Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses

Damian Bowness

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

June 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Damian Bowness**

Entitled: **Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Abdelhak Bentaleb

_____ Examiner
Dr. Yiming Xiao

_____ Supervisor
Dr. Charalambos Poullis

Approved by _____
Dr. Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses

Damian Bowness

3D reconstruction is a foundational component in robotics and autonomous systems, enabling machines to perceive and interpret their environment for tasks such as navigation, obstacle avoidance, and motion planning. As these systems increasingly operate in real-time and dynamic environments, the need for efficient and robust scene understanding becomes paramount. Among recent innovations, 3D Gaussian Splatting (3DGS) has gained attention for its ability to reconstruct photorealistic 3D scenes with high efficiency and support for real-time novel view synthesis. Unlike traditional mesh- or voxel-based representations, 3DGS models scenes using a compact set of anisotropic Gaussians, each carrying spatial and appearance information, which are then splatted into the image plane during rendering.

However, a persistent challenge in such learned representations is uncertainty due to insufficient, ambiguous, or occluded data in the input views i.e. epistemic uncertainty. This can result in visual artifacts, especially when rendering novel viewpoints that diverge significantly from the training set. Existing methods such as BayesRays, designed for NeRF-based models, address this issue via probabilistic ray-based sampling and post-hoc filtering, but require retraining.

In this work, we propose a gradient sensitivity-based filtering framework for 3DGS that mitigates epistemic artifacts in real-time without the need for model retraining. Specifically, we introduce a novel sensitivity score that quantifies the directional gradient of pixel color with respect to spatial perturbations at the point of ray-Gaussian intersection. This score captures the local instability in the rendering process caused by insufficient coverage or ambiguity in training views. By computing this score directly within the existing rendering pipeline, we enable on-the-fly filtering of Gaussians whose contributions are deemed unstable or unreliable.

Our approach can be applied to any pre-trained 3DGS model, making it highly practical for deployment in real-time systems. We evaluate our method on challenging indoor and outdoor scenes, including those from the Deep Blending and NeRF-On-the-Go datasets, and show that it effectively suppresses rendering artifacts. Notably, our filtering substantially improves visual quality, realism, and consistency compared to BayesRays, while avoiding the overhead of additional training or scene-specific tuning. This makes our method particularly suited for robotics and AR/VR applications where rapid adaptation and robustness to viewpoint changes are critical.

Acknowledgments

I would like to begin by expressing my deep gratitude to my supervisor, Charalambos (Charis) Poullis, for offering me the opportunity to pursue this research and for trusting me to work independently. That trust gave me the space to grow as a researcher and problem-solver, and I am truly thankful for it.

To my parents, thank you for your unwavering support. Mom, your snacks and late-night company turned stressful deadlines into memories I'll always cherish. Dad, thank-you for patiently listening to me as I stumbled through code, and being there to catch the moment when the lightbulb finally went on.

To my sister, Camille, thank you for the grad school commiseration and solidarity. It really helped to know someone else had been through the emotional rollercoaster and didn't fall off. And to my brother-in-law, Motoki, thank you for your steadfast belief that "I'll be fine", maybe even better than fine. Your quiet confidence gave me a lot of strength when I needed it.

To my partner, Lindsay, thank you for all the ways you stood by me: through the restless nights, the second-guessing, and the times I almost lost my footing. (Speaking of footing, shout out to Wren for getting me out climbing and the midnight Mario Kart sessions that helped me stay sane when deadlines and stress were piling up!)

And to my friends, Bobby, Rohan, Mikey G, and everyone else who stuck around through all the years I kept saying "Next year, I'll be done", thank you. I'm incredibly grateful for your patience. I can't wait to finally make good on all those missed weekends!

This work is as much a product of my community's support as it is of my own efforts. This thesis wouldn't exist without all of you. Thank you all.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

3D reconstruction is the process of generating digital three-dimensional representations of real-world objects, environments, or scenes based on information acquired from two-dimensional (2D) inputs such as photographs, videos, or sensor data such as LiDAR [18, 42, 51, 56] (see Figure 1.1). This process entails estimating the geometric structure, spatial configuration, and sometimes photometric attributes, such as texture and color, of physical scenes. The resulting digital models can be visualized, analyzed, or manipulated in a range of downstream applications.

A key difficulty in 3D reconstruction lies in the joint estimation of both spatial and photometric properties of the scene [70]. This includes recovering depth, orientation, and scale, along with capturing photometric details such as color, reflectance, and surface texture, all from inherently flattened sensor measurements [10]. Imaging devices typically record 2D projections of the 3D world, effectively discarding not only direct depth cues but also view-dependent appearance variations that are critical to realistic modeling and rendering [59]. Accurately capturing these properties is essential for generating photo-realistic images and ensuring consistency in applications like augmented reality, where synthetic content must seamlessly blend with real-world visuals. As a result, 3D reconstruction involves multiple interconnected inference problems. Spatial inference focuses on recovering the lost third dimension from 2D projections using geometric cues across multiple views, calibrated camera parameters, or learned priors [58]. While, photometric inference entails estimating appearance-related properties such as surface reflectance, material-dependent shading,

and view-dependent effects like specular highlights and reflections. These photometric components are essential for ensuring realistic rendering, physical accuracy, and perceptual consistency in downstream applications. As a foundational challenge in computer vision, 3D reconstruction enables intelligent systems to perceive, interpret, and interact with the physical world in three dimensions, supporting advanced reasoning, simulation, and decision-making across domains such as autonomous robotics, augmented and virtual reality (AR/VR), medical diagnostics, and cultural heritage digitization [46, 50, 53, 62].

This capacity to digitize spatial structure with high fidelity underpins the broader paradigm of digital twinning. A digital twin is a virtual, dynamic replica of a physical object, system, or environment that reflects real-time or near-real-time conditions [57]. These virtual counterparts enable comprehensive simulation, predictive modeling, and intelligent control across a range of domains [13]. In robotics, for example, digital twins facilitate system testing, teleoperation, and adaptive motion planning by mirroring the physical robot and its environment in a virtual space [2]. In smart cities, digital twins of infrastructure and traffic systems support energy optimization and emergency response planning [4]. In industrial manufacturing, digital twins enable predictive maintenance and process optimization by monitoring equipment state and simulating failures before they occur [61]. In the medical domain, patient-specific digital twins derived from CT or MRI scans are used to simulate surgical interventions or assess treatment response [24, 63]. Furthermore, in emerging applications such as AR/VR and the metaverse, digital twins furnish the spatial scaffolding necessary for immersive and responsive user experiences [46].

3D reconstruction plays an indispensable role in the creation and maintenance of digital twins. It provides the geometric backbone: the shape, topology, and spatial arrangement of the physical entities being digitally mirrored. Without accurate and high-resolution 3D models, digital twins lack the fidelity required for spatial analysis and realistic simulation [35]. Modern 3D reconstruction pipelines not only capture static geometry but also support dynamic updates, enabling virtual models to evolve in real time alongside their physical counterparts [2], [4]. For instance, in the construction industry, progressive 3D scans of a building site can be used to update its digital twin and detect deviations from architectural plans [5]. Additionally, 3D reconstruction supports semantic enrichment by integrating object categories, material properties, and environmental context into the

model, thereby enhancing interpretability and functionality [27, 31, 69].

Moreover, reconstructed models serve as inputs to simulation engines, where spatial fidelity is essential for tasks such as physics-based reasoning, robotic manipulation, or visibility estimation. For example, in autonomous driving, reconstructed 3D maps enable accurate path planning and obstacle avoidance by modeling drivable surfaces, traffic infrastructure, and dynamic agents [33]. In surgical robotics, 3D reconstructions of patient anatomy allow for preoperative rehearsal and intraoperative guidance, reducing the risk of procedural error [47].

Ultimately, 3D reconstruction serves as the critical bridge between physical perception and digital representation. By enabling the translation of real-world spatial data into actionable digital models, it underlies the practical realization of high-resolution, semantically rich digital twins. As the demand for intelligent, spatially aware systems continues to grow, the role of 3D reconstruction will remain central to future advancements across scientific, industrial, and societal domains.

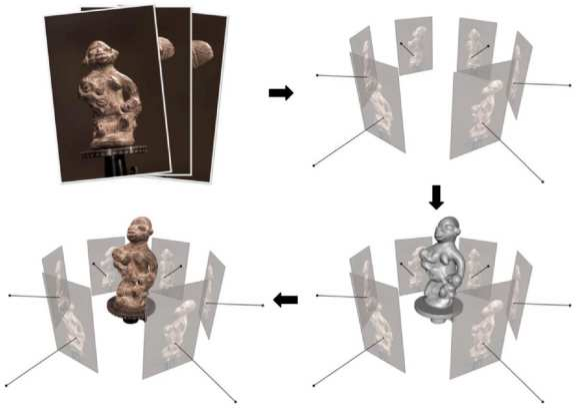


Figure 1.1: Image-based geometry reconstruction pipeline. Source: [22].

1.0.1 Contributions

The result of this work is the paper "Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses", which is currently under review. In this work, we propose a novel real-time, render-aware filtering method that improves the visual quality of 3D Gaussian Splatting (3DGS) under extreme out-of-distribution (OOD) camera poses. Our key insight is that epistemic artifacts in novel-view synthesis, such as directional blurring, smearing, and ghosting, are

often linked to unstable or ambiguously placed Gaussians, particularly in anisotropic regions poorly supported by training views.

To address this, we introduce a gradient sensitivity score that quantifies the directional instability of each Gaussian by measuring the norm of the pixel color gradient with respect to the 3D position of ray-Gaussian intersections. This sensitivity is computed in a rotation-aligned and scale-invariant space, allowing us to identify and filter Gaussians whose anisotropic geometry makes them particularly prone to hallucinating geometry under view shifts. The method is integrated directly into the rendering pipeline and applied in real-time without requiring retraining or external supervision.

To validate the effectiveness of our filtering approach, we also introduce two complementary non-reference evaluation metrics:

- **CLIP Alignment Score (CAS)**: A semantic fidelity measure based on prompt-conditioned CLIP similarity to detect alignment between rendered and expected scene content.
- **SAM Reference Similarity Score (SRSS)**: A spatial coherence measure that matches masked scene regions against reference images using SAM and CLIP embeddings.

Together, these metrics allow us to capture both high-level semantic consistency and fine-grained visual stability revealing how our render-aware filtering method preserves object identity through selective feature suppression.

Our technical contributions are:

- A rotation-aligned, scale-invariant filtering module that computes directional gradient sensitivity at ray-Gaussian intersections to detect and suppress unstable Gaussians in real-time. This is integrated directly into the 3DGS pipeline without requiring retraining.
- A pair of complementary semantic evaluation metrics, SRSS and CAS, designed to assess object-centric fidelity and alignment with semantic intent in open-world settings, enabling evaluation without pixel-level ground truth by focusing on semantically coherent regions.
- Extensive quantitative and qualitative evaluation across real-world datasets (Deep Blending and Nerf-on-the-go), showing improved performance under extreme OOD viewpoints relative to state-of-the-art baselines such as BayesRays.

Chapter 2

Preliminaries and Related Work

In computational photography, computer vision, and graphics, the plenoptic function models how light varies with position, direction, wavelength, and time. Introduced in 1991, it formalizes the concept of capturing all light in a scene by representing the observed intensity and chromaticity as a function of these variables. [1]. Mathematically, it is defined as

$$P = P(V_x, V_y, V_z, \theta, \phi, \lambda, t) \tag{1}$$

where (x, y, z) specifies the spatial position of the observer, (θ, ϕ) denote the angular direction of the ray, λ encodes wavelength (or color), and t captures temporal variation. This seven-dimensional function encapsulates all possible visual information from every conceivable viewpoint.

Due to its high dimensionality, direct representations of the plenoptic function are computationally infeasible. In practice, simplifications are often employed, such as fixing the wavelength or time dimensions, resulting in lower-dimensional approximations like light fields or lumigraphs [16, 29]. These models enable applications like image-based rendering and novel view synthesis by capturing how light radiates from the scene into the camera. Importantly, the plenoptic function provides a theoretical underpinning for recent advances in neural rendering, where high-fidelity image generation is achieved without explicit 3D geometry.

One prominent neural rendering method is Neural Radiance Fields (NeRF) [36]. NeRFs model 3D scenes implicitly using multilayer perceptrons (MLPs), which map 3D spatial coordinates and

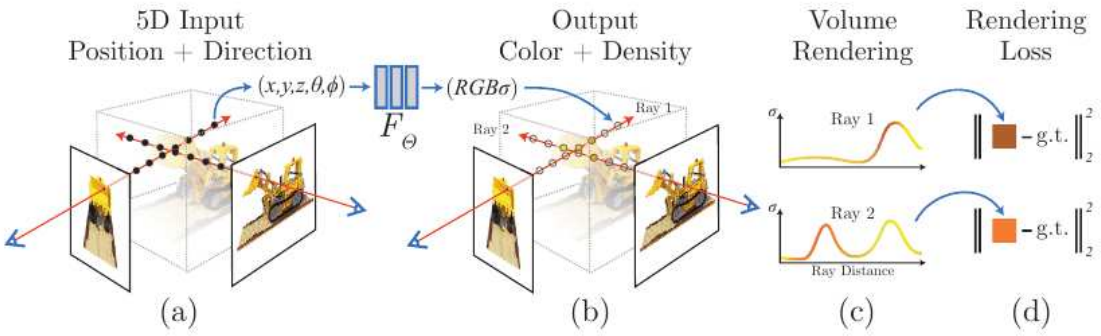


Figure 2.1: **NeRF Training Pipeline**: **(a)** A ray cast from a camera samples 3D points (x, y, z) along its direction (θ, ϕ), forming a five-dimensional coordinate input to the neural network F_Θ. **(b)** The MLP predicts a volume density σ and view-dependent radiance c at each sampled point; these values are composited to produce a synthetic pixel color for each input view. **(c)** For two example rays, the predict density σ_t is plotted against ray distance t; these profiles determine transmittance weights used in the rendering equation. **(d)** The integrated color C_{pred} for each ray is compared to the ground-truth pixel C_{gt} via an ℓ_2 photometric loss $||C_{\text{pred}} - C_{\text{gt}}||^2$, guiding gradient updates F_Θ. Source: [9].

2D viewing directions to volumetric density and color. Formally, a NeRF approximates the function

$$f(x, y, z, \theta, \phi) \rightarrow \sigma, \mathbf{c} \tag{2}$$

where σ is the volumetric density and $\mathbf{c}$ is the emitted color. By optimizing this function using a differentiable volumetric rendering technique, NeRF can synthesize photorealistic views of a scene from arbitrary perspectives (see Figure 2.1).

