Recovering Visual Saliency from Intrinsic Properties of 3D Gaussian Splatting

Xinya Bi

Recovering Visual Saliency from Intrinsic Properties of 3D Gaussian Splatting

by

Xinya Bi

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on June 10, 2026 at 16:30.

Student number: 6195350
Project duration: November, 2025 – June, 2026
Thesis committee: Liangliang Nan, Martijn Meijers, Weixiao Gao

Cover: Self made

TU Delft

Preface

This thesis grows out of my interest in understanding 3D scene representations using methods beyond LiDAR-collected point clouds. At first, I worked with the CLIP model, attempting to embed semantic information into 3D Gaussian Splatting. Later, I delved into the intrinsic properties of 3D Gaussian Splatting, which surprisingly exhibited behaviours I did not expect at first. Perhaps that is the essence of research: exploring things you might never have paid attention to.

During this journey, I would like to express my sincere gratitude to my supervisors, Liangliang Nan and Martijn Meijers. In our bi-weekly meetings, they always provided valuable insights on my research and helped clear away the confusions I encountered along the way.

Last but not least, I would like to thank my mom and dad for their support through this process. Without your company, it would have been impossible for me to complete this project.

Xinya Bi
Delft, June 2026

Contents

Preface i

1 Introduction 1

2 Preliminary Materials 3
- 2.1 Fundamentals of 3D Gaussian Splatting . . . 3
- 2.2 Visual Saliency Detection . . . 4
- 2.3 Machine Learning and Statistics Evaluation . . . 5
- 2.4 Capture Intent in Multi-view Reconstruction . . . 7
- 2.5 Scene Understanding in 3D Gaussian Splatting . . . 8

3 Scientific Article 10
- 3.1 Abstract . . . 11
- 3.2 Introduction . . . 11
- 3.3 Related Works . . . 12
- 3.4 Method . . . 13
  - 3.4.1 Gaussian Properties as Saliency Indicators . . . 13
  - 3.4.2 Feature Extraction . . . 13
  - 3.4.3 Machine Learning Classifier . . . 14
- 3.5 Experiments . . . 15
  - 3.5.1 Experiment Setup . . . 15
  - 3.5.2 Feature Discriminability . . . 16
  - 3.5.3 Saliency Alignment Evaluation . . . 16
  - 3.5.4 Foreground Extraction . . . 18
  - 3.5.5 Ablation Study . . . 19
- 3.6 Conclusion and Future Work . . . 20
  - 3.6.1 Conclusion . . . 20
  - 3.6.2 Limitations and Future Work . . . 20

4 Conclusion 24
- 4.1 Reflection on the Research . . . 24
- 4.2 Relation to MSc Geomatics for the Built Environment . . . 24
- 4.3 Limitations Revisited . . . 26
- 4.4 Future Work . . . 26

References 27

A AI Declaration 30

B Reproducibility self-assessment 31

1

Introduction

3D scene reconstruction is an important task in Geomatics, enabling applications such as urban mapping (Herold & Hecht, 2018), infrastructure monitoring (Fathi et al., 2015), and cultural heritage documentation (Gomes et al., 2014; Llull et al., 2023). Traditionally, this has relied on LiDAR-based point cloud, which provides accurate geometry but requires expensive hardware and time-consuming data collection. Novel view synthesis offers a compelling alternative: using only photographs, it reconstructs a scene and enables rendering from arbitrary viewpoints. 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has recently emerged as a state-of-the-art method, representing scenes as millions of Gaussian primitives that can be rendered in real time at high quality. Its application to large-scale urban environments has been demonstrated in recent work (Li et al., 2024; Y. Liu, Luo, et al., 2024; Miao et al., 2025), showing promising results for scene reconstruction from image data alone.

Unlike LiDAR point clouds, where each point records only geometric position, each Gaussian primitive in a 3DGS model carries a rich set of attributes: position, size, opacity, orientation, and appearance, all of which are determined entirely by the photographs used during reconstruction. This means that Gaussian properties do not only reflect the geometry and appearance of the scene, but also implicitly encode how the scene was observed and, by extension, what the photographer chose to focus on.

This raises a natural question: **Can this capture intent be recovered from the intrinsic properties of a trained 3DGS model, without any external supervision?** More specifically, these following sub research questions guide this work:

1. Do the intrinsic properties of 3DGS primitives serve as meaningful indicators of visual saliency and capture intent?
2. Do the recovered 3D saliency masks align with human visual perception as measured by 2D saliency detection models?
3. Can 3D saliency derived from Gaussian properties generalize to unseen scenes and support downstream tasks such as foreground extraction?

This thesis investigates these questions and proposes a method to do so. We extract a set of features from the properties of each Gaussian primitive and train a machine learning classifier to distinguish between the photographer's primary subject and the background. Crucially, our method requires only the reconstructed 3D model and camera parameters, no original photographs or external 2D foundation models are needed at inference time.

We experiment on 16 annotated scenes from three datasets, including two self-captured scenes around landmarks in Delft and demonstrates that the method successfully recovers the photographer's region of interest. The recovered 3D saliency shows strong alignment with human visual perception and generalizes to unseen scenes. These results suggest that 3DGS models encode richer information than what is strictly needed for rendering. This has direct implications for Geomatics workflows: as image-based 3D reconstruction becomes more accessible and cost-effective compared to LiDAR, understanding what a scene captures and what it misses becomes increasingly important. Automatically recovering

the photographer's region of interest could guide more targeted data acquisition, prioritize high-saliency regions for denser sampling, and enable perceptual quality-driven compression of large-scale urban models.

This thesis is structured around a scientific article, supplemented by preliminary materials and an extended discussion. The four parts are organized as follows.

Part 1 (this chapter) provides a general introduction to the research for readers without a specialist background in 3DGS and machine learning. It explains the research context, the high-level approach, and the structure of the thesis.

