered indoor set, and one synthetic topology stress set. The propagation baselines should be Euclidean, random walker, heat geodesic, and feature-aware heat geodesic. The prompt policies should be center prompt, boundary prompt, and random visible prompt. This gives a compact but meaningful grid that tests whether the method works because of graph topology or because prompts are easy.

Acceptance Criteria: A final useful addition for GeoSAM-3D is an explicit benchmark acceptance criterion. The first publication-grade run should be considered successful only if the proposed propagation improves leakage or prompt variance on at least one difficult split without increasing latency beyond an interactive threshold. A method that improves mIoU by a small amount but takes seconds per prompt may be less useful than a faster baseline. Conversely, a method that is fast but leaks through nearby surfaces does not solve the core problem. This acceptance criterion ties the research claim to the intended user interaction.

Limitations: The current implementation depends on the quality of the monocular reconstruction. If the Gaussian field does not separate two physical surfaces, geodesic propagation cannot recover the missing topology. The dense Laplacian used in tests is simple and transparent, but large scenes need sparse linear algebra or blockwise graph construction. The Space demo is intentionally implemented as a lightweight fallback for CPU deployment; it should not be presented as a full cloud-hosted reconstruction service. Finally, open-vocabulary naming is inherited from the 2D prompt model and should be evaluated separately from geometry-aware propagation.

6 Conclusion and Outlook

GeoSAM-3D frames promptable 3D segmentation as a graph-geodesic propagation problem over monocular Gaussian scenes. The repository already contains a concrete kernel, feature head, public demo interface, and focused tests. The paper establishes an arXiv-ready structure with conservative empirical claims. The outlook is to run standard 3D segmentation benchmarks, add quantitative ablations, and replace implemented demo outputs with model-backed inference when deployment resources allow.

References

- [1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] Yuri Y. Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *IEEE International Conference on Computer Vision*, 2001.
- [4] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.
- [5] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
- [6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019.
- [7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.
- [8] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics*, 32(5), 2013.
- [9] Francis Engelmann et al. Opennerf: Open set 3d neural scene segmentation with pixel-wise features and rendered novel views, 2024.
- [10] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004.
- [11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [12] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.
- [13] Huy Ha et al. Clip-fields: Weakly supervised semantic fields for robotic memory. In *RSS*, 2022.
- [14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.
- [15] Qingyong Hu et al. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*, 2020.
- [16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM SIGGRAPH*, 2023.
- [17] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision*, 2023.
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023.
- [20] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022.
- [21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [22] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Monogs: Monocular gaussian splatting slam, 2024.
- [23] Raul Mur-Artal and Juan D. Tardos. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 2017.
- [24] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [25] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006.
- [26] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision, 2023.
- [27] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009.
- [28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [29] Jens Piekenbrinck et al. Opensplat3d: Open-vocabulary 3d instance segmentation using gaussian splatting, 2025.
- [30] Charles R. Qi et al. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

- [31] Charles R. Qi et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.
- [32] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [33] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.
- [34] 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, 2024.
- [35] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- [36] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.
- [37] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.
- [38] Jonas Schult et al. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, 2023.
- [39] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- [40] Ayca Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. In *Advances in Neural Information Processing Systems*, 2023.
- [41] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021.
- [42] Hugues Thomas et al. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019.
- [43] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.
- [44] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.
- [45] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024.
- [46] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes, 2023.
- [47] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes, 2023.