NeRF's main innovation is its ability to model continuous volumetric fields without relying on discrete geometry like meshes or point clouds. Instead, the entire scene is encoded in the neural network weights, enabling compact representation and impressive visual quality. However, NeRF faces several limitations, including slow training and rendering times, high computational overhead, and difficulties with generalization to unobserved viewpoints or dynamic scenes. Recent enhancements such as FastNeRF, Instant-NGP, and Dynamic NeRF have aimed to address these shortcomings through architectural improvements and more efficient sampling techniques [37, 44, 48]. Despite these efforts, NeRF's reliance on implicit representations can make integration with existing geometry pipelines more challenging.

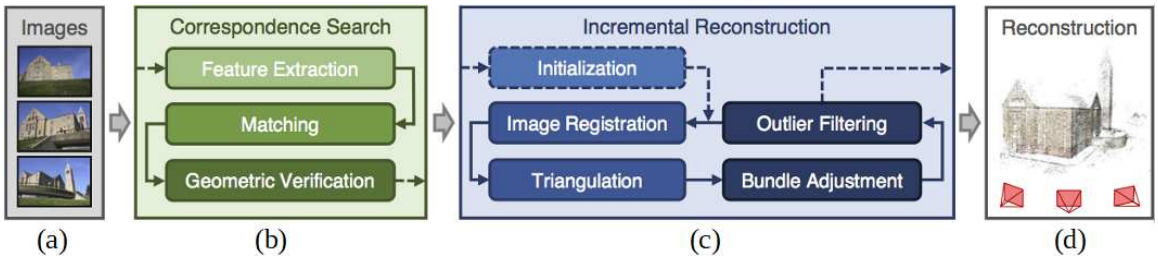


Figure 2.2: **COLMAP's incremental SfM pipeline**: **(a)** A set of overlapping images of a static scene are used as input, with or without known camera intrinsics. **(b)** Features are extracted and matched across image pairs, followed by geometric verification to filter inconsistent correspondences. **(c)** A sparse 3D model is built by initializing with a seed image pair and incrementally registering new images via triangulation and bundle adjustment. **(d)** Final Output. The resulting scene is a globally consistent sparse point cloud with registered camera poses, suitable for further dense reconstruction. Source: [54].

In contrast, COLMAP is a mature, geometry-based pipeline widely used for Structure-from-Motion (SfM) and Multi-View Stereo (MVS), see Figure 2.2. COLMAP constructs sparse and dense 3D reconstructions from unordered image collections [54]. It begins with keypoint detection and matching, typically using Scale-Invariant Feature Transform (SIFT) features, followed by incremental SfM to estimate camera intrinsics and extrinsics. A global optimization step called bundle adjustment is employed to refine camera poses and 3D point positions by minimizing reprojection error across multiple views. For high-fidelity reconstructions, COLMAP also incorporates MVS techniques that estimate per-pixel depth and normals, resulting in dense point clouds and textured 3D models.

COLMAP's robustness to varying scene conditions, wide adoption in academic and industrial settings, and compatibility with a range of sensors make it a powerful tool in photogrammetry and robotics [43, 55]. However, its reliance on discrete point representations can face limitations in flexibility due to their inability to smoothly interpolate geometry, model view-dependent appearance, or easily adapt to scene modifications [20]. They are also inefficient at scale, requiring large storage and exhaustive processing of points during rendering [3]. Neural methods address these issues by offering continuous, compressed, and adaptive representations that better support high-quality, real-time, and editable rendering. Nonetheless, COLMAP remains an essential component in hybrid pipelines that aim to bridge geometric and learning-based techniques.

3D Gaussian Splatting (3DGS) [25] represents a compelling hybrid approach that combines

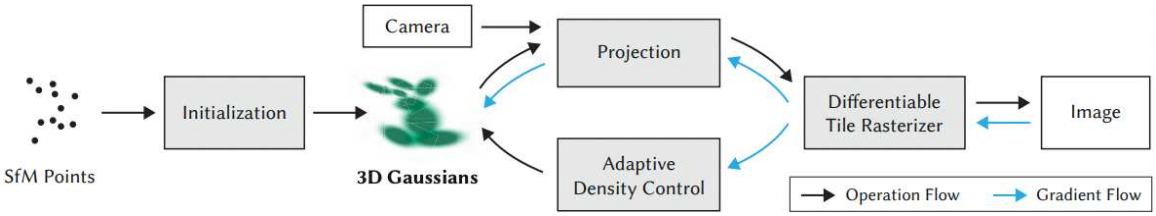


Figure 2.3: **3DGS pipeline**: **(a) Initialization** The pipeline begins with a sparse point cloud from SfM, which is used to initialize a set of 3D Gaussian primitives with position, scale, color, and opacity. **(b) Differentiable Rendering**: Each Gaussian is projected onto the image plane using a camera model and rendered via a tile-based differentiable rasterizer. **(c) Adaptive Optimization**: The parameters of each Gaussian are refined using gradient-based optimization, while an adaptive density control module prunes or duplicates Gaussians to improve coverage and efficiency. **(d) Real-Time Output** The result is a compact and optimized set of 3D Gaussians enabling real-time, high-fidelity rendering across novel views. Source: [25].

geometric initialization with data-driven rendering. Rather than modeling scenes with voxels or meshes, 3DGS uses a sparse set of anisotropic 3D Gaussian primitives to encode spatial and appearance information. Each Gaussian is parameterized by its position, orientation, scale (reflecting anisotropy), color, and opacity. These Gaussians are then projected to the image plane in a process known as splatting, where their contributions to each pixel are accumulated using a differentiable rendering function.

The 3DGS pipeline relies heavily on a strong geometric initialization, for which COLMAP is commonly used [12]. The SfM step in COLMAP provides accurate camera intrinsics, extrinsics, and a sparse point cloud, which are used to initialize the position and covariance structure of each Gaussian. Once initialized, the splatting-based rendering allows for photometric optimization, refining Gaussian parameters to minimize rendering error across input views (see Figure 2.3).

Unlike NeRF, which learns the plenoptic function implicitly through neural network parameters, 3DGS approximates it explicitly. Each Gaussian acts as a localized light emitter with directional and spatial extent, approximating the contribution of a small region in the plenoptic domain. When combined, these Gaussians form a piecewise-continuous and differentiable approximation of the plenoptic function: continuous and smooth within local regions, with well-defined gradients that enable optimization through backpropagation [66]. During rendering, Gaussians project radiance along cones defined by their anisotropy and orientation, efficiently simulating how light propagates through space and contributes to different camera rays [67].

This explicit and compact representation allows 3DGS to achieve real-time rendering speeds while maintaining high visual fidelity [25]. It also makes the method well-suited for dynamic and interactive applications such as virtual reality, digital twins, and autonomous systems. Nevertheless, 3DGS shares some challenges with its neural counterparts, including difficulties modeling fine textures, transparent surfaces, and thin structures [8, 30, 65]. These limitations are often exacerbated by epistemic uncertainty in underconstrained regions i.e. areas where limited or ambiguous input data leads to unstable parameter estimates. Small perturbations in Gaussian position, scale, or orientation can produce noticeable shifts in rendered appearance. To mitigate this, we introduce a gradient sensitivity analysis that quantifies these instabilities and filters unreliable contributions during rendering, improving both robustness and visual consistency.

Chapter 3

Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses

The following is a verbatim copy of the manuscript currently under review, titled Extreme Views: 3DGS Filter for Novel View Synthesis from Out-of-Distribution Camera Poses, authored by Damian Bowness and Charalambos Poullis.

3.1 Abstract

When viewing a 3D Gaussian Splatting (3DGS) model from camera positions significantly outside the training data distribution, substantial visual noise commonly occurs. These artifacts result from the model's lack of training data and supervision in these extrapolated regions, leading to uncertain density, color, and geometry predictions. Such visual instability disrupts user immersion and realism, posing significant challenges to interactive multimedia applications such as gaming, immersive virtual environments, and personalized content creation.

To address these challenges and enhance user experiences, especially in interactive and personalized systems, we propose a novel real-time render-aware filtering method. Our approach leverages sensitivity scores derived from intermediate gradients, explicitly targeting instabilities caused

Figure 3.1: **Left**: An unfiltered render of the 3DGS model [25] of the Playroom scene, **Right**: A filtered render using our method from the same viewpoint ($\tau_{grad.} = 10^{-5}, \tau_{ratio} = .5$). To facilitate direct comparisons and highlight the impact of individual components, we consistently reference this representative scene and viewpoint throughout the main text. Additional examples can be found in Appendix B.

by anisotropic orientations rather than isotropic variance. This filtering method directly addresses the core issue of generative uncertainty, allowing generative multimedia systems to maintain high visual fidelity even when users freely navigate outside the original training viewpoints.

Experimental evaluation demonstrates that our method substantially improves visual quality, realism, and consistency compared to existing Neural Radiance Field (NeRF)-based approaches such as BayesRays. Critically, unlike methods requiring extensive post-hoc retraining or fine-tuning, our filter seamlessly integrates into existing 3DGS rendering pipelines in real-time. Consequently, it empowers generative multimedia applications to produce robust, realistic, and personalized visual content dynamically, thereby enhancing user immersion and interactivity.

3.2 Introduction

Robust and consistent novel-view synthesis is critical for immersive applications such as gaming, virtual reality, and personalized media creation. High-quality scene rendering from arbitrary viewpoints enhances user experience and realism, enabling more engaging and believable environments. However, the presence of visual artifacts, particularly in regions not well covered by training

data, can significantly reduce perceived quality, breaking immersion and degrading interaction. Ensuring stability and coherence in novel-view synthesis is therefore essential for delivering reliable and high-fidelity visual content in generative multimedia systems.

Recent advances in neural rendering, such as Neural Radiance Fields (NeRF)[8] and 3D Gaussian Splatting (3DGS)[6], have enabled high-quality photorealistic view synthesis. Among these, 3DGS is particularly attractive due to its explicit scene representation and real-time rendering capabilities. However, when rendering from viewpoints significantly outside the bounds of the original training images, both NeRF and 3DGS models often produce substantial visual artifacts, including scattered points, flickering colors, and blurry or inconsistent geometry. These artifacts arise because the models lack training data and supervision in these extrapolated regions, resulting in high uncertainty in predicted density, color, and geometry. For NeRF, this may manifest as unstable radiance field outputs, while in 3DGS, it appears as disorganized splats floating in space or visually implausible surfaces. Essentially, the model is forced to hallucinate content without reliable guidance, leading to incoherent and noisy renderings. Existing solutions attempt to address this through additional training data or supervision [3, 12, 14], but such approaches are computationally expensive and impractical for real-time applications. This highlights the need for lightweight, render-time techniques that can dynamically suppress uncertainty-driven artifacts and enhance robustness in novel-view synthesis.

We propose a novel real-time, render-aware filtering method that enhances 3DGS rendering quality in extreme out-of-distribution (OOD) views (see Fig. 3.3) by leveraging a sensitivity score derived from intermediate color gradients. Our method computes the gradient of the pixel color with respect to the 3D position of each ray-Gaussian intersection point. A score based on this gradient captures the directional instability of Gaussians and identifies those that are likely to produce artifacts, particularly due to anisotropic elongation. By applying filtering based on this sensitivity measure, our approach dynamically suppresses visually unstable Gaussians without requiring retraining or auxiliary data. The filter operates in a rotation-aligned, scale-independent space, targeting instabilities caused by orientation rather than magnitude.

We integrate our sensitivity-based filter into the standard 3DGS rendering pipeline and evaluate its effectiveness across several real-world datasets. Quantitative and qualitative results show

that our method significantly improves visual quality compared to state-of-the-art NeRF-based approaches such as BayesRays, without requiring additional post-hoc training. This lightweight, drop-in solution enhances the robustness, realism, and consistency of novel-view synthesis, making it well-suited for interactive and performance-sensitive applications in generative multimedia environments.

Our technical contributions are:

- A novel sensitivity-based filtering approach explicitly designed for real-time suppression of anisotropic instabilities in generative multimedia
- Integration of a render-aware gradient sensitivity filtering directly into the 3D Gaussian Splatting pipeline without requiring retraining or auxiliary supervision.
- Comprehensive quantitative and qualitative validation across diverse scenes demonstrating significant improvements in perceptual quality under extreme viewpoints and occlusions.

Our method substantially advances generative multimedia systems' robustness, realism, and interactivity, making it especially suited for performance-sensitive and immersive applications.

3.3 Related Work

3.3.1 Neural Rendering and NeRF

Neural rendering techniques, particularly Neural Radiance Fields (NeRF) [36], have significantly advanced novel-view synthesis by implicitly representing 3D scenes as continuous volumetric fields encoded via neural networks. NeRF optimizes a multi-layer perceptron (MLP) by sampling along camera rays to predict color and opacity values, enabling photorealistic image rendering from unseen viewpoints. However, this implicit representation poses challenges such as high computational cost and slow rendering times, limiting their real-time application potential.