Part 2 provides the technical background needed to understand the scientific article in Part 3. It is organized into three chapters. Chapter 2.1 introduces 3D Gaussian Splatting, covering the representation and rendering pipeline. This background directly supports the Method section of the article. Chapter 2.2 introduces visual saliency detection, explaining the concept of saliency, the distinction between bottom-up and top-down attention, and the deep learning models used as a 2D reference in our evaluation. This background supports the Related Work and Experiment sections. Chapter 2.3 introduces the machine learning and statistical methods used in this thesis, including Random Forest, cross-validation, and Cohen's d effect size. This background supports both the Method and Experiment sections. Chapter 2.4 introduces the relationship between photographer behaviour and 3DGS optimisation, explaining how top-down capture intent is implicitly transformed into measurable bottom-up geometric statistics during the reconstruction process. This background provides the theoretical motivation for the approach presented in the article. Chapter 2.5 reviews existing approaches to scene understanding in 3DGS, covering methods that integrate 2D foundation models as external semantic supervision, and discusses their shared limitations. It further explains why using 2D saliency maps as direct supervision for 3DGS is an ill-posed problem due to the view-dependent nature of 2D saliency signals. Together, these two chapters establish the research gap this thesis addresses.

Part 3 is the scientific article, written in the style of a conference publication. It contains an abstract, introduction, related work, methodology, experimental results, conclusion and limitations.

Part 4 provides a closing discussion that expands on the article. It reflects on the research process, connects the work to the MSc Geomatics for the Built Environment programme, revisits limitations in more depth, and outlines directions for future work.

2

Preliminary Materials

This chapter gives the background knowledge of 3D Gaussian Splatting, saliency detection, the used machine learning method and research gap analysis.

2.1. Fundamentals of 3D Gaussian Splatting

3D Gaussian Splatting is an explicit radiance field based method that represents the reconstructed scene using a large number of anisotropic 3D Gaussian ellipsoids (Kerbl et al., 2023). The main flow chart is demonstrated in Figure 2.1

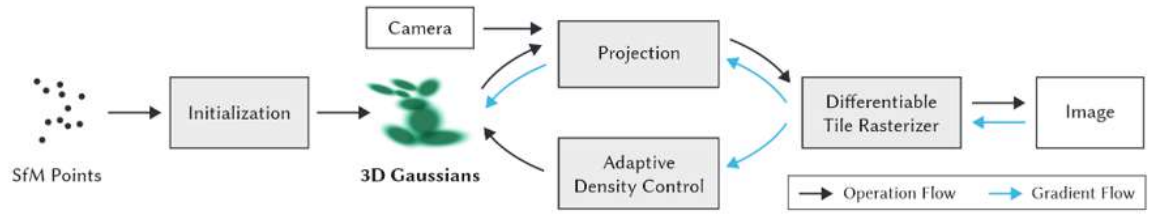


Figure 2.1: 3DGS flow chart

In the initialisation step, SfM reconstructs a sparse point cloud from the training views, which is then used to initialise the Gaussians. Each Gaussian is defined by its mean (μ), covariance (Σ), opacity (α), and spherical harmonics (SH). They are spatially modelled using a 3D Gaussian distribution:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.1}$$

The covariance matrix Σ can be further decomposed into rotation (R) and scaling (S) matrices:

$$\Sigma = RSS^TR^T \tag{2.2}$$

During the optimisation process, for a given camera pose, a differentiable rasterizer renders a 2D image by projecting all Gaussians observed from the camera pose onto the image plane. This process consists of several sub-steps:

(1) 3D to 2D projection: Using the intrinsic and extrinsic camera parameters, the 3D Gaussians are projected onto the 2D image plane. The covariance matrix is computed as

$$\Sigma' = JW\Sigma W^TJ^T \tag{2.3}$$

where J is the Jacobian matrix providing an affine approximation of the projective transformation, and W denotes the view transformation from world coordinates to camera coordinates.

(2) Depth sorting: For every 16×16 pixel tile of the image, the projected Gaussians that intersect with the tile are sorted by depth in a front-to-back order.

(3) Colour and opacity accumulation: For each pixel, the final colour is computed by blending the depth-sorted Gaussians that overlap with the pixel:

$$C_p = \sum_{i \in \mathcal{N}_p} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{2.4}$$

where $\mathcal{N}_p$ is the sorted set of Gaussians overlapping pixel p, c_i is the view-dependent colour computed from spherical harmonics, and α_i denotes the opacity of Gaussian i. $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ represents the accumulated transmittance contributed by all Gaussians in front of Gaussian i.

The optimisation process iteratively compares the rendered image with the training views. The Gaussian parameters are optimised using stochastic, gradient descent optimisation. The loss function is defined as:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{D\text{-}SSIM} \tag{2.5}$$

$\mathcal{L}_1$ denotes the mean absolute error between the rendered pixel colours and the ground truth, ensuring colour accuracy at the pixel level. $\mathcal{L}_{D-SSIM}$ represents the structural similarity (SSIM) based loss, which helps maintain the structural and perceptual consistency of the reconstructed scene.
Optimisation is interleaved with adaptive density control, which densifies under-reconstructed regions guided by gradient information and prunes Gaussians with low opacity, while also limiting oversized splats based on screen-space radius.

2.2. Visual Saliency Detection

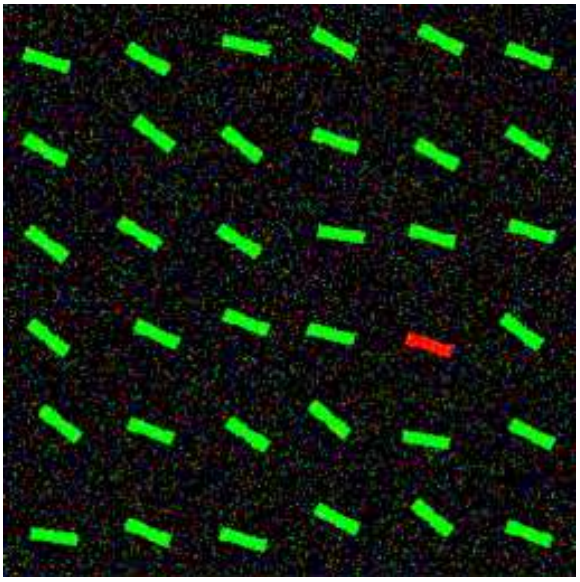
Definition of Visual Saliency Visual saliency is defined as the distinct perceptual quality that makes certain parts of an image or environment, such as objects or regions, stand out from their surroundings (Itti et al., 1998). It is driven by two main factors (Connor et al., 2004; Zhang et al., 2025):

- **Bottom-up:** A stimulus-driven signal that directs attention towards regions with distinct low-level features, such as color contrast, sharp edges, or unusual textures.
- **Top-down:** A goal-driven signal that directs attention towards objects that align with the observer's internal goals, knowledge, or expectations.

Below are examples of bottom-up (Figure 2.2a) and top-down (Figure 2.2b) saliency signals.

The Advancement of Saliency Detection 2D saliency detection is fundamentally based on local contrast within a single image frame. Early methods, such as Itti-Koch's method (Itti et al., 1998), decomposes an image into multiple spatial scales using Gaussian pyramids, extracting low-level features such as color, brightness, and orientation. For each feature, local regions are compared against their surroundings to identify distinctive areas, which are then combined into a single saliency map where brighter regions indicate higher salience. This approach successfully models early primate visual processing. Hou and Zhang (Hou & Zhang, 2007) propose an alternative approach that operates in the frequency domain, extracting the spectral residual of an image's log spectrum to identify statistically unexpected regions as salient, offering a computationally efficient alternative to feature-based methods.

The emergence of deep neural networks has greatly advanced the field. Early deep learning methods separate the input images into small regions then use multi-layer perceptrons to calculate each small regions saliency score, but this method overlooks spatial relations between these regions (Dhara & Kumar, 2025). DHSNet (N. Liu & Han, 2016) addresses this by combining both paradigms: it first generates a coarse global saliency map from the entire image to capture global structure, then progressively refines local details through a hierarchical recurrent convolutional network. U^2-Net (X. Qin et al., 2020) further advances this direction with a two-level nested U-structure that captures both local and global contextual information at multiple scales, achieving accurate saliency prediction with sharp object boundaries.



(a) The red bar catches attention effortlessly due to its distinct color, illustrating bottom-up saliency.

(b) The banana does not stand out unless the observer is actively looking for it, illustrating top-down saliency.

Figure 2.2: Two categories of visual saliency. Bottom-up saliency is driven by low-level stimulus features, while top-down saliency is driven by the observer's goals and expectations.

Though salient object detection utilizing deep neural networks has achieved high accuracy on benchmark datasets, several open problems remain in real-world scenarios. First, when a scene contains multiple objects with similar visual properties, existing methods struggle to identify the true subject of interest. Second, object occlusion is a common challenge: when the primary subject is partially or fully occluded by other objects, current saliency detection methods often fail to correctly detect it. Third, complex lighting conditions, such as strong shadows or specular highlights, can introduce local contrast that misleads the detector towards irrelevant regions.

These challenges are further compounded in 3D scenes. A single 2D frame captures only one viewpoint of the scene, and the salience of an object in that frame may not reflect its importance across all viewpoints. An object that is partially occluded or poorly lit in one view may be the primary subject of the scene. This information is only revealed through multi-view consistency. This limitation motivates the use of 3D representations, where information from multiple viewpoints is aggregated, as a more reliable basis for recovering the photographer's capture intent.

2.3. Machine Learning and Statistics Evaluation

Random Forest Random Forest is an ensemble machine learning algorithm that builds multiple decision trees on random subsets of the training data and random subsets of the features. Each tree independently learns a decision boundary, and the final prediction is determined by majority voting across all trees. This strategy reduces overfitting compared to a single decision tree, making Random Forest particularly well-suited for small datasets where individual trees might otherwise memorize the training data. The following image Figure 2.3 shows an example workflow of Random Forest.

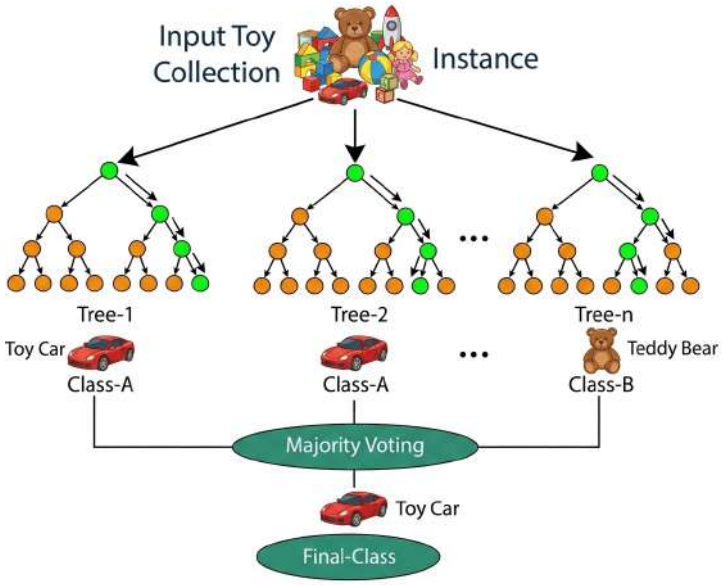


Figure 2.3: Random Forest mechanism

Several hyperparameters control the depth and complexity of the trees, directly impacting the trade-off between overfitting and underfitting. In this thesis, grid search is used to find the optimal hyperparameters on the validation scene prior to final model training.

Random Forest is chosen as the classifier in this thesis for three reasons. First, it handles small training sets without overfitting due to its ensemble nature. Second, it provides feature importance scores that reflect each feature's contribution to classification decisions, offering interpretable insights into which Gaussian properties are most discriminative for recovering visual saliency. Third, it is computationally efficient at inference time, requiring only a single forward pass through the trained trees.