To mitigate uncertainties in NeRF reconstructions, post-hoc uncertainty quantification methods have been introduced. Notably, Goli et al. [15] proposed BayesRays, a Bayesian framework using a Laplace approximation of the Hessian matrix to quantify systematic uncertainties inherent in NeRF reconstructions. Their method employs a spatial deformation field, addressing the highly correlated

Figure 3.2: A frame (0640) from the Arc-de-Triomphe scene. **Top:** unfiltered 3DGS render, **Bottom:** Our filter: τ_{grad}= .00001, τ_{ratio}= .5. Additional examples of dynamic scene distractors being removed and comparisons to ground truth and BayesRays can be found in Appendix A.1.

parameters within NeRF models. However, BayesRays requires post-hoc training and remains exclusive to NeRF-based methods. Moreover, the prior distribution in BayesRays is modeled as an isotropic Gaussian, which regularizes the posterior toward uniform uncertainty across directions in

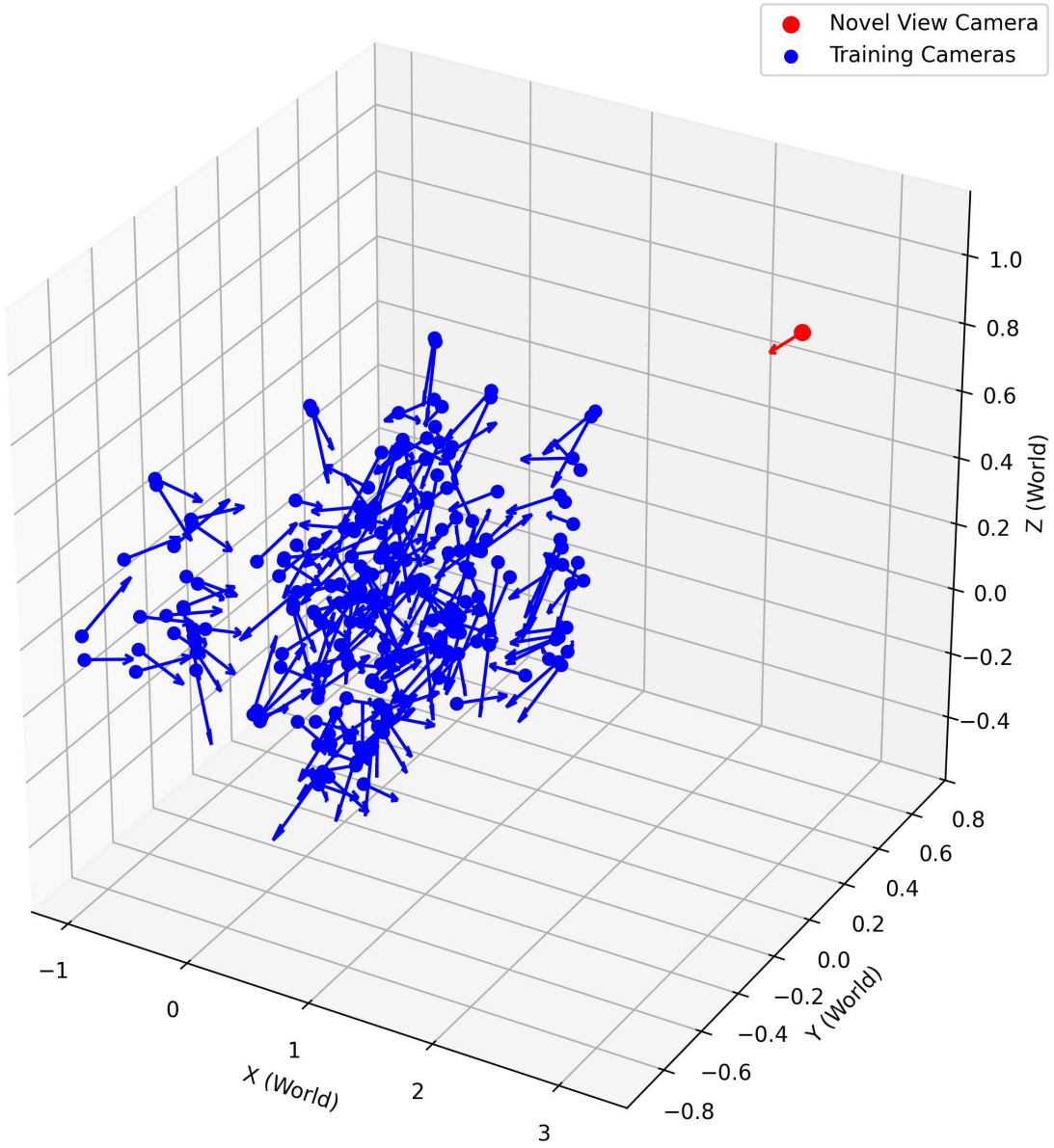


Figure 3.3: Novel view synthesis from an OOD camera pose(red) compared to the training cameras poses (blue) from the Playroom dataset shown in Figure 3.1.

parameter space. While the Laplace approximation is theoretically capable of capturing anisotropic local curvature through the Hessian of the loss function, practical implementations often rely on simplifications, such as diagonal approximations or low-rank estimates, that limit its effectiveness. As a result, directional instabilities in the rendered output are not captured during inference and must instead be approximated through post-hoc training procedures, making real-time uncertainty estimation infeasible.

3.3.2 Explicit Representations and 3D Gaussian Splatting

Recent efforts shifted toward explicit volumetric representations such as Plenoxels [11] and 3DGS, addressing computational inefficiencies and enabling real-time rendering. Specifically, 3DGS uses a set of explicit Gaussian primitives to represent radiance fields efficiently, making it highly suitable for interactive and personalized multimedia applications.

Jieng et al. [23] generalized uncertainty quantification beyond NeRF to explicit representations, calculating a Fisher information-based approximation across training images. Their method aims to maximize information gain for optimal view selection during active training but does not directly address rendering-time instabilities or anisotropic uncertainties.

Hanson et al. [17] explicitly integrated uncertainty attributes into 3DGS, representing each Gaussian primitive with an uncertainty measure utilized for pruning redundant primitives. While effective for reducing model complexity, their approach primarily targets spatial redundancy and does not specifically address directional instabilities during rendering.

3.3.3 Gaussian Rendering and Ray-Based Techniques

Methods exploiting ray-Gaussian intersections have emerged to improve rendering accuracy and facilitate geometry extraction. Leonides and Hebert [26] developed a differentiable ray-based renderer that integrates algebraic surfaces with Gaussian mixture models (GMMs). Their work provides analytic solutions for computing intersections between rays and Gaussian primitives.

Extending these concepts, Gao et al. [14] adapted ray-Gaussian intersection solutions for 3DGS for more accurate lighting effects and improved rendering quality in scenes represented with explicit

Gaussians. While Yu et al. [68] constructed volumetric opacity fields from ray-Gaussian intersections. The opacity fields enabled high-quality mesh extraction directly from 3DGS models. While this method significantly improves geometric fidelity and facilitate mesh extraction, it does not explicitly address uncertainty or sensitivity related to directional rendering instabilities.

3.3.4 Limitations and Gap in the Literature

Despite significant advances, existing uncertainty quantification methods primarily target isotropic noise, redundant primitives, or implicit representations. There is a notable gap in efficiently addressing anisotropic instabilities and directional sensitivity inherent in explicit representations such as 3DGS, particularly in generative multimedia contexts. Current approaches either incur substantial computational overhead, require retraining, or inadequately address directional instabilities that significantly degrade rendering quality and temporal coherence.

3.4 Background

3DGS is a hybrid volumetric rendering technique that models radiance fields using a discrete set of explicit Gaussian volumes. By representing radiance fields as a set of 3D Gaussians, 3DGS defines the probability distribution of the radiance of a point in space. Each 3D Gaussian primitive, $\mathcal{G}$, is independently parameterized by its density parameters and spatial parameters. The density parameters are color, c, and opacity, α. The spatial parameters are the mean, μ, and covariance, Σ.

$$\mathcal{G}(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{3}$$

To ensure the covariance matrix remains positive semi-definite throughout optimization the 3D Gaussians are represented as ellipsoids. Therefore the Gaussian mean is represented by the ellipsoid's center while the covariance matrix is decomposed into a scaling matrix, S, and rotation matrix, R,

$$\mathcal{G}(x) = e^{-\frac{1}{2}(x-\mu)^T(RSS^TR^T)^{-1}(x-\mu)} \tag{4}$$

For rendering, 3D Gaussians are depth-sorted and projected onto a 2D image plane. To minimize the perspective distortion of the projection of the covariance matrix, the Jacobian, J, of the affine approximation of the projective transformation is used [71]. Therefore, given a viewing transformation W, the 3D covariance matrix Σ is transformed into camera coordinates, Σ' by:

$$\Sigma' = JW\Sigma W^T J^T \tag{5}$$

The final 2D covariance matrix is obtained by removing the third row and column of Σ', resulting in a 2×2 matrix describing the Gaussian projected into screen space.

The color, C, of pixel, x, is computed by alpha compositing:

$$C(x) = \sum_{i=1}^{N} c_i \alpha_i \mathcal{G}_i(x_i) \left[\prod_{j=1}^{i-1} \left(1 - \alpha_j \mathcal{G}_j(x_j)\right) \right] \tag{6}$$

The parameters of each 3D Gaussian are optimized from back propagating the loss between the rendered image and ground-truth image.

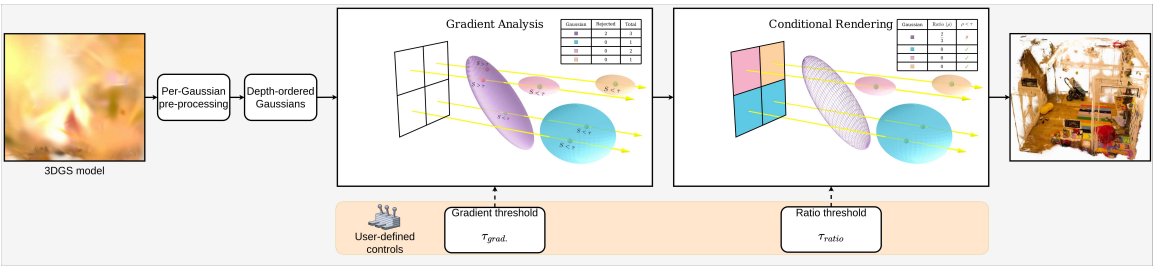


Figure 3.4: 3DGS render pipeline with two-pass filtering

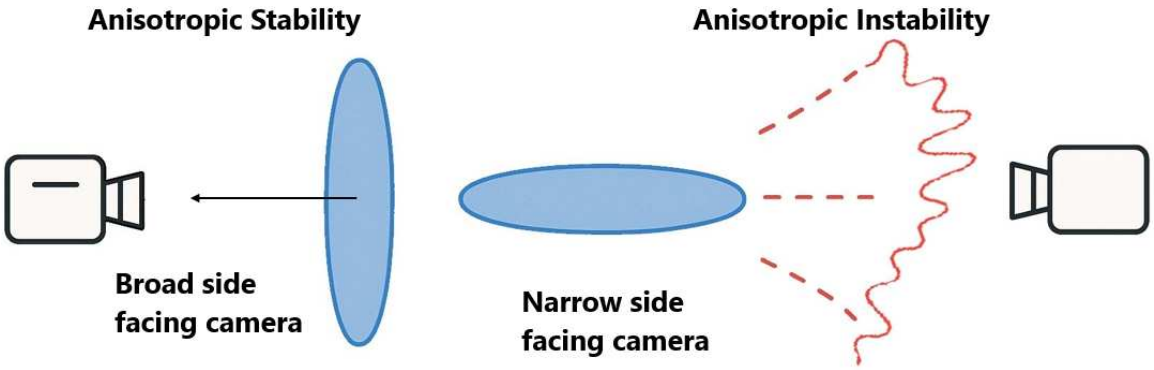


Figure 3.5: Anisotropic stability vs. instability. Red lines illustrate high gradient ray intersections, leading to view-dependent visual artifacts.

3.5 Methodology

We introduce a gradient sensitivity analysis tailored specifically for 3DGS. The primary goal is to systematically assess and improve the stability and reliability of rendered colors by addressing the challenges posed by spatial perturbations. Gaussian sensitivity analysis quantifies the degree to which variations in the input parameters of each Gaussian influence the rendered output. Specifically, this analysis emphasizes the sensitivity of pixel colors to minute perturbations in the spatial configurations of Gaussians intersected by camera rays. This provides deeper insight into regions that might be susceptible to visual artifacts or inconsistencies.

Formally, sensitivity is expressed through the magnitude of the gradient of a pixel's color relative to changes in parameters at the intersection between a ray and a Gaussian. This intersection point is defined as the depth along a camera ray where the Gaussian achieves its maximum contribution to the rendered output. Importantly, the calculation of this point considers all essential spatial parameters of the Gaussian (i.e. position, orientation, and scale). A low gradient magnitude signifies that the rendered color is minimally affected by small parameter changes, thus indicating a stable and reliable region in the rendering space. Conversely, a high gradient magnitude highlights substantial sensitivities, signaling areas that are unstable and potentially prone to visual anomalies.

A contribution of a Gaussian to the pixel color depends on the intersection point with the pixel

ray. When this intersection occurs near the Gaussian's mean, the sensitivity to parameter perturbations is low, as the gradient of the Gaussian is minimal near its center. Consequently, such contributions are rarely rejected by a per-contribution sensitivity threshold. To assess the overall stability of a Gaussian, we introduce a two-pass approach: the first pass records how often each Gaussian's contributions are rejected across all pixels (τ_{grad}), and the second pass computes the proportion of rejections. If this proportion exceeds a user-defined threshold (τ_{ratio}), the Gaussian is eliminated from the final rendering for that viewpoint.

Conducting systematic analyses of these gradient sensitivities allows for targeted and informed filtering of problematic Gaussians per viewpoint. This selective filtering significantly enhances visual quality, reduces rendering artifacts, and ensures greater temporal consistency across frames, ultimately leading to more robust and immersive visual experiences.

3.5.1 Ray-Marching

To analyze the radiance field with fine spatial resolution, we adopt a ray-marching approach. Unlike screen-space projection methods that evaluate Gaussians in 2D after rasterization, ray-marching enables direct computation of the ray-Gaussian interaction in 3D space, leading to more precise control over rendering dynamics. This method is particularly suited for computing pointwise sensitivity, as it allows us to localize the analysis to specific ray-Gaussian intersections.

Given a ray defined by its origin $o \in R^3$ and direction $r \in R^3$, any point x on the ray is parameterized as:

$$x = o + tr \tag{7}$$

where t represents the distance along the ray. To evaluate the contribution of a k^{th} 3D Gaussian to the ray, we first transform the ray into the Gaussian's canonical coordinate system. This transformation normalizes spatial variation using the Gaussian's scale S_k and orientation R_k, and positions the ray relative to the Gaussian's mean μ_k:

$$o_g = S_k^{-1} R_k (o - \mu_k) \tag{8}$$

$$r_g = S_k^{-1} R_k r \tag{9}$$

$$x_g = o_g + t r_g \tag{10}$$

In this normalized space, the Gaussian contribution simplifies to a 1D univariate Gaussian along the ray:

$$G^{1D}(t) = e^{-\frac{1}{2} x_g^T x_g} \tag{11}$$

$$= e^{-\frac{1}{2}(r_g^T r_g t^2 + 2 o_g^T r_g t + o_g^T o_g)} \tag{12}$$

This quadratic form of the exponent provides analytical tractability and numerical stability, facilitating efficient determination of the maximum Gaussian contribution without needing to solve higher-order equations.

The peak contribution occurs at the depth t_{min}, where the exponent reaches its minimum:

$$t_{\text{min}} = -\frac{o_g^T r_g}{r_g^T r_g} \tag{13}$$

After determining the depths for all Gaussians intersected by the ray, the final pixel color is computed via depth-ordered alpha compositing across all Gaussians intersected by the ray:

$$C(o, r) = \sum_{k=1}^{K} c_k \alpha_k \mathcal{G}_k^{1D}(t_{k,\text{min}}) \prod_{j=1}^{k-1} (1 - \alpha_j \mathcal{G}_j^{1D}(t_{j,\text{min}})) \tag{14}$$

This formulation allows precise control over how Gaussians influence each pixel, by ensuring accurate modeling of cumulative optical effects along viewing rays, laying the groundwork for analyzing how sensitive each pixel is to local changes in the 3D radiance field.