Cross Validation Cross-validation is used to assess how well a machine learning model generalizes to unseen data while reducing the risk of overfitting. The simplest form is holdout validation, where the dataset is split into a training set and a test set, and the model is evaluated on the test set after training. The usual training test split ratio is 70:30, 80:20. The division depends on the dataset size and problem type. While straightforward, this approach is sensitive to how the data is split, particularly when the dataset is small. Other common strategies are k-fold cross-validation and leave-one-out cross-validation (LOOCV). In k-fold cross-validation, the dataset is divided into k equal-sized folds, the model is trained on $k-1$ folds and tested on the remaining fold, repeating this process k times. LOOCV is a special case where k equals the total number of samples, meaning the model is trained on all samples except one, which is used for testing. This is repeated for every sample in the dataset.

When the dataset is small, LOOCV is generally preferred over k-fold cross-validation because it maximizes the use of available training data. In this thesis, only 12 scenes are available for training, making LOOCV the appropriate choice. LOOCV is applied within the training set to evaluate the stability of the model across different scene types: a low standard deviation of F1 scores across folds indicates that the model generalizes consistently rather than overfitting to specific scene characteristics.

However, LOOCV alone cannot assess generalization to entirely new scenes, since all scenes used in LOOCV belong to the training set. A separate held-out test set of three scenes, unseen during both training and LOOCV, is therefore used to provide an unbiased estimate of the model's performance on new scenes. Together, LOOCV and the held-out test set answer two complementary questions: whether the model is stable within the training distribution, and whether it generalizes in new cases.

Cohen's d Cohen's d is a measure of effect size that quantifies how well a variable separates two groups. While statistical significance tests such as the p-value indicate whether a difference between

groups is likely to be real, they do not convey how large that difference is. Cohen's d addresses this by expressing the separation between two group means in units of their pooled standard deviation:

$$d = \frac{|M_2 - M_1|}{\sigma_{\text{pooled}}} \tag{2.6}$$

where M_1 and M_2 are the means of the two groups and σ_{pooled} is their pooled standard deviation. Values of $d \leq 0.2$, $0.2 < d \leq 0.5$, $0.5 < d \leq 0.8$, and $d > 0.8$ indicate negligible, small, medium, and large separation respectively (Diener, 2010).

In this thesis, Cohen's d is used to evaluate the discriminability of each Gaussian feature between foreground and background primitives. This allows us to systematically identify which features provide the strongest separation signal before training the classifier, providing a principled basis for feature selection and ablation study design.

Figure 2.4 illustrates the feature distributions of primary object and background Gaussians for a representative scene. Features where the two distributions show little overlap are strong discriminators, corresponding to a high Cohen's d value. Features with heavily overlapping distributions contribute less to separating the two classes.

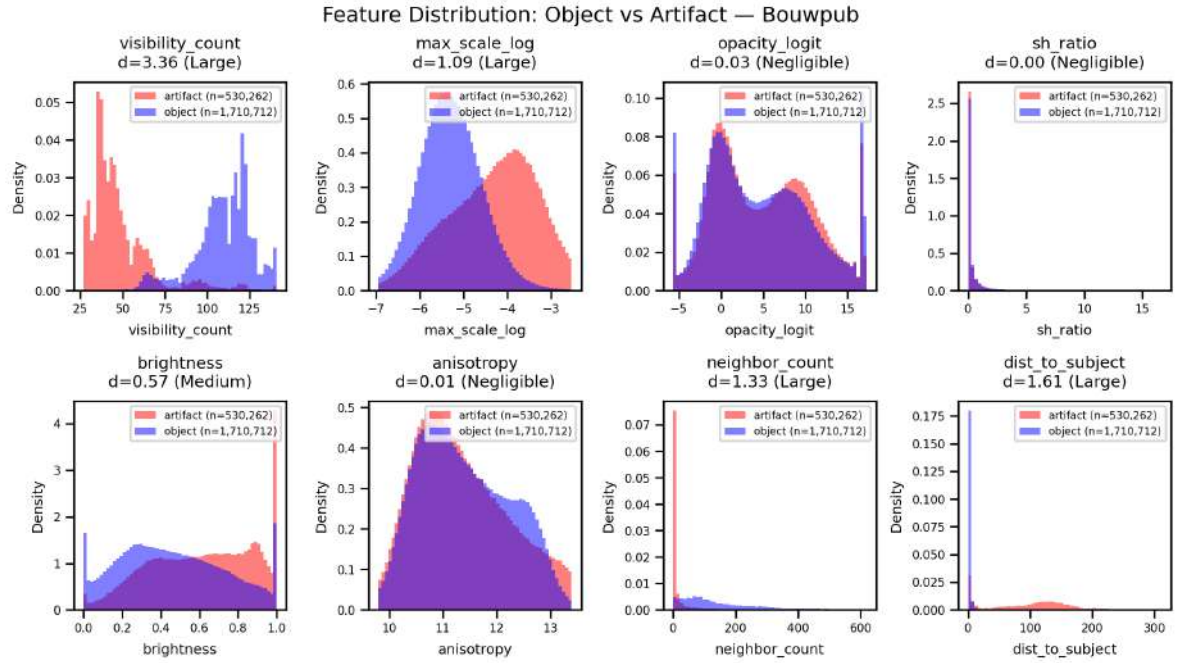


Figure 2.4: Feature distribution of primary object and background Gaussians in the Bouwpub scene (see Figure 2.5 left column for a visual reference of the scene)

2.4. Capture Intent in Multi-view Reconstruction

As mentioned in 2.2, visual saliency can be broadly categorised into two mechanisms: bottom-up, driven by low-level features such as colour contrast and edge density, and top-down attention, driven by the observer's internal goals and intentions (Connor et al., 2004). Bottom-up saliency is modelled using computational approaches, early methods such as Itti-Koch extract features including colour, intensity, and orientation to identify locally distinctive regions (Itti et al., 1998), Harel et al. (Harel et al., 2006) propose Graph-Based Visual Saliency (GBVS), which treats the image as a Markov chain where edge weights reflect feature dissimilarity between image patches, saliency emerges from the stationary distribution of a random walk over this graph, capturing global distinctiveness rather than purely local contrast. Deep learning methods such as U^2-Net (X. Qin et al., 2020) learn to predict salient regions from datasets annotated by human labellers, capturing patterns of human visual attention without requiring task-specific supervision at inference time. Top-down saliency is usually measured using

descriptive and qualitative method by using eye-tracking machines or behavioural studies (Zhu et al., 2014).

While top-down saliency has traditionally been studied through controlled behavioural experiments, it also manifests in everyday photography. The photographer's capture intent is a form of top-down attention: when composing a scene, the photographer makes deliberate, goal-driven decisions about which subject to focus on, resulting in disproportionately dense viewpoint coverage of the primary subject in the training image collection. When these photographs are used to train a 3DGS model, the optimisation process transforms this top-down behavioural signal into bottom-up geometric statistics that are directly measurable from the trained model. Regions with dense viewpoint coverage receive more gradient updates during optimisation, and the adaptive density control mechanism densifies under-reconstructed regions that are consistently observed across views. As a consequence, Gaussian primitives corresponding to the primary subject differ from background primitives in their geometric properties, which can be directly calculated from the trained model without relying on external supervision.

This observation suggests that multi-view reconstruction acts as a passive aggregator of top-down attention signals, converting the photographer's focus into measurable bottom-up geometric statistics. Moreover, unlike 2D saliency methods that are susceptible to single-frame illumination bias and occlusion, these 3D statistics reflect the photographer's intent across all training viewpoints, making them more robust to per-frame ambiguities.

To our knowledge, no existing work has systematically investigated whether bottom-up geometric statistics implicitly encode top-down visual saliency, nor whether such saliency is recoverable without external supervision. This gap motivates the approach presented in Chapter 3.

2.5. Scene Understanding in 3D Gaussian Splatting

3D Gaussian Splatting has demonstrated remarkable performance in novel view synthesis. However, many downstream applications in robotics and geomatics require not only photorealistic rendering but also an understanding of what is present in the scene (Bao et al., 2025; Wu et al., 2024). Scene understanding tasks encompass a broad range of topics, include object segmentation, object editing and salient object detection.

Existing approaches to scene understanding in 3DGS predominantly rely on 2D foundation models as an external source of supervision. SAGA (Cen et al., 2023) and LangSplat (M. Qin et al., 2023) both explicitly embed learnable semantic features directly onto Gaussians. SAGA targets real-time, interactive 3D segmentation by distilling the capabilities of the Segment Anything Model (SAM). It attaches a learnable affinity feature to each Gaussian and optimises via contrastive learning against 2D SAM masks. To address multi-granularity ambiguity, SAGA introduces a scale-gated mechanism that dynamically adjusts features based on observation distance. Similarly, LangSplat constructs a 3D language field by using a scene-wise autoencoder to compress high-dimensional CLIP embeddings into compact latent vectors attached to Gaussians, enabling efficient open-vocabulary querying directly within 3D space. More recently, GSsplat (Xiao et al., 2025) extends this direction by proposing a generalizable semantic Gaussian Splatting framework that predicts semantic Gaussian parameters directly from multi-view inputs without scene-specific retraining.

Among the scene understanding tasks listed above, salient object detection: the task of identifying perceptually important regions that attract human visual attention is particularly relevant for applications such as foreground extraction, scene compression, and content-aware rendering. Given the success of 2D saliency detection methods in identifying visually prominent regions from single images, a natural extension would be to use these 2D saliency maps (such as U^2-Net) as supervision to identify perceptually important regions in 3DGS. However, this approach faces a fundamental challenge: 2D saliency is inherently view-dependent. The same 3D region may appear salient in one viewpoint but be occluded, poorly lit, or outside the frame in another. Aggregating inconsistent per-frame saliency signals into a stable 3D representation is therefore an ill-posed problem. As shown in Figure 2.5, While the building is the primary subject of the scene, U^2-Net identifies the parked cars as the salient object in the second viewpoint, part of the rooftop and dustbins as salient objects in the third viewpoint, demonstrating that 2D saliency predictions vary inconsistently across viewpoints.

These limitations suggest that injecting external semantic supervision into 3DGS is constrained by the quality of 2D signals. This motivates an alternative direction: rather than imposing external labels onto Gaussian primitives, whether the intrinsic properties of the trained model already encode sufficient information for scene understanding, which will be further invesitigated in Chapter 3

(a) Viewpoint 1 original image

(b) Viewpoint 1 U^2-Net prediction

(c) Viewpoint 2 original image

(d) Viewpoint 2 U^2-Net prediction

(e) Viewpoint 3 original image

(f) Viewpoint 3 U^2-Net prediction

Figure 2.5: Illustration of the view-dependency problem in 2D saliency detection on the Bouwpub scene. Left column shows training images from three different viewpoints. Right column shows the corresponding U^2-Net predictions.

3

Scientific Article

Recovering Visaul Saliency from Intrinsic Properties of 3D Gaussian Splatting

Xinya Bi

Delft University of Technology

3.1. Abstract

3D Gaussian Splatting (3DGS) represents scenes as collections of Gaussian primitives whose attributes are shaped by multi-view photographic supervision. This raises a natural question: does the photographer's visual focus leave a measurable imprint on these intrinsic properties? While prior work has explored segmentation and scene decomposition in 3DGS, no existing method has investigated whether Gaussian attributes alone encode visual saliency. We propose a mask-free, post-hoc classifier that recovers the photographer's region of interest from Gaussian attributes, requiring neither the original training images nor any 2D foundation model. Trained on scenes from Tanks and Temples and MipNeRF360, our method achieves a mean LOOCV F1 of 0.957 and generalizes to unseen scenes with a mean test F1 of 0.929. Projected 3D saliency masks show strong alignment with U^2-Net predictions on original training images, confirming that multi-view Gaussian intrinsic properties capture a geometrically consistent, view-stable notion of saliency that single-frame 2D methods cannot provide. These properties make our method applicable to automatic foreground extraction, capture intent analysis, and perceptual quality-driven compression for bandwidth-efficient streaming.