3.5.2 Color Gradient Sensitivity

To measure the sensitivity of the rendered color to spatial perturbations, we derive the gradient of the composite color $C(o, r)$ with respect to the 3D position x. This involves computing the gradient of the alpha-blended color contribution from each Gaussian, taking into account how both direct and accumulated transmittance change under spatial variation.

Starting from the definition, we express the gradient of the composite color as:

$$\nabla C(x) = \sum_{k=1}^{K} c_k a_k \nabla \prod_{j=1}^{k-1}(1 - a_j) + \prod_{j=1}^{k-1}(1 - a_j)\nabla \mathcal{G}_k(x_k) \tag{15}$$

where $a_i = \alpha_i \mathcal{G}_i(x_i)$. The gradient of a single Gaussian term is given by:

$$\nabla_x \mathcal{G}(x) = e^{-\frac{1}{2}x^T \Sigma^{-1} x}(-\tfrac{1}{2})\big(2\Sigma^{-1}x\big) = -\mathcal{G}(x)\Sigma^{-1}x \tag{16}$$

and the gradient of the accumulated transmittance product becomes:

$$\nabla_x \prod_j (1 - a_j) = \left(\prod_j (1 - a_j)\right) \sum_j \frac{a_j \Sigma_j^{-1} x_j}{1 - a_j} \tag{17}$$

Substituting these expressions into the original gradient equation yields:

$$\nabla C = \sum_{k=1}^{K} c_k a_k \prod_{j=1}^{k-1}(1 - a_j) \left(\sum_{j=1}^{k-1} \frac{a_j \Sigma_j^{-1} x_j}{1 - a_j} - \Sigma_k^{-1} x_k\right) \tag{18}$$

This Jacobian-like quantity describes how sensitive the final color is to the underlying spatial configuration. However, due to the involvement of matrix inversions and eigenvalue computations, evaluating this fully is computationally expensive and impractical for real-time rendering.

To improve computational practicality, we decouple the gradient from the color vector c_k by

replacing it with the scalar constant $c_k = 1$. This removes color-specific variation and focuses on transmittance dynamics. This form captures the structural sensitivity of the scene to spatial changes, identifying regions of high rendering instability while avoiding unnecessary per-color calculations.

Figure 3.6: Example of the loss of detail when including scale (left) vs. no scale (right)

3.5.3 Directional Sensitivity and Rotation-Based Gradient Filtering

While scalar gradient magnitudes effectively quantify overall spatial sensitivity, they do not capture directional instabilities. That is, situations in which the rendered output is disproportionately sensitive to perturbations along specific directions or axes. To address this limitation, we extend our analysis by calculating directional gradients within a rotation-aligned coordinate system. By isolating the influence of Gaussian orientation from its scale, this method enables a precise assessment of sensitivity relative to the Gaussian's principal axes. Consequently, we can independently evaluate directional instabilities without the confounding effects introduced by anisotropic scaling.

The core intuition is that Gaussians with strong anisotropic properties (i.e. those elongated along 1 or 2 of the 3 axes) exhibit direction-dependent instability. For example, a long, thin Gaussian may be stable along its major axis but highly sensitive to perturbations along its minor axes. Traditional geometric analyses such as Principal Component Analysis (PCA) use eigenvalue ratios of the covariance matrix to characterize such anisotropy, but these are not render-aware and they do not directly measure the impact on rendered output.

Instead, by applying the gradient filter in a rotation-only space, we can expose directional rendering instabilities. This is achieved by transforming the covariance matrix using only its rotation matrix R_k, excluding the scale S_k. In this aligned space, we compute sensitivity gradients relative

to each principal axis of the Gaussian, revealing how rendering behavior changes with orientation. Therefore, we drop the scale transformation in Equations (8)–(10) and derive our sensitivity metric:

$$S = \sum_{k=1}^{K} a_k \prod_{j=1}^{k-1} (1 - a_j) \left(\sum_{j=1}^{k-1} \frac{a_j x_j}{1 - a_j} - x_k \right) \tag{19}$$

where $x = R_k(o - \mu_k) + tR_k r$, which represent the ray-Gaussian intersection point in the rotation-aligned Gaussian space.

This approach offers several advantages:

- **Isolates rotational sensitivity**: By excluding scale, we ensure that the gradient reflects only changes due to orientation, not magnitude.
- **Highlights unstable orientations**: High directional sensitivity indicates that minor misalignments can significantly affect rendering, flagging Gaussians prone to producing visual artifacts. (See Figure 3.5)
- **Complementary to PCA filtering**: While PCA identifies noise due to isotropic variations, our method detects and corrects instabilities due to anisotropic orientation.

3.5.4 Aggregate Sensitivity Analysis

We introduce a two-pass filtering pipeline to evaluate and reject unstable Gaussians based on their aggregate sensitivity scores (see Algorithm 1 and Figure 3.4).

In the first pass, we compute gradient-based sensitivity at each ray-Gaussian intersection, as defined in Equation 19. A Gaussian's contribution to a pixel's color is conditionally accepted or rejected according to a user-defined sensitivity threshold (τ_{grad}). For each Gaussian, a ray intersections contribution is either accepted or rejected. We track 2 counts: rejected and used, where used is the sum of accepted and rejected counts. These statistics serve as inputs for the second pass.

In the second pass, we compute the aggregate sensitivity score for each Gaussian as the ratio of its rejection count to its total usage count. The aggregate sensitivity score for a Gaussian with a

zero usage count is 1. Gaussians with a rejection ratio exceeding a user-defined threshold (τ_{ratio}) are excluded from the final rendering.

Ultimately, this filtering mechanism enables robust scene reconstruction by attenuating the influence of unstable Gaussians. By selectively removing Gaussians with high aggregate sensitivity, we reduce noise and enhance the spatial consistency of the rendered view. This targeted use of rotation-aligned gradient analysis allows us to remove distractors and improve rendering quality.

3.6 Results

To evaluate the impact of our modifications, we conducted a series of experiments using a modified version of the Nerfstudio Splatfacto framework [60]. Specifically, we replaced the default projection mechanism with ray marching and incorporated a gradient sensitivity computation and filtering mechanism to improve perceptual rendering quality in OOD views.

To mitigate the effects of high-gradient sensitivity during ray-Gaussian intersections, we applied a thresholding mechanism. Threshold values were selected heuristically, with all experiments utilizing a fixed gradient threshold of 0.0001 and a ratio threshold of .5 (unless stated otherwise). These unitless thresholds were used to exclude regions exhibiting extreme sensitivity, which could lead to instability in uncertainty estimation.

We applied our method to evaluate rendering quality across two datasets, Deep Blending and NeRF On-the-go, which feature complex, real-world scenes with diverse environment and lighting conditions. For Deep Blending, we generated extreme out-of-distribution (OOD) views by extrapolating beyond the training camera poses. This allowed us to test our filtering mechanism's ability to eliminate noisy or unstable Gaussians during rendering. In contrast, NeRF-On-the-go was evaluated using renders from the original training camera trajectories. The rendered outputs were then compared to their corresponding ground truth images to assess how effectively our filtering mechanism suppresses transient artifacts introduced by dynamic scene elements. Since the goal of our filtering mechanism is to remove noisy rendering artifacts from novel viewpoints without an available ground truth, we assess perceptual quality using standard no-reference image quality (NR-IQA) metrics: NIQE, BRISQUE, and PIQE.

- **Natural Image Quality Evaluator (NIQE)**:
  assesses image quality based on deviations from a learned natural scene statistics model. [40]
- **Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)**:
  uses statistical features from locally normalized luminance coefficients to predict perceived distortion. [39]
- **Perception based Image Quality Evaluator (PIQE)**:
  quantifies perceptual distortions by analyzing block-wise degradation in an image, emphasizing spatially significant distortions. [45]

These metrics are designed to capture human-perceived image quality by penalizing unnatural textures, noise, and distortions in the absence of reference images. For all three, lower scores indicate better perceptual quality.

(a) NR-IQA scores for Playroom. $\tau_{grad.} = 10^{-5}$, $\tau_{ratio} = 0.5$

	NIQE ↓	BRISQUE ↓	PIQE ↓
BayesRays 0.1	11.74	44.45	83.84
BayesRays 0.2	9.18	47.11	74.47
BayesRays 0.5	10.67	64.24	58.31
Ours	**3.41**	**41.38**	**52.99**

(b) NR-IQA scores for CreepyAttic. $\tau_{grad.} = 10^{-5}$, $\tau_{ratio} = 0.01$

	NIQE ↓	BRISQUE ↓	PIQE ↓
BayesRays 0.1	10.30	45.16	77.42
BayesRays 0.2	10.17	52.12	69.64
BayesRays 0.5	14.02	53.57	46.87
Ours	**5.94**	**42.30**	**40.24**

(c) NR-IQA scores for DrJohnson. $\tau_{grad.} = 10^{-5}$, $\tau_{ratio} = 0.5$

	NIQE ↓	BRISQUE ↓	PIQE ↓
BayesRays 0.1	10.02	46.84	70.38
BayesRays 0.2	9.71	57.88	59.58
BayesRays 0.5	11.34	55.91	59.76
Ours	**4.49**	**37.37**	**51.60**

Table 3.1: NR-IQA scores for different scenes.

In both NeRF and 3D Gaussian Splatting (3DGS) settings, we render ”white space” for camera viewpoints far outside the training distribution. This occurs because these models only learn

to reconstruct content within the well-observed bounds of the training images. Extrapolated views lack sufficient supervision, resulting in default white outputs due to the model's inability to infer meaningful scene content. However, large white regions introduce challenges for NR-IQA metrics, which rely on statistical texture and spatial detail. Uniform areas lacking gradient or edge information skew these metrics and may be incorrectly penalized as over-smoothed or distorted, especially in patch-based approaches such as PIQE. To address this, we cropped all rendered images before evaluation. Evaluation images were captured by rendering animations that begin within the training

Figure 3.7: Single-Pass render of Figure 3.1. $\tau_{grad.}$ = .00001

camera distribution and gradually transition to an extreme OOD orbital trajectory. Each animation

followed a smooth camera path traversing from the center of the learned scene volume to viewpoints significantly outside the bounds of the training data. For every frame, we computed NIQE, BRISQUE, and PIQE scores to assess the evolution of perceptual quality as the model moves from familiar to extrapolated regions. Average scores across all frames are reported in Table 3.1a to 3.1c. These tables also present results for different scenes from the Deep Blending dataset [19], which was used to train both 3DGS and NeRF models. The selected frames and their corresponding unfiltered counterparts are included in the appendix for visual comparison. As this represents, to our knowledge, the first real-time filtering method for 3DGS, we compare our results to BayesRays [15], a real-time filtering approach for NeRFs that leverages a pre-trained uncertainty field.

Furthermore, experiments with scenes from the NeRF On-the-go dataset [52] suggest that our intermediate gradient metric, calculated explicitly within a rotation-only space, can effectively serve as an alternative approximation of the Jacobian-based Hessian of the Fisher Information Matrix (FIM). Standard practice approximates the Hessian using an expected value of Jacobian products across multiple samples. However, we employ a decoupled, single-sample gradient evaluation which efficiently captures the essential directional sensitivity encoded in the Jacobian's partial derivatives. This provides a computationally efficient alternative for epistemic uncertainty. High gradient magnitudes highlight significant directional sensitivities, indicating regions of heightened epistemic uncertainty due to limited or ambiguous training data. By identifying and conditionally filtering these Gaussians, our method effectively suppresses Gaussians exhibiting substantial anisotropic instability, significantly reducing visual artifacts related to epistemic uncertainty, see Figure 3.2.

All experiments were conducted on an NVIDIA RTX 2080 Ti GPU. For consistency, Splatfacto models were trained with COLMAP initialization for 30,000 iterations. Following BayesRays, we also trained a Nerfacto model for 30,000 iterations and extracted uncertainty fields over an additional 1,000 iterations. Images were resized by 30% to 50% such that they were approximately 1500x700 px due to hardware and time constraints.

3.7 Ablation

We analyze the contribution of individual filtering components through a series of ablation studies.

3.7.1 Single-Pass

We first implement a single-pass approach in which Gaussians are not fully removed; instead, only their contributions are selectively filtered based on the intermediate gradient at the ray-Gaussian intersection point. As shown in Figure 3.7, the centers of many Gaussians remain visible in the render. This occurs because the gradient magnitude is low near a Gaussian's center. From Equation 19, this behavior is expected: the x_k term is near zero when the ray intersects the Gaussian near its mean. For the first Gaussian, there are no prior contributions along the ray, therefore the cumulative sensitivity remains at its initialized value of zero. As a result, initial Gaussians in depth ordered list (and often responsible for occluding distant geometry in extreme viewpoints) are preserved when intersected near their centers because their sensitivities are near zero.

3.7.2 Scale-Incorporated

Next, we examine the impact of incorporating scale into the intermediate gradient calculation, as described in Section 3.5.3. This leads to a loss of anisotropic shape information, effectively normalizing Gaussians into unit spheres in their local coordinate space. Figure 3.6 illustrates how this transformation causes elongated, noisy Gaussians to persist, while small Gaussians with scales less than 1 (which are critical for scene detail) are over-filtered. This is due to the normalization process that increases the magnitude of the gradient vector which makes them more likely to exceed the rejection threshold.

3.7.3 Filter Parameters

Finally, the impact of the two parameters (gradient $\tau_{grad.}$, and ratio threshold τ_{ratio}) in isolation are demonstrated in Figures 3.8 and 3.9. Through visual inspection it is clear that exclusively using one or the other parameter results in differing visual quality with the best performance derived from

a scene-specific combination of the two set by the user.

Figure 3.8: Same frame as Figure 3.1 while holding the ratio threshold τ_{ratio} at .5. **Left**: $\tau_{grad.}$ = .0001, **Right**: $\tau_{grad.}$ = .0005

Figure 3.9: Same frame as Figure 3.1 while holding the gradient threshold $\tau_{grad.}$ at .00001. **Left**: τ_{ratio} = .25, **Right**: τ_{ratio} = .75

3.8 Conclusion

We introduced a novel sensitivity measurement for 3DGS that identifies and filters anisotropic instabilities during rendering, without requiring retraining or scene-specific tuning. By analyzing

intermediate gradient responses from the differentiable rasterization pipeline, our method targets the core source of generative uncertainty: Directional instability arising from anisotropic orientations. This filtering mechanism enables robust rendering even when users navigate freely beyond the original training views, a setting where standard 3DGS models often produce severe visual artifacts. Experimental results across complex, photorealistic datasets demonstrate consistent improvements in perceptual quality metrics surpassing baseline 3DGS and NeRF-based methods such as BayesRays. By integrating seamlessly into existing 3DGS pipelines, our method empowers generative multimedia systems to dynamically produce realistic, stable, and personalized content. This substantially enhances user interactivity and immersion in real-time generative 3D environments.