3.2. Introduction

Novel view synthesis aims to render new views from a given set of input images and their respective camera poses (Mildenhall et al., 2020). It has evolved from implicit neural representations such as NeRF (Mildenhall et al., 2020), where scene information is encoded in the weights of a multilayer perceptron, to explicit representations such as 3D Gaussian Splatting (Kerbl et al., 2023), where scenes are modelled as discrete Gaussian primitives with interpretable attributes, such as position, shape and appearance. Unlike implicit representations, these attributes can be directly read, filtered, and analyzed without inverting the network. Since they are determined by multi-view image supervision during training, their values reflect not only the geometry and appearance of the scene, but also which parts of the scene were consistently observed across viewpoints, and by extension, where the photographer chose to focus.

Despite this potential, existing approaches to scene understanding in 3DGS do not exploit these intrinsic properties directly. They instead inject external semantic supervision into the representation by distilling features from 2D foundation models such as SAM (Kirillov et al., 2023) or CLIP (Radford et al., 2021), enabling open-vocabulary querying (M. Qin et al., 2023) and instance segmentation(Ye et al., 2023). While effective, these methods require either the original training images, a pre-trained 2D model, or modifications to the training pipeline, and are designed to address a different question: what semantic category each region belongs to. A complementary dimension of scene understanding is visual saliency, which means identifying which objects or regions attract human attention. It has direct applications in foreground extraction and perceptual scene compression. This problem has received comparatively little attention in 3DGS. No existing method has asked whether the intrinsic attributes of Gaussian primitives alone are sufficient to recover visual saliency, without any external supervision or modifications to the reconstruction pipeline.

We investigate this question and propose a mask-free, post-hoc approach that recovers the photographer's region of interest using only the intrinsic attributes of a trained 3DGS model. Rather than injecting external supervision, we treat Gaussian attributes as indicators of capture intent and show that they can be used to recover 3D visual saliency directly. The resulting saliency representation requires no training images or 2D segmentation models at inference time, and can be directly applied to tasks such as foreground extraction. The main contributions of this work are as follows:

- We demonstrate that the intrinsic properties of 3DGS primitives serve as meaningful indicators of 3D visual saliency.

- We show that projecting the recovered 3D saliency onto 2D image planes produces masks that align with human visual perception as measured by U^2-Net saliency predictions.
- We demonstrate that 3D saliency derived from Gaussian properties generalizes to unseen scenes and can be applied to downstream tasks such as foreground extraction.

3.3. Related Works

This section reviews prior work on 3D Gaussian Splatting, Gaussian primitive analysis, scene understanding in 3DGS, and 2D visual saliency detection, highlighting the gap that motivates our approach.

3D Gaussian Splatting 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) represents scenes as a collection of anisotropic 3D Gaussian primitives, each defined by a mean position μ, covariance $\Sigma = RSS^TR^T$, R and S are the rotation and scale matrices, opacity α, and spherical harmonics (SH) for view-dependent appearance. During rendering, Gaussians are projected onto the image plane and composited via alpha blending, where the contribution of each primitive is weighted by its opacity and the accumulated transmittance of all primitives in front of it. Parameters are optimised via differentiable rasterization against training views, interleaved with adaptive density control that densifies under-reconstructed regions and prunes low opacity primitives.
Original 3DGS relies on appearance-driven optimisation and lacks explicit geometric constraints, which can lead to floaters and inaccurate surface reconstruction. Several methods solve this issue by introducing geometry-aware regularisation, such as SuGAR (Guédon & Lepetit, 2024), which incorporated regularised terms that encouraged Gaussians to align with scene surfaces, while PGSR (Chen et al., 2025) introduced unbiased depth rendering for improved geometric accuracy.
In this work, we build upon PGSR as our reconstruction backbone, as its depth regularization helps to produce cleaner 3DGS models that are more amenable to our analysis.

Gaussian primitive analysis Prior works have investigated the intrinsic properties of Gaussian primitives mainly for pruning and compression. The core idea shared across these methods is to estimate an importance score for each Gaussian based on its contribution to rendering quality, then discard low-importance primitives. LightGaussian (Z. Fan et al., 2023) uses opacity O_i and normalized Gaussian volume γ_i to formulate the global significance score then remove the Gaussians that are below certain threshold. TrimGS (L. Fan et al., 2024) uses alpha-blending weights as the importance score measurement. TIDI-GS (Yang et al., 2026) extends this idea in indoor scenes by leveraging consistency across multi-views and spatial isolation.
While these methods demonstrate that Gaussian attributes carry meaningful signals about primitive importance, they uniformly define importance as rendering fidelity. We instead ask whether these same attributes encode different information: the photographer's visual saliency.

Scene Understanding in 3DGS Existing methods for scene understanding in 3DGS primarily rely on 2D semantic priors and can be categorized into two paradigms: implicit semantic fields and explicit semantic binding. FMGS (Zuo et al., 2024) distills semantic features from 2D foundation models into multi- resolution hash encodings (MHE) to achieve open vocabulary object detection. In contrast to implicit semantic field, SAGA (Cen et al., 2023) adopt an explicit binding strategy by distilling the capabilities of the Segment Anything Model (SAM), then attach a learnable affinity feature to each Gaussian, which enables interactive object segmentation in 3DGS. GaussianCut (Gilitschenski et al., 2024) feed user annotations to 2D segmentation models to generate coarse mask for foreground object isolation.
Despite their effectiveness, these methods share a common limitation: they require either the original training images, 2D foundational model or modifications to the training process. In contrast, our method operates purely on the intrinsic properties of a trained 3DGS model.

Visual Saliency Detection Visual saliency detection aims to identify regions in an image that attract human visual attention. Some of its applications include background removal, image and video compression, context-aware rendering. Early methods relied on features such as color, intensity and orientation (Itti et al., 1998). Deep neural networks have advanced the field: BASNet (X. Qin et al., 2021) introduces a boundary-aware salient object detection network that produces sharp foreground boundaries, while U2Net (X. Qin et al., 2020) proposes

a two-level nested U-structure that captures both local and global contextual information for accurate saliency prediction. These methods are trained on datasets such as DUTS (Wang et al., 2017), where ground truth masks are annotated by human labelers to reflect visual attention.

However, 2D saliency detection operates on individual frames and cannot capture the photographer's intent completely. A salient object in a single frame may be a transient object that is absent in most viewpoints. The true object of interest is only revealed through multi-view consistency. We hypothesize that this capture intent is implicitly encoded in the intrinsic properties of trained 3DGS models, and propose to recover it without any 2D supervision.

3.4. Method

Our method consists of three stages: feature extraction from Gaussian primitives, classifier training on annotated scenes, and inference on unseen scenes to recover the photographer's region of visual interest. Figure 3.1 illustrates the overall pipeline.

3.4.1. Gaussian Properties as Saliency Indicators

The properties of a Gaussian primitive encode information about how the scene was observed during training, which motivates their use as saliency indicators. Specifically, we identify three key observations:

(a) **Multi-view visibility reflects photographer's focus:** Gaussians corresponding to the photographer's focus receive consistent supervision across training views, resulting in higher visibility counts. In contrast, background regions are observed from fewer viewpoints, yielding lower visibility.

(b) **Local density reflects geometric coherence:** The primary subject typically forms a geometrically coherent surface with high local point density, while background artifacts exhibit sparse neighborhoods.

(c) **Scale and anisotropy reflect surface structure:** Gaussians in well-reconstructed regions tend to be compact and planar, while background Gaussians are often large and needle-like with high anisotropy.

These observations motivate the construction of a feature representation that embeds intrinsic Gaussian properties and local neighborhood information. The detailed feature vector construction is described in Section 3.4.2.

3.4.2. Feature Extraction

We compute an 8-dimensional raw feature vector $f_i^{\text{raw}} = [V_i,\, S_{\max},\, O_i,\, SH_i,\, Br_i,\, An_i,\, \text{Neighbor}_i,\, \text{Dist}_i]$ for each Gaussian primitive i, grouped into four categories: visibility, geometry, spatial, and appearance. Each feature is defined as follows.

Feature Computation

(a) **Visibility: Visibility count (V_i)** measures the number of training views in which a Gaussian projects onto the image plane with a non-zero screen-space radius, serving as a proxy for multi-view coverage:

$$v_i = \sum_{k=1}^{M} \mathbf{1}[r_i^k > 0] \tag{3.1}$$

where M is the total number of training views and r_i^k is the screen-space radius of Gaussian i in view k.

(b) **Geometry: Maximum scale ($S_{\max}$)** captures the spatial extent of a Gaussian along its dominant axis:

$$s_i^{\max} = \max(s_i^0, s_i^1, s_i^2) \tag{3.2}$$

where s_i^0, s_i^1, s_i^2 are the log-scale parameters of Gaussian i.

Anisotropy (An_i) quantifies the elongation of a Gaussian by comparing its largest and smallest axes:

$$a_i = \log_{10}\left(\frac{\exp(s_i^{\max})}{\exp(s_i^{\min}) + \epsilon}\right) \tag{3.3}$$

where ϵ is a small constant for numerical stability. Background Gaussians tend to exhibit large anisotropy values, corresponding to needle-like or flat shapes.

(c) **Spatial: Neighbor count (Neighbor_i)** measures local point density by counting the number of Gaussians within a radius r:

$$n_i = |\{j \neq i : \|p_i - p_j\|_2 \leq r\}| \tag{3.4}$$

where $p_i \in \mathbb{R}^3$ is the position of Gaussian i, and the radius is set adaptively as $r = 5 \cdot \text{median}(\exp(s^{\max}))$.

Distance to subject (Dist_i) measures the spatial distance from each Gaussian to

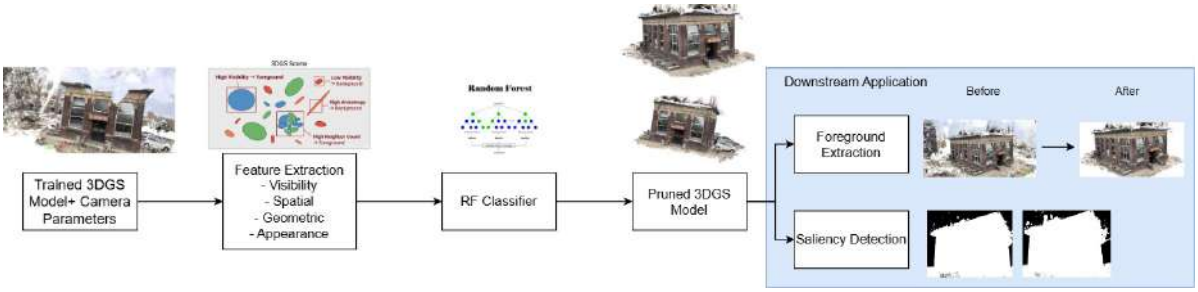


Figure 3.1: Overview of the proposed pipeline

the high-density region, normalized by the adaptive search radius r to ensure scale invariance:

$$\tilde{d}_i = \frac{\min_{j \in \mathcal{S}} \|p_i - p_j\|_2}{r} \tag{3.5}$$

where $\mathcal{S} = \{j : n_j > \text{median}(\mathbf{n})\}$ denotes the set of Gaussians with above-median neighbor counts, and r is the adaptive search radius shared with neighbor count.

(d) **Appearance: Brightness (Br_i)** is computed from the zeroth-order spherical harmonic (DC) coefficients following the ITU-R BT.709 standard:

$$b_i = 0.2126\, R_i + 0.7152\, G_i + 0.0722\, B_i \tag{3.6}$$

where $(R_i, G_i, B_i) = \text{clip}(\mathbf{f}_i^{\text{dc}} \cdot C_0 + 0.5,\ 0,\ 1)$ and $C_0 = 0.2821$ is the zeroth-order SH coefficient.

SH complexity ratio (SH_i) measures the relative energy of higher-order spherical harmonics compared to the DC component, reflecting view-dependent color variation:

$$h_i = \frac{\sum_{l>0} \|\mathbf{f}_i^{\text{rest}}\|^2}{\|\mathbf{f}_i^{\text{dc}}\|^2 + \epsilon} \tag{3.7}$$

Opacity logit (O_i) is the raw opacity parameter before sigmoid activation:

$$o_i = \text{logit}(\alpha_i) = \log \frac{\alpha_i}{1 - \alpha_i} \tag{3.8}$$

where $\alpha_i \in (0, 1)$ is the opacity of Gaussian i.