Acknowledgement

This research was undertaken, in part, based on support from the Natural Sciences and Engineering Research Council of Canada Grants RGPIN-2021-03479 (NSERC DG) and ALLRP 571887 - 2021 (NSERC Alliance).

Chapter 4

Additional Metrics

We introduce and compare two complementary evaluation metrics for object-centric image analysis: CLIP Alignment Score (CAS) and SAM Reference Similarity Score (SRSS). Both metrics leverage a contrastive CLIP filter to select relevant regions, but differ in their final scoring stage. One uses Segment Anything Model (SAM) [28], embeddings against a reference set of exemplar regions, while the other relies solely on Contrastive Language-Image Pre-Training (CLIP) [49] image–text similarity without any external references. We describe each metric in turn, analyze their strengths and failure modes, and present guidance on when to apply each in practice.

4.1 Background and Motivation

In many real-world scenarios there is no canonical reference image available to verify whether a distractor has been successfully removed. In such cases, evaluation must rely on semantic understanding rather than pixel-level comparison. This shift presents two central challenges:

- **Semantic Context Dependence**: The plausibility of an edit often depends on the broader spatial and structural coherence of the scene. For instance, removing a foreground statue may reveal content that only appears realistic if it aligns with contextual structures, such as the continuation of an arch or consistent lighting. However, classical no-reference image quality assessment (NR-IQA) metrics like BRISQUE [38] and NIQE [41] are rooted in low-level natural scene statistics (NSS) and are not equipped to reason about global semantic layout or

object relationships.

- **Visual vs. Semantic Plausibility**: An image may appear artifact-free and perceptually natural, yet remain semantically implausible. For example, a generative model may inpaint a region with structurally coherent textures but introduce impossible object placements or illogical scene arrangements. Such failures elude NSS-based methods and even many deep NR-IQA models, which can be misled by irrelevant semantic features [32]. Recent studies demonstrate that deep features may encode "semantic noise" that confounds quality prediction when content varies but distortion remains constant [32].

To overcome these limitations, modern vision-language frameworks are increasingly being used to evaluate image quality with an emphasis on semantic plausibility. These include a constellation of tools that jointly provide object-level reasoning, fine-grained localization, and language-grounded verification. For instance, Distillation with No Labels (DINO) [6] and its open-vocabulary variant Grounding DINO [34] enable zero-shot object proposals from natural language prompts by leveraging transformer-based attention and alignment with CLIP embeddings. Meanwhile, the Segment Anything Model (SAM) [28] provides high-quality, promptable segmentation masks that generalize across object categories, enabling spatial precision in evaluating whether editing artifacts persist.

CLIP-based metrics such as CLIPScore [21] and CLIP-IQA [64] offer strong tools for measuring the semantic alignment between an image and a textual prompt, particularly in tasks like captioning or prompt-conditioned image generation. Their ability to function without a ground-truth reference makes them appealing for evaluating generated content in open-domain settings. However, these models are fundamentally designed to detect the presence of target concepts, not the absence of undesired content.

This makes them poorly suited for evaluating distractor removal or semantic editing. CLIPScore, for instance, will remain high as long as the image retains elements of the prompt, even if the prompt-relevant object is partially occluded by clutter. Similarly, CLIP-IQA does not explicitly penalize artifacts or distractors unless they severely disrupt global semantics. Neither metric is contrastive or selective enough to verify whether edits successfully eliminate unwanted elements. As a result, both are limited in applications where semantic precision rather than broad alignment, is the

primary evaluation goal.

While global semantic metrics such as CLIPScore and CLIP-IQA focus on text-image alignment, they are often insensitive to localized changes and cannot directly verify whether an unwanted object has been successfully removed. To address this gap, ReMOVE (Reference-Free Object Removal Verification) [7] introduces a region-focused alternative that evaluates edits based on visual coherence and residual objectness within the masked area. Rather than relying on external prompts or reference views, ReMOVE analyzes the edited region in relation to the unaltered background, using self-supervised features to detect whether the fill is both plausible and free of unintended replacements.

This approach makes ReMOVE particularly well-suited for object removal tasks where the goal is to eliminate all semantic and visual traces of a distractor. It can detect subtle failures, such as partial removal, replacement with another object, or unnatural boundaries, that may go unnoticed by prompt-driven metrics. However, because ReMOVE operates solely on visual cues within the image, it cannot assess whether the removal preserves broader scene consistency or aligns with the semantic structure of the surrounding context.

To address these limitations, visual understanding components can be integrated to form a more semantically aware evaluation pipeline. Grounding DINO enables zero-shot object proposals from natural language prompts, SAM produces high-quality region masks, and CLIP embeds both images and text into a shared semantic space. When combined, these tools enable the construction of metrics that (a) filter distractors through contrastive verification and (b) assess region fidelity either via a curated reference set or directly through textual similarity. This bridges the gap between perceptual and semantic evaluation.

4.2 Contrastive CLIP filter

The contrastive filtering process selects semantically relevant regions by integrating open-vocabulary proposals from Grounding DINO, fine-grained masks from SAM, and semantic similarity scoring via CLIP. It follows a three-stage pipeline: region proposal, semantic extraction, and contrastive

scoring. This pipeline is designed to maximize region coverage while minimizing the risk of overlooking valid objects, even those with low initial confidence. Together, these stages identify regions that most faithfully represent the intended object and discard ambiguous or misleading content.

Region Proposal (Grounding DINO)

Given an image and a text prompt, Grounding DINO is used to generate open-vocabulary bounding boxes. The model is invoked with a low threshold to return up to K region proposals, each associated with a raw confidence score. Rather than filtering out low-confidence boxes prematurely, all K hypotheses are retained at this stage to allow for more precise semantic filtering downstream.

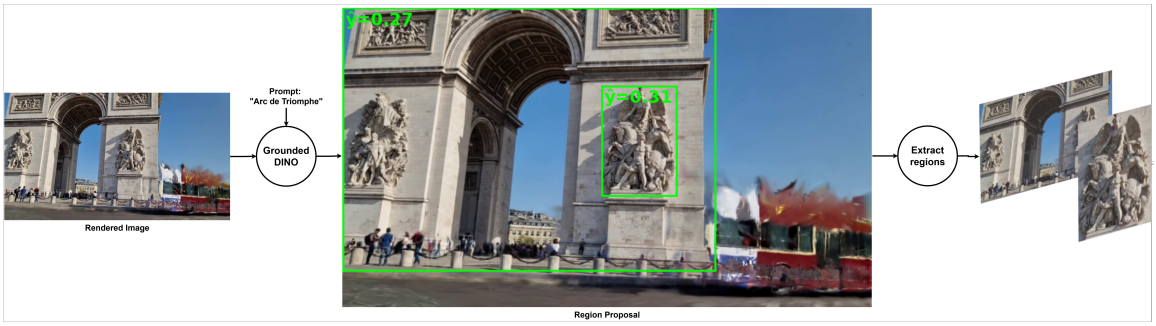


Figure 4.1: Region proposal and extraction pipeline

4.2.1 Semantic Extraction (SAM)

Each bounding box is then refined using SAM, which produces a high-resolution binary segmentation mask for the proposed region. To ensure quality and spatial coherence, any region with a degenerate box (width or height < 2 pixels) or a degenerate mask (e.g., containing fewer than two pixels in a row or column) is immediately discarded. This ensures that only well-formed candidate regions proceed to semantic verification.

4.2.2 Contrastive Scoring

Each candidate region is then segmented using SAM, and the resulting masked crop is encoded into a CLIP image embedding. These embeddings are evaluated based on their similarity to a positive class prompt and a negative distractor class prompt. Only regions for which the positive

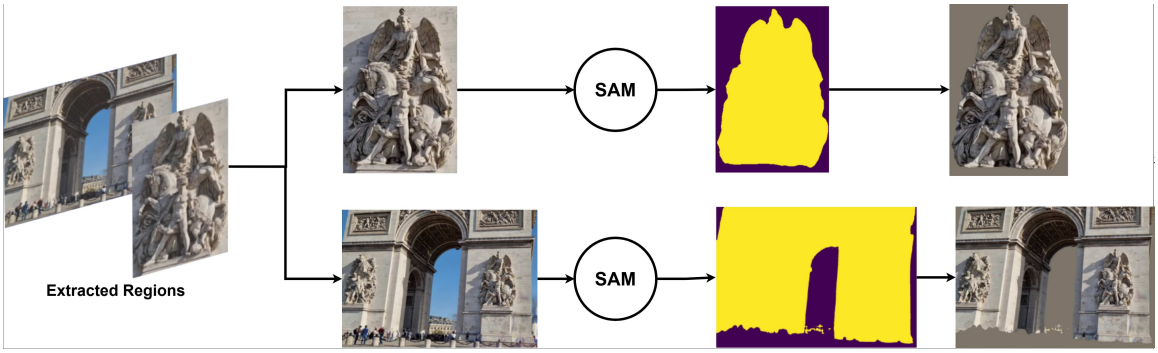


Figure 4.2: Semantic segmentation pipeline

similarity exceeds the negative similarity by a predefined margin are retained.

$$\text{sim}_{\text{pos}} - \text{sim}_{\text{neg}} \geq \text{margin}$$

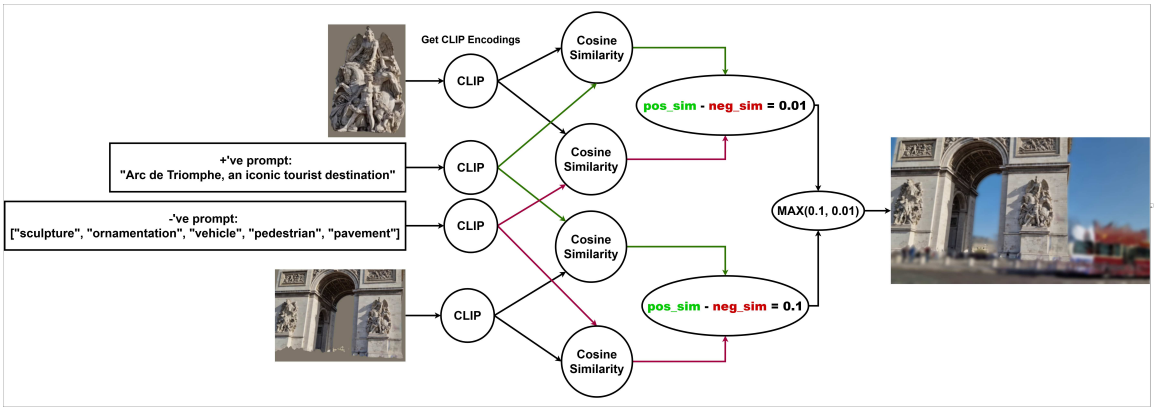


Figure 4.3: Contrastive scoring pipeline

To encode these textual concepts into CLIP embeddings, we employ two distinct strategies, each selected for its semantic properties and intended role in the filtering process:

- **Single Rich Prompt Encoding** (for positives): Positive concepts such as ”Arc de Triomphe” are often unique and highly structured, making them amenable to holistic descriptions. A compound phrase like ”Arc de Triomphe, a monument in Paris” is encoded as a single sentence, allowing CLIP to leverage syntax and relational cues (e.g., place, type, structure) for a coherent semantic representation. This approach produces a single, normalized text feature used to verify if a region depicts the intended subject.

- **Multiple Prompt Averaging** (for negatives): Distractor concepts like ”sculpture” or ”ornamentation” are more ambiguous and visually diverse. Instead of relying on a single sentence, we use a list of short prompts, each capturing a different facet of potential distractors. Each is encoded individually and L2-normalized before averaging, yielding a composite embedding that reflects the full spectrum of undesired content. This makes the negative signal more robust across varied backgrounds or occlusions.

In short, this filtering relies on distinct text encoding strategies: single rich prompts are used for positives to maximize specificity and capture structured semantic meaning, while prompt averaging is applied to negatives to ensure generality and robustness across a range of distractor types. A region is retained only if its alignment with the positive concept exceeds its alignment with the negative concept by a specified margin. This contrastive condition helps exclude ambiguous or misleading regions (such as close-up sculptures or occluded fragments). Only regions that clearly reflect the intended object are preserved. If no region passes the contrastive check, the function returns a value of –1, indicating a failure to detect the object under the current parameters.

Together, these strategies mitigate false positives caused by generic saliency or visual similarity and enable both CAS and SRSS to focus evaluation on semantically coherent candidates. Crucially, this process allows the metrics to operate without exhaustive supervision, making them applicable in open-world settings where object categories are not fixed and distractor types are highly variable.

4.3 SAM Reference Similarity Score (SRSS)

The SAM Reference Similarity Score (SRSS) is a semantic evaluation metric designed to assess whether a region in a test image corresponds to a known object, based on feature similarity rather than pixel-level matching. It operates by comparing region-level features extracted via SAM from the test image against a set of reference embeddings derived from curated images of the target object to establish a semantic anchor for the target object.

4.3.1 Reference Embedding Construction

Each reference image is first segmented using SAM, isolating the object of interest. SAM's internal image encoder produces a dense, high-resolution feature map of the image, and the masked region is then average pooled to yield a compact feature vector. These vectors, drawn from multiple reference views or conditions, are normalized and stacked to form a bank of reference embeddings that capture the semantic diversity of the object.

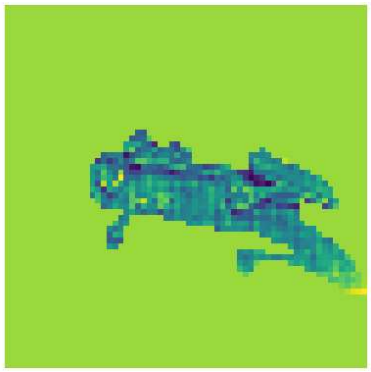


Figure 4.4: Heatmap of a reference embedding for the Spot scene from the Nerf-on-the-go dataset.

4.3.2 Test Image Scoring

When evaluating a test image, the same contrastive CLIP filtering pipeline is applied to identify semantically plausible candidate regions. For each retained region, SAM is again used to generate a segmentation mask, and a corresponding region embedding is extracted using the same average pooling method. The final SRSS score is computed by measuring the cosine similarity between the test region's embedding and each reference embedding, taking the maximum as the similarity score.

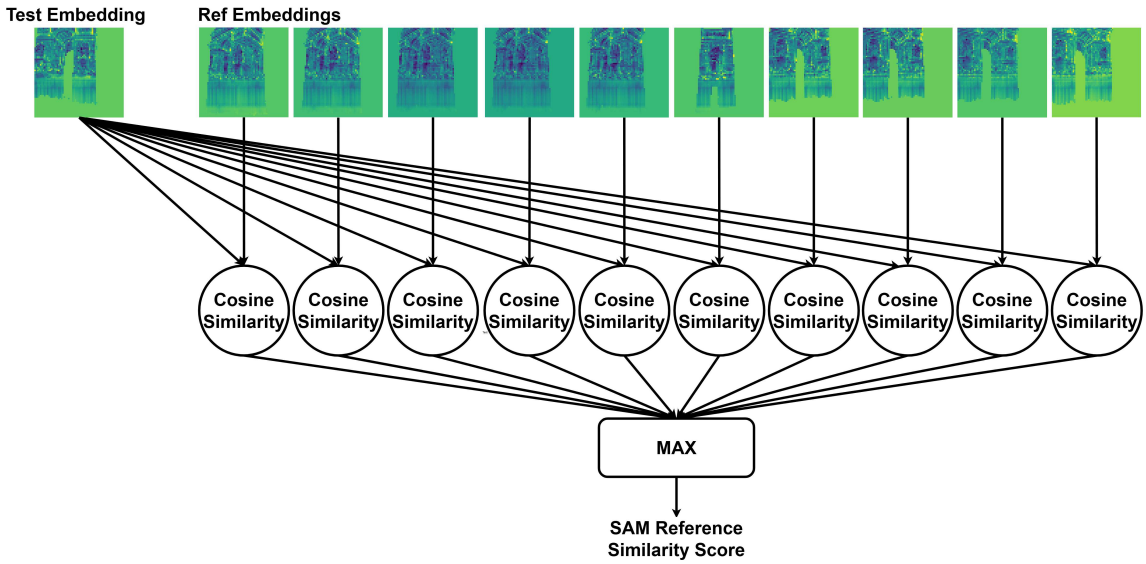


Figure 4.5: SRSS evaluation head

SRSS evaluates semantic correspondence under open-world conditions. By comparing each candidate region to a set of reference embeddings from varied viewpoints, it maintains robustness to changes in pose, lighting, and occlusion. Unlike traditional image quality assessment methods, it does not require pixel-level alignment between test and reference images, making it flexible and easy to deploy. Its reliance on high-level feature similarity allows it to capture whether the masked region genuinely resembles the target object, even in the presence of noise or visual ambiguity. Finally, because it builds entirely on pretrained SAM and CLIP models, SRSS is zero-shot and scalable, requiring no additional training to generalize across new concepts.

4.4 Clip Alignment Score (CAS)

From the contrastive shortlist, the region with the largest margin between positive and negative CLIP similarity scores ($\text{sim}_{\text{pos}} - \text{sim}_{\text{neg}}$) is selected as the most semantically aligned candidate. To isolate this region for final scoring, a composite image is generated by applying a background masking strategy to the original image. This composite image preserves the identified object while altering or neutralizing the surrounding pixels, ensuring that CLIP's embedding focuses primarily on the salient region of interest. Three masking strategies are supported:

- **Padding**: centers the cropped region on a neutral canvas.
- **Gaussian blurring**: suppresses contextual features by smoothing the background
- **Mean-filling**: replaces background pixels with an average RGB value

Each of these strategies reduces background interference and enforces attention to the object itself, which is critical because CLIP embeds entire images holistically.

This masked composite image is then re-encoded using CLIP, and its embedding is compared to the positive prompt embedding using cosine similarity. The resulting value defines the final CAS score, which quantifies how well the selected region semantically aligns with the intended object description. Importantly, this score is produced without any reference images, relying solely on language supervision. This makes CAS particularly suited to open-world or dynamic evaluation scenarios where reference images are not available, not reliable, or too costly to curate.

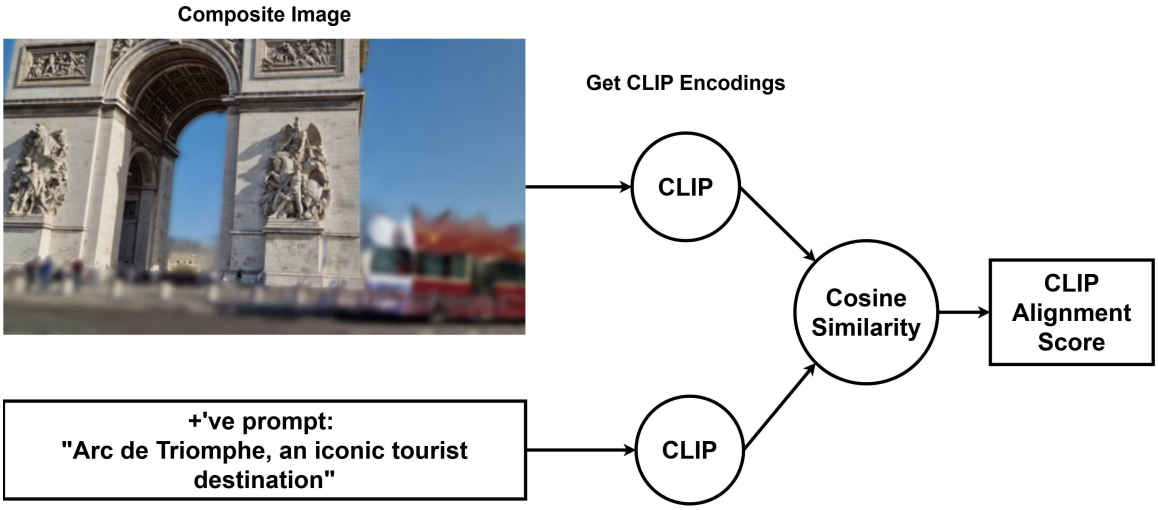


Figure 4.6: CAS evaluation head

Leveraging CLIP's ability to embed both images and text into a shared semantic space, the metric evaluates alignment based on conceptual similarity rather than pixel-level correspondence. This allows it to recognize whether a region plausibly depicts an object like the "Arc de Triomphe" across variations in lighting, perspective, or occlusion. By re-encoding a composite image where the background is masked, CAS evaluates the masked region in a controlled semantic context. This partial suppression of the background helps isolate the object while retaining enough structural layout to preserve spatial cues, allowing CLIP to focus on the intended region without being misled by surrounding distractors.

Moreover, the use of contrastive filtering upstream ensures that only semantically relevant candidates are ever evaluated, reducing false positives from visually salient but semantically incorrect regions. The final score thus reflects not only the presence of a plausible object but also its distinctiveness from distractors, as defined by the positive–negative prompt contrast.

Taken together, CAS offers a lightweight, zero-shot, and scalable alternative to reference-based quality metrics. It is especially valuable for evaluating image edits, generation, or object removal tasks in unconstrained environments, where visual plausibility and semantic coherence must be judged relative to intent rather than reference images.

4.5 Qualitative Analysis of Scored Regions

Across the four illustrative case studies that follow, each experiment is conducted using the same set of scene renders produced by the reconstruction models developed in Chapter 3. To ensure consistency in zero-shot, text-driven evaluation, we apply the blur masking mode when computing the CAS, allowing the masked object to remain in focus while background features are suppressed. For SRSS, the evaluation relies on a set of 10 curated reference images from the scene, each selected to clearly depict the main (canonical) object at the focal point of the scene from a diverse range of viewing angles. The curated reference images act purely as a semantic anchor for SRSS, they are never seen during model training. All positive and negative prompts used for both SRSS and CAS, ranging from minimal identifiers to rich descriptive sentences, are documented in Appendix C. The numerical results discussed in each case study are drawn from Table 4.1 and Table 4.2, which summarizes the average scores for SRSS and CAS across scenes and rendering configurations. Together, these case studies offer qualitative insight into how each metric responds to different visual configurations, as well as how our rotation-aligned gradient filter affects semantic fidelity and region selection. By examining not just the score values but also the spatial regions and semantic features deemed important, we gain a deeper understanding of the comparative behavior of SRSS and CAS, and how our proposed filtering strategy influences their performance in preserving meaningful content while attenuating noise.

	Splatfacto		**Nerfacto (Bayes-Rays)**			
Scene	**No filter**	**Our Filter**	**No filter**	**0.25**	**0.50**	**0.75**
Arc	0.985	0.982	0.908	0.905	0.930	0.930
Patio	0.742	0.804	0.416	0.437	0.411	0.403
Spot	0.955	0.945	0.677	0.599	0.622	0.657

Table 4.1: Average SRSS by scene and method. Our filter: τ_{grad}= .00001, τ_{ratio}= .5.

4.5.1 Case Study 1: High Agreement on a Canonical Object

In the Arc de Triomphe scene, where the target object is clearly visible and centrally positioned, both SRSS and CAS yield high similarity scores, demonstrating strong agreement in their

	Splatfacto		Nerfacto (Bayes-Rays)			
Scene	**No filter**	**Our Filter**	**No filter**	**0.25**	**0.50**	**0.75**
Arc	0.3158	0.3094	0.2685	0.2293	0.2463	0.2293
Patio	0.0989	0.2120	-0.1578	-0.1082	-0.1670	-0.1619
Spot	0.2697	0.2655	0.0857	-0.0310	0.0006	-0.1619

Table 4.2: Average CAS by scene and method. Our filter: τ_{grad}= .00001, τ_{ratio}= .5.

semantic evaluations. SRSS, which compares the test region's SAM-derived embedding to a curated bank of reference features from ten distinct viewpoints, registers a near-perfect average score of 0.985 under standard Splatfacto rendering. CAS, which computes cosine similarity between a CLIP-encoded masked region and a descriptive text prompt such as "Arc de Triomphe," similarly confirms strong alignment, validating the region's semantic coherence through language supervision. The contrastive CLIP filter in both cases effectively isolates the correct object by excluding visually salient but semantically irrelevant content, ensuring that evaluation is focused on meaningful structure (see Figure 4.7).

Figure 4.7: Case Study 1: frame_200.jpg, arcdetriomphe
Left (Unfiltered Splatfacto): CAS = 0.3188, SRSS = 0.9935 The full Arc de Triomphe is presented against a blurred crowd and street; both metrics register near-ceiling similarity, reflecting clear, unambiguous depiction.
Right (Our Filter): CAS = 0.3193, SRSS = 0.9940 After applying our filter to suppress unstable, anisotropic artifacts, the core structure remains intact and both scores hold steady. This demonstrates that the object's semantic embedding is unaffected by background suppression.

When our rotation-aligned gradient filter is applied, SRSS remains essentially unchanged, holding at an average of 0.982 across the dataset, indicating that the filter preserves the object's semantic integrity while selectively suppressing unstable, anisotropic artifacts. This minimal drop shows that we can remove noisy, low-confidence pixels without disrupting the core identity of the target. At

the same time, the filter consistently boosts CAS in even the most challenging cases: for example, on frame_640.jpg in Appendix A.1, CAS rises from 0.282 to 0.297 under Splatfacto, compared with 0.316 for Nerfacto and 0.296/0.294 for Bayes-Rays thresholds 0.25/0.50 (see Appendix A.1).

4.5.2 Case Study 2: Divergence Under Visual Ambiguity and Prompt Instability

The Patio scene reveals a nuanced divergence between prompt-based and structure-based evaluation metrics, offering a compelling illustration of how visual ambiguity challenges semantic alignment. Under standard Splatfacto rendering, SRSS registers an average of 0.742, suggesting moderate alignment with the reference set of canonical views, while the CAS average remains much lower at 0.099, reflecting difficulty in satisfying the prompt under occlusion, clutter, or fragmentary visibility. This discrepancy is particularly evident in scenes where the gondola is partially obscured or crowded by distractors.

This divergence is diagnostic. CAS relies on global text-to-image alignment and penalizes regions that only partially reflect the prompt. In cluttered scenes like Patio, where the gondola may be partially occluded, distorted, or surrounded by people, CAS remains conservative, even when structural cues are strong. SRSS, by contrast, compares localized SAM-derived features to curated embeddings from ten canonical views. It tolerates variability in pose and occlusion, resulting in more stable semantic alignment (see Figure 4.8.

Notably, CAS and SRSS respond differently because they evaluate different aspects of the same region: CAS emphasizes how fully the prompt is satisfied, while SRSS tests resemblance to a learned visual prototype. The Patio results illustrate both the failure modes of prompt-based evaluation and the robustness of instance-level reference matching (see Figure 4.9). Our filter mitigates both by preserving semantically meaningful structure while suppressing anisotropic artifacts, leading to converging improvements across metrics without compromising either.

When our rotation-aligned gradient filter is applied, both metrics improve substantially: SRSS rises to an average of 0.804 and the CAS average more than doubles to 0.212. These gains highlight the filter's dual role: it preserves salient object structure for SRSS, while removing distracting regions that weaken CAS's prompt alignment. Crucially, this improvement does not come at the cost of discarding meaningful structure. Rather, the filter enhances signal fidelity by removing

Figure 4.8: Case Study 2.1: IMG_8668.JPG, patio_high
Left (Unfiltered Splatfacto): CAS = 0.200, SRSS = .893 A small graffiti panel of the cabin peeks through a cluttered foreground. SRSS remains high because the fragment matches the reference views, but CAS stays low because prompt alignment is diluted by distractors.
Right (Our Filter): CAS = 0.244, SRSS = .955 After removing clutter, the full object is revealed. Both scores rise, yet CAS still lags SRSS, underscoring that text-driven alignment falters under visual ambiguity even when reference-based matching succeeds.

content that is visually unstable without sacrificing semantic cues.

This sharply contrasts with Bayes-Rays filtering in Nerfacto, which struggles to balance this tradeoff. While SRSS scores for Nerfacto remain low across all thresholds (ranging from an average of 0.403 to 0.437), CAS scores degrade even further into the negative (from an average of –0.108 down to –0.178), revealing that its filtering method often removes or misidentifies object-level content entirely.

In contrast, our filter actively improves CAS, a prompt-sensitive metric, without compromising SRSS, which is grounded in local visual embeddings. This shows that, unlike Nerfacto, our method can disambiguate scenes where the prompt is underspecified or overly broad. The Patio scene thus serves as a compelling case of metric divergence under visual ambiguity. It highlights not only the limitations of prompt-based evaluation, but also the importance of instance-grounded structural matching. In this context, our rotation-aligned gradient filter bridges the gap between these two evaluation paradigms, delivering reliable improvements where Bayes-Rays filtering consistently falls short.

Figure 4.9: Case Study 2.2: IMG_8545.JPG, patio_high
Left (Unfiltered Splatfacto): CAS = 0.244, SRSS = 0.926 A toy in the foreground partially obscures the gondola mini-cabin. SRSS remains high by matching the fragment to reference views, but CAS stays low because the text prompt cannot reliably pick out the occluded object among clutter.
Right (Our Filter): CAS = 0.244, SRSS = 0.942 Removing the occluding toy exposes the full cabin. SRSS improves while CAS stays relatively flat. This demonstrate the failure of prompt-based evaluation because the prompt-based alignment still struggles without more specific textual cues, even when the scene is fully visible.

4.5.3 Case Study 3: Edge Case with Occlusion or Partial View

In scenes where the target object is partially occluded or viewed from an uncommon angle, the metrics CAS and SRSS often diverge in their sensitivity and robustness. This is clearly observed in the ”Spot” scene, which features Boston Dynamics’ yellow quadruped robot. The SRSS scores for Splatfacto, both with and without our rotation-aligned gradient filter, remain consistently high with an average of 0.955 and 0.945 respectively. These values indicate that even when some object features are occluded or distorted, the SRSS evaluation can still successfully identify semantically correct regions. This robustness is due to the curated reference set used by SRSS, which includes 10 diverse views of the object and enables matching based on localized, instance-level features extracted via SAM.

In contrast, CAS, which relies on holistic text-to-image similarity, shows more variability. For Splatfacto, average CAS is 0.270 without filtering and 0.266 with our rotation-aligned gradient filter applied. These scores, while lower than the SRSS equivalents, still indicate moderate alignment with the positive prompt. However, for Nerfacto, CAS scores degrade considerably, especially at higher Bayes-Ray thresholds (e.g., average of 0.0005 at 0.5 and –0.162 at 0.75), reflecting CAS’s difficulty

in maintaining alignment when the full semantic concept (“robot dog lying on the floor”) is partially obscured or flattened by over smoothing. These low CAS scores suggest that the prompt fails to capture partial or ambiguous configurations of the robot, even if they would still be recognizable to a reference-based system like SRSS.

Importantly, our rotation-aligned gradient filter plays a valuable role here. It reduces the presence of visually salient but semantically ambiguous distractors, such as background clutter or anisotropically smeared textures, without erasing key structural features of the robot (see Figure 4.10). This helps preserve CAS performance in borderline cases while maintaining the high SRSS that depends on faithful region preservation. The slight CAS drop from an average of 0.270 to 0.266 is negligible and well within expected tolerances, reinforcing that the filter does not harm semantic alignment under occlusion. Together, these results highlight how SRSS can maintain accuracy under visual uncertainty, and how our filtering strategy balances aggressive noise suppression with semantic retention.

Figure 4.10: Case Study 3: IMG_8295.JPG, spot
Left (Unfiltered Splatfacto): CAS = 0.274, SRSS = 0.964 A slightly transparent red line cuts across the scene, partially obscuring the robot dog. SRSS stays high because the fragment still matches reference embeddings from varied angles, while CAS remains moderate under the distractor’s influence.
Right (Our Filter): CAS = 0.276, SRSS = 0.976 The red artifact is removed, leaving only the blurred background around the robot. SRSS edges up further and CAS improves slightly, illustrating that our filter can suppress partial occlusions without disrupting the core semantic identity.

4.6 Summary

The SAM Reference Similarity Score and the CLIP Alignment Score represent two complementary approaches to evaluating region-level object identity without relying on pixel-wise ground truth. SRSS operates by comparing a test region's SAM-derived feature embedding to a curated set of reference embeddings, enabling semantic verification against known visual instances. This approach offers fine-grained discrimination and robustness to viewpoint or occlusion, particularly when reference images capture a range of appearances. However, its reliance on pre-collected references can limit flexibility in open-ended settings, and performance is sensitive to the diversity and quality of the reference set. In contrast, CAS performs evaluation in a fully zero-shot setting by comparing the CLIP image embedding of a masked region directly to a text-encoded prompt. This makes it highly adaptable and lightweight, with no requirement for reference images. CAS is effective in open-world scenarios where semantic categories are fluid or user-defined, but it may struggle to distinguish between visually similar distractors and target objects, particularly when prompt specificity is lacking. While SRSS excels in scenarios requiring precise instance-level alignment with curated references, CAS is better suited for no-reference image quality assessment, where reference images are unavailable or impractical to obtain. Its reliance on text-driven semantic similarity allows CAS to flexibly evaluate object presence and alignment in open-ended or exploratory settings. Together, the two metrics strike a balance between precision and generality, supporting semantic evaluation across a wide spectrum of data regimes and deployment constraints.

Crucially, our filter introduced in this work consistently improves or preserves metric performance across all scenes and configurations. In SRSS, it enhances semantic correspondence by suppressing anisotropic or structurally unstable regions that often interfere with SAM's feature extraction, yielding more focused and reliable region embeddings. In CAS, it mitigates the risk of prompt misalignment by reducing background distractions that may dilute the text-image similarity computation. Importantly, these gains are achieved without compromising the core identity of the object; the filter retains sufficient structure and context to maintain fidelity under both evaluation strategies. For difficult frames involving occlusions, distracting textures, or semantic ambiguity, the filter helps isolate meaningful content while discarding misleading signals, effectively improving

the contrastive filtering stage and reinforcing both SRSS and CAS outputs.

In summary, our filtering strategy meaningfully improves semantic evaluation under challenging conditions, while SRSS and CAS offer orthogonal but complementary views of semantic correctness. Used to evaluate the effects of our filter, these metrics reveal how selective feature suppression can preserve object identity and enhance alignment with semantic intent, without relying on pixel-wise ground truth or retraining.

Chapter 5

Future Work and Conclusion

5.1 Future Work

There are several promising directions for extending this work, both in terms of enhancing the proposed gradient sensitivity filtering mechanism and in refining the evaluation metrics introduced alongside it. One avenue involves extending our current single-view sensitivity model into a multi-view framework. While the present pipeline computes directional instability based on gradients from a single rendering viewpoint, aggregating sensitivity scores across adjacent camera poses would provide a richer, more stable estimate of each Gaussian's reliability. This multi-view accumulation could generate volumetric uncertainty maps that highlight persistently unstable regions, such as thin structures, occluded geometry, or reflective surfaces, and enable more informed filtering decisions. These maps could also serve as input to downstream tasks such as adaptive mesh simplification, visibility-aware reconstruction, or uncertainty-aware SLAM, where understanding where and why the model is uncertain is critical for robustness and decision-making.

Beyond rendering-time filtering, gradient sensitivity can also be leveraged as a signal for data-driven pruning or regularization. Because sensitivity scores reflect directional instability, they can reveal which Gaussians repeatedly behave erratically across multiple views (i.e. those in poorly observed, ambiguous, or textureless regions). These unstable components can be pruned or down-weighted, resulting in leaner, more efficient models. This stands in contrast to traditional filtering heuristics based on alpha thresholds or visibility counts, offering a more principled, geometry-

and view-aware criterion. Additionally, this same sensitivity signal could be integrated directly into training as a form of structure-aware regularization. Penalizing unstable Gaussians during optimization would encourage the model to suppress the emergence of spurious structure, improving generalization and visual coherence, especially in sparse or occlusion-heavy datasets.

Another compelling direction is to use gradient sensitivity for view planning and active reconstruction. By identifying regions of high epistemic uncertainty, a system could actively select future camera poses that are most likely to reduce uncertainty and improve fidelity. This would support applications such as aerial photogrammetry, robotic navigation, or AR/VR scene completion, where minimizing uncertainty in semantically meaningful regions is essential for operational effectiveness.

A critical area of future work lies in the development of more semantically grounding no-reference image quality assessment (NR-IQA) metrics. Existing tools like NIQE, BRISQUE, and PIQE, while widely adopted, are not designed to capture the kinds of semantic or geometric distortions that plague view synthesis such as ghosting, hallucinated edges, or occlusion-related artifacts. They often assign misleadingly high quality scores to visually implausible or semantically incoherent renderings. To address this, our work proposed two complementary evaluation metrics: the SAM Reference Similarity Score (SRSS) and the CLIP Alignment Score (CAS). SRSS quantifies alignment with canonical object structure based on feature similarity to a reference bank of instance-level embeddings, while CAS assesses prompt-based alignment using CLIP-derived cosine similarity with masked image regions.

These two metrics offer orthogonal insights into semantic correctness: SRSS captures precise structural correspondence with known references and CAS caputres alignment to flexible, zero-shot textual prompts. However, they too can be extended. For SRSS, future work could involve dynamically selecting or learning reference exemplars to improve generalization across occlusion, scale, or pose changes. It may also be beneficial to explore how SRSS could be used in temporally consistent settings, i.e. evaluating video sequences or navigation paths, to detect semantic drift or identity loss over time. CAS, by contrast, could benefit from improved prompt engineering, multi-prompt ensembles, or the use of spatially grounded language models that better reflect complex object relationships and context (e.g., “a yellow robot on a concrete floor”).

Looking forward, both SRSS and CAS could be integrated into learning-based NR-IQA models

that are jointly trained with preference data to predict human judgments of view synthesis quality. Such hybrid metrics could balance low-level fidelity and high-level recognizability, addressing the gap between photometric error and perceptual understanding. They could also support the fine-tuning of generative pipelines with loss functions that directly reflect human preferences for clarity, plausibility, and semantic salience.

Ultimately, this future work points toward a broader goal: building rendering pipelines and evaluation tools that are not only real-time and data-efficient, but also perceptually trustworthy and semantically aware. By combining gradient-based uncertainty filtering, semantically grounded evaluation metrics, and task-driven utility (e.g., pruning, view planning, or downstream analysis), we can move toward systems that not only look good but also know what they're looking at.

5.2 Conclusion

Gradient sensitivity filtering for 3D Gaussian Splatting (3DGS) was introduced in this work as a principled response to rendering artifacts that arise under epistemic uncertainty, particularly in under-constrained or out-of-distribution viewpoints. While 3DGS represents a breakthrough in real-time, compact scene representation, it remains vulnerable to view-dependent artifacts in settings with sparse, ambiguous, or occluded supervision. These issues manifest as anisotropic instabilities, where Gaussians with poorly estimated appearance or geometry project unreliable radiance that distorts the rendered image.

To address this challenge, we propose a novel two-pass filtering pipeline that quantifies per-Gaussian instability using a gradient sensitivity score derived from intermediate rotation-space gradients. This score acts as an efficient, differentiable surrogate for directional uncertainty, capturing local anisotropic behavior in a way that approximates the diagonal of the Fisher Information Matrix. Leveraging this score, our rasterization pipeline first computes pixel-level contribution ratios and then filters out visually unstable components in a second pass based on per-pixel sensitivity thresholds. This enables dynamic, view-aware filtering without retraining the scene representation and provides precise control over artifact suppression.

Critically, this filtering strategy operates entirely post hoc, without modifying the trained 3DGS

model or requiring any additional learning. It generalizes well across both indoor and outdoor environments, particularly excelling in scenes characterized by occlusion, visual clutter, or narrow fields of view. Compared to alternative uncertainty-aware filtering techniques like BayesRays, our method achieves similar or superior artifact suppression with significantly less computational overhead, and without the need for ensemble sampling or variational approximations.

To evaluate the efficacy of our approach, we introduced two complementary semantic evaluation metrics: the SAM Reference Similarity Score (SRSS) and the CLIP Alignment Score (CAS). These metrics move beyond traditional perceptual measures by enabling localized, object-centric semantic assessment. SRSS compares masked region features against a bank of reference embeddings derived from diverse views of the canonical object, enabling robust instance-level matching. CAS evaluates the semantic alignment of a masked region directly against a text prompt using contrastive CLIP embeddings, providing a lightweight, zero-shot method for prompt-driven validation. Together, these metrics offer orthogonal but complementary insight: SRSS is robust to partial visibility and viewpoint variation, while CAS is sensitive to prompt specificity and holistic appearance alignment.

Our experiments on diverse datasets such as NeRF-on-the-Go and Deep Blending demonstrated that this filtering strategy reliably suppresses noise and improves both perceptual and semantic fidelity. Quantitative evaluations using NR-IQA metrics (NIQE, BRISQUE, PIQE) confirmed consistent gains in image quality, while qualitative and metric-driven analysis using SRSS and CAS validated that semantic alignment is preserved, or even enhanced, after filtering. These improvements are especially notable in scenarios where traditional radiance field methods or probabilistic filters like BayesRays struggle, particularly in indoor environments with frequent occlusions and ambiguous geometry.

This work contributes a lightweight and scalable mechanism for improving view-dependent fidelity in 3D Gaussian Splatting without retraining or sacrificing real-time performance. It bridges the gap between uncertainty modeling and efficient rendering by showing that directional gradient sensitivity can serve as a powerful proxy for epistemic confidence. The introduction of SRSS and CAS further provides the community with practical tools for semantic evaluation in novel view synthesis, tools that are aligned with both human perception and task-relevant structure. As real-time rendering systems increasingly move into open-world, zero-shot applications, our method offers a

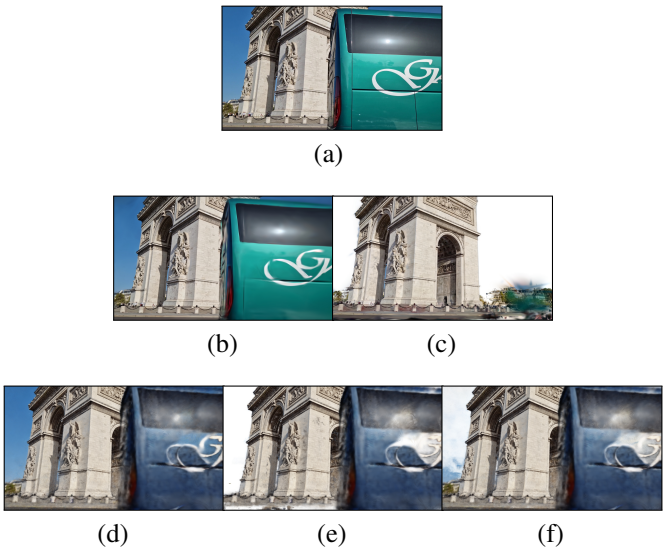
practical, interpretable, and effective strategy for mitigating the risks of uncertainty while preserving high-quality outputs.

Appendix A

A.1 Epistemic Artifact Removal - NeRF-on-the-go Dataset

Frame	Scene	Original 3DGS	Our Filter	Orignal NeRF	BR = .25	BR = .5
frame 640.jpg	arcdetriomphe	.282	0.297	0.316	0.296	0.294
IMG 8631.jpg	patio high	0.260986328	0.266113281	0.207519531	0.209228516	0.182006836
IMG 8295.jpg	spot	0.273925781	0.275878906	0.231567383	0.259521484	0.205078125

Table A.1: CAS results for Figures A.1, A.2, A.3



(a) (b) (c) (d) (e) (f)

Figure A.1: frame_0640.jpg, arcdetriomphe (a) Ground truth (b) Original 3DGS reconstruction (c) Our filter: $\tau_{grad.}$ = .00001, τ_{ratio} = .5 (d) Original NeRF reconstruction (e) BayesRays: filter = .25 (d) BayesRays: filter = .5

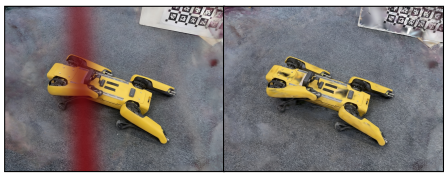
(a)

(b) (c)

(d) (e) (f)

Figure A.2: IMG_8631.jpg, patio_high (a) Ground truth (b) Original 3DGS reconstruction (c) Our filter: $\tau_{grad.}$ = .00001, τ_{ratio} = .5 (d) Original NeRF reconstruction (e) BayesRays: filter = .25 (d) BayesRays: filter = .5

(a)



(b) (c)

(d) (e) (f)

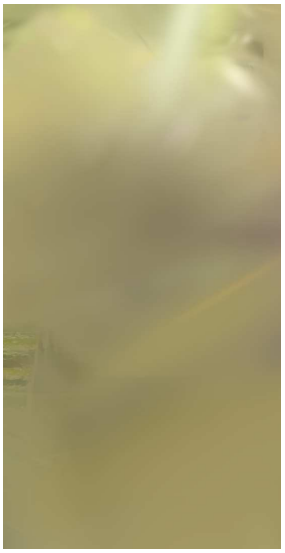
Figure A.3: IMG_8295.jpg, spot scene (a) Ground truth (b) Original 3DGS reconstruction (c) Our filter: $\tau_{grad.}$ = .00001, τ_{ratio} = .5 (d) Original NeRF reconstruction (e) BayesRays: filter = .25 (d) BayesRays: filter = .5

A.2 Additional Examples - Deep Blending Dataset

(a)

(b)

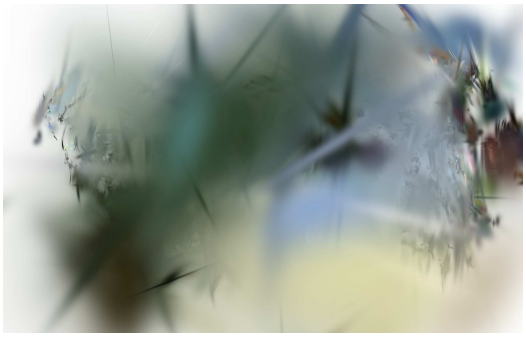
Figure A.4: DrJohnson novel view from 3DGS reconstruction (a) and filtered output (b). Threshold = .00001, ratio = .5.



(a)

(b)

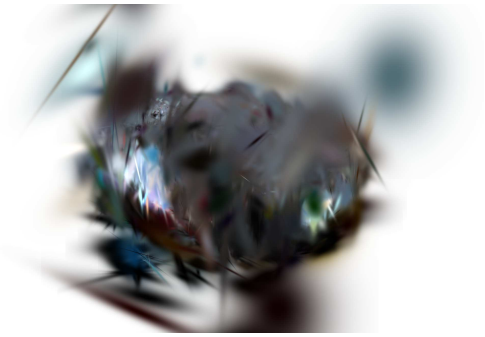
Figure A.5: Bedroom novel view from 3DGS reconstruction (a) and filtered output (b). Threshold = .00001, ratio = .5.



(a) Aquarium-20 (no filter)

(b) Aquarium-20 (filtered: .00001, .5)

Figure A.6: Side-by-side comparisons of unfiltered and filtered views: gradient threshold, ratio threshold



(c) CreepyAttic (no filter)

(d) CreepyAttic (filtered: .00002, .5)

(c) SainteAnne (no filter)

(d) SainteAnne (filtered: .00002, .5)

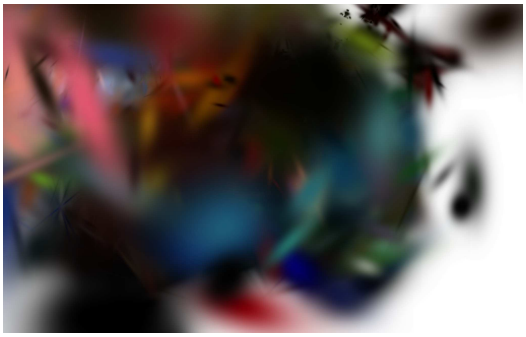
(a) Ponche (no filter)

(b) Ponche (filtered: .0001, .5)

(a) NightSnow (cropped, no filter)

(b) NightSnow (cropped, filtered: .00001, .5)

Figure A.7: Side-by-side comparisons of unfiltered and filtered views: gradient threshold, ratio threshold



(a) Library (cropped, no filter)

(b) Library (cropped, filtered: .00001, .5)

(a) Shed (no filter, noise from snowy scene)

(b) Shed (filtered: .00025, .5)

Figure A.8: Side-by-side comparisons of unfiltered and filtered views: gradient threshold, ratio threshold

Appendix B

Pseudo-code

Algorithm 1 Gradient-based Filtering of Noisy Gaussians in 3DGS Rendering

Require: 3D Gaussian model $\mathcal{G}$, Gradient threshold $\tau_{grad.}$, Ratio threshold τ_{ratio}
Ensure: Filtered Image I

Init.: RejectedCount$[G_i] \leftarrow 0$, UsedCount$[G_i] \leftarrow 0$, $\forall G_i \in \mathcal{G}$
for each Gaussian $G_i \in \mathcal{G}$ **do**
    Per-Gaussian pre-processing (transform into camera space)
end for
Sort Gaussians by depth (depth-ordered Gaussian list)
First Pass: Gradient Analysis
for each pixel ray r **do**
    for each Gaussian G_i intersected by r **do**
        Compute gradient magnitude $\|\nabla C\|$ at ray-Gaussian intersection using Eq. 19
        UsedCount$[G_i] \leftarrow$ UsedCount$[G_i] + 1$
        if $\|\nabla C\| > \tau_{grad.}$ **then**
            RejectedCount$[G_i] \leftarrow$ RejectedCount$[G_i] + 1$
        end if
    end for
end for
Second Pass: Conditional Rendering
for each pixel ray r **do**
    for each Gaussian G_i intersected by r **do**
        Compute rejection ratio $R_i \leftarrow \frac{\text{RejectedCount}[G_i]}{\text{UsedCount}[G_i]}$
        if $R_i > \tau_{ratio}$ **then**
            Remove G_i from rendering for this viewpoint
        end if
    end for
end for
Render the filtered image I using remaining Gaussians

Appendix C

Prompts

Grounding DINO prompt for object detection:

Scene	Prompt
arc de triomphe:	”Arc de Triomphe”
patio:	”Gondola”
spot:	”robot dog”

Positive prompts for contrastive filter:

Scene	Prompt
arc de triomphe:	”Arc de Triomphe”
patio:	”a brightly painted graffiti gondola cabin exterior as a courtyard art installation”
spot:	”A yellow quadruped robot dog (Boston Dynamics Spot) sitting on a carpeted floor.”

Negative prompts for contrastive filter

Scene	Prompt
arc de triomphe:	[”sculpture”, ”ornamentation”, ”vehicle”, ”pedestrian”, ”pavement”]
patio:	[”people”, ”person”, ”teddy”, ”man”, ”woman”, ”crowd”, ”ground”, ”floor”, ”wall”, ”windows”]
spot:	[”person”, ”leg”, ”chair”, ”desk”, ”cable”, ”box”, ”red couch”]

Bibliography

[1] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[2] Tran Anh, Nguyen Tan, Than Le, Chi Hieu Le, Jamaluddin Mahmud, Mohd Abd Latif, and Quang Nguyen Ho. *Digital Twins of Robotic Systems: Increasing Capability for Industrial Applications*, pages 241–258. 02 2023.

[3] Ascard, Marcus and Movahedi, Farjam. Assessing the Efficiency of COLMAP, DROID-SLAM, and NeRF-SLAM in 3D Road Scene Reconstruction, 2023. Student Paper.

[4] Michael Batty. Digital twins. *Environment and Planning B*, 45(5):817–820, 2018.

[5] Frédéric Bosché, Mahmoud Ahmed, Yelda Turkan, Carl T. Haas, and Ralph Haas. The value of integrating scan-to-bim and scan-vs-bim techniques for construction monitoring using laser scanning and bim: The case of cylindrical mep components. *Automation in Construction*, 49:201–213, 2015. 30th ISARC Special Issue.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[7] Aditya Chandrasekar, Goirik Chakrabarty, Jai Bardhan, Ramya Hebbalaguppe, and Prathosh AP. Remove: A reference-free metric for object erasure, 2024.

[8] Hanlin Chen, Chen Li, Yunsong Wang, and Gim Hee Lee. Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance, 2025.

[9] Mohamed Debbagh. Neural radiance fields (nerfs): A review and some recent developments, 04 2023.

[10] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, page 11–20, New York, NY, USA, 1996. Association for Computing Machinery.

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022.

[12] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, June 2024.

[13] Aidan Fuller, Zhong Fan, Charles Day, and Chris Barlow. Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, 8:108952–108971, 2020.

[14] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing, 2005.

[15] Lily Goli, Cody Reading, Silvia Sell´an, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification in neural radiance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. *The Lumigraph*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023.

[17] Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. Pup 3d-gs: Principled uncertainty pruning for 3d gaussian splatting. *arXiv*, 2024.

[18] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[19] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. 37(6):257:1–257:15, 2018.

[20] Max Hermann, Hyovin Kwak, Boitumelo Ruf, and Martin Weinmann. Leveraging neural radiance fields for large-scale 3d reconstruction from aerial imagery. *Remote Sensing*, 16(24), 2024.

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[22] Qixing Huang and Yasutaka Furukawa. Image based reconstruction ii.

[23] Wen Jiang, Boshu Lei, , and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information, 2023.

[24] Maged N Kamel Boulos and Peng Zhang. Digital twins: From personalised medicine to precision public health. *J. Pers. Med.*, 11(8):745, July 2021.

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimk¨uhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

[26] Leonid Keselman and Martial Hebert. Approximate differentiable rendering with algebraic surfaces. In *European Conference on Computer Vision (ECCV)*, 2022.

[27] Ye Keyang, Hou Qiming, and Zhou Kun. 3d gaussian splatting with deferred reflection. 2024.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[29] Marc Levoy and Pat Hanrahan. *Light Field Rendering*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023.

[30] Mingwei Li, Pu Pang, Hehe Fan, Hua Huang, and Yi Yang. Tsgs: Improving gaussian splatting for transparent surface reconstruction via normal and de-lighting priors, 2025.

[31] Shaoming Li, Qing Cai, Songqi Kong, Runqing Tan, Heng Tong, Shiji Qiu, Yongguo Jiang, and Zhi Liu. Mesc-3d:mining effective semantic cues for 3d reconstruction from a single image, 2025.

[32] Xudong Li, Timin Gao, Runze Hu, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Jingyuan Zheng, Yunhang Shen, Ke Li, Yutao Liu, Pingyang Dai, and Rongrong Ji. Adaptive feature selection for no-reference image quality assessment by mitigating semantic noise sensitivity, 2024.

[33] Liewen Liao, Weihao Yan, Ming Yang, and Songan Zhang. Learning-based 3d reconstruction in autonomous driving: A comprehensive survey, 2025.

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[35] Alwyn Mathew, Shuyan Li, Kacper Pluta, Rahima Djahel, and Ioannis Brilakis. Digital twin enabled construction progress monitoring. 07 2024.

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

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

[38] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012.

[39] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[40] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[41] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[42] Theo Moons, Luc Van Gool, and Maarten Vergauwen. 3d reconstruction from multiple images part 1: Principles. *Found. Trends. Comput. Graph. Vis.*, 4(4):287–404, April 2010.

[43] L. Morelli, F. Ioli, R. Beber, F. Menna, F. Remondino, and A. Vitti. Colmap-slam: A framework for visual odometry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1/W1-2023:317–324, 2023.

[44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022.

[45] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, 2015.

[46] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011.

[47] Sung Hyun Park, Ki-Yoon Kim, Yoo Min Kim, and Woo Jin Hyung. Patient-specific virtual three-dimensional surgical navigation for gastric cancer surgery: A prospective study for preoperative planning and intraoperative guidance. *Front. Oncol.*, 13:1140175, February 2023.

[48] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[50] Fabio Remondino. Heritage recording and 3d modeling with photogrammetry and 3d scanning. *Remote Sensing*, 3(6):1104–1138, 2011.

[51] Fabio Remondino and Sabry El-Hakim. Image-based 3d modelling: A review. *The Photogrammetric Record*, 21:269 – 291, 09 2006.

[52] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[53] Henry Robb, Gemma Scrimgeour, Piers Boshier, Anna Przedlacka, Svetlana Balyasnikova, Gina Brown, Fernando Bello, and Christos Kontovounisios. The current and possible future role of 3D modelling within oesophagogastric surgery: a scoping review. *Surg. Endosc.*, 36(8):5907–5920, August 2022.

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

[55] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[56] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528, 2006.

[57] Concetta Semeraro, Mario Lezoche, Hervé Panetto, and Michele Dassisti. Digital twin paradigm: A systematic literature review. *Computers in Industry*, 130, 05 2021.

[58] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.

[59] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer Cham, 2 edition, 2022. Includes 374 b/w and 144 color illustrations.

[60] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

[61] Fei Tao, Jiangfeng Cheng, Qinglin Qi, Meng Zhang, He Zhang, and Fangyuan Sui. Digital twin-driven product design, manufacturing and service with big data. *Int. J. Adv. Manuf. Technol.*, 94(9-12):3563–3576, February 2018.

[62] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. Intelligent Robotics and Autonomous Agents. The MIT Press, Cambridge, MA, August 2005. Hardcover, 8 x 9 in.

[63] Alexandre Vallee. Digital twin for healthcare systems. *Front. Digit. Health*, 5:1253050, September 2023.

[64] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

[65] Haato Watanabe, Kenji Tojo, and Nobuyuki Umetani. 3d gabor splatting: Reconstruction of high-frequency surface texture using gabor noise, 2025.

[66] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *The Twelfth International Conference on Learning Representations*, 2024.

[67] Ziyi Yang, Xinyu Gao, Yangtian Sun, Yihua Huang, Xiaoyang Lyu, Wen Zhou, Shaohui Jiao, Xiaojuan Qi, and Xiaogang Jin. Spec-gaussian: Anisotropic view-dependent appearance for 3d gaussian splatting, 2024.

[68] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient high-quality compact surface reconstruction in unbounded scenes. *SIGGRAPH ASIA*, 2024.

[69] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[70] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum*, 37(2):625–652, 2018.

[71] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002.