Feature Normalization All features are standardized using a standard scaler. For each feature dimension, the normalized value is:

$$\hat{f}_i = \frac{f_i - \mu}{\sigma} \tag{3.9}$$

where μ and σ are the mean and standard deviation computed across all Gaussians in the scene.

Feature Analysis and Classifier Motivation Figure 3.2 illustrates the distribution of each feature for primary object and background Gaussians in one of the scenes. While features such as visibility count, neighbour count, and distance to subject show clear separation between the two classes, no single feature provides a clean linear decision boundary sufficient for classification.

3.4.3. Machine Learning Classifier

The features described in Section 3.4.2 capture diverse cues of Gaussian primitives, but their interactions are complex and difficult to model with hand-crafted rules. We therefore train a supervised classifier to learn which Gaussian attributes are predictive of visual saliency directly from annotated data.

Model Selection We adopt Random Forest (RF) as our classifier for three reasons: it handles small training sets without overfitting, it provides feature importance scores that offer interpretable insights into which Gaussian properties are most discriminative for capturing saliency, and it is computationally efficient at inference time.

Training Protocol Ground truth foreground labels are obtained through semi-automatic annotation. We first estimate a coarse bounding box around the primary subject in 3D space to provide an initial selection, followed by manual refinement in SuperSplat to further remove floaters and correct boundary errors. Foreground is consistently defined as geometrically coherent surfaces of the primary subject, excluding distant background and floaters.

Our dataset comprises 16 fully annotated scenes, split into 12 training scenes, 1 independent validation scene, and 3 held-out test scenes. Class imbalance is addressed using balanced class weights. To evaluate cross-

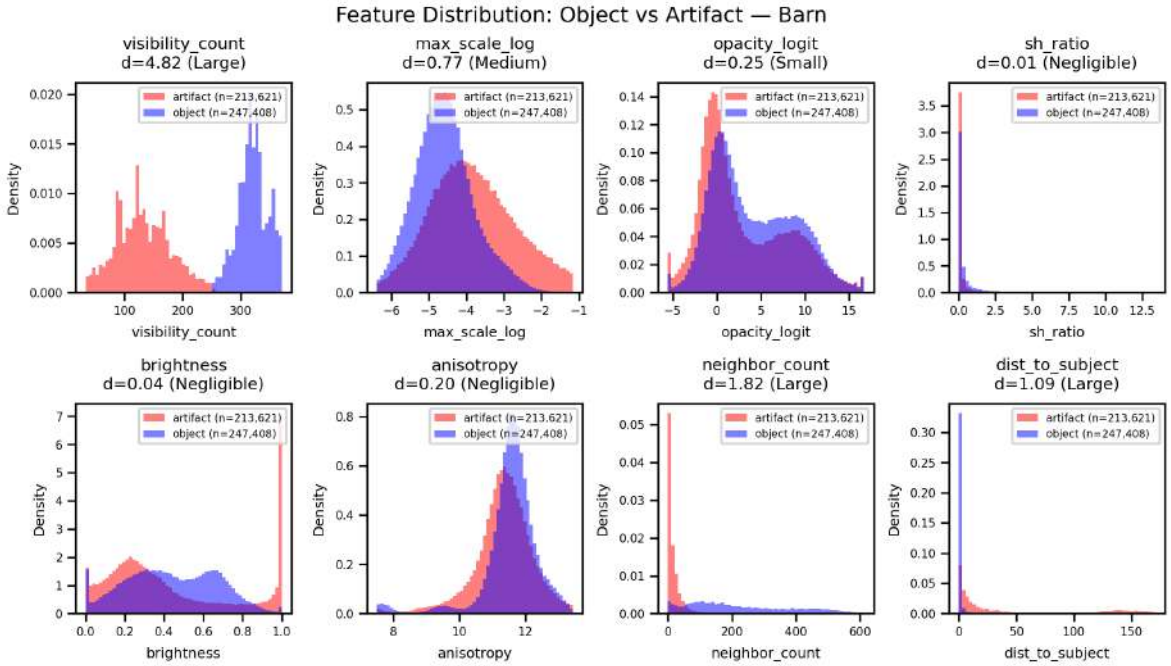


Figure 3.2: Feature distribution of foreground and background Gaussians in the Barn scene. While features such as visibility count, neighbour count, and distance to subject show clear separation between the two classes, no single feature provides a clean linear decision boundary sufficient for classification.

scene generalization, we perform leave-one-out cross-validation (LOOCV) within the 12 training scenes. The validation scene is subsequently used to tune and freeze the optimal hyperparameters. Finally, the model is trained on all 13 training and validation scenes combined, and evaluated on the 3 held-out test scenes to report the final F_1-score and qualitative results.

Inference At inference time, only the trained `.ply` file and camera parameters are required. The classifier predicts a per-Gaussian foreground probability, which is thresholded at 0.5, corresponding to the default decision boundary of the Random Forest posterior probability and not tuned on the test set. No segmentation masks or retraining are needed.

3.5. Experiments

3.5.1. Experiment Setup

Datasets We evaluate our method on three datasets spanning 16 diverse scenes in total. Tanks and Temples (Knapitsch et al., 2017) contains large-scale outdoor environments; we utilize 8 scenes from this dataset, including *Barn*, *Caterpillar*, and *Ignatius*. MipNeRF360 (Barron et al., 2021) spans both unbounded indoor and outdoor scenes at varying scales, from which we select 4 scenes (e.g. *Bicycle*, *Bonsai*, and *Flowers*). Finally, our self-captured dataset consists of two scenes captured around landmarks in Delft, Netherlands (*House1* and *Bouwpub*), demonstrating applicability to casually captured real-world data.

For the experimental split, 12 scenes from Tanks and Temples and MipNeRF360 comprise the training set, while *Lighthouse* is selected as the independent validation scene. The remaining 3 scenes (the 2 self-captured scenes along with *Courthouse* from Tanks and Temples) are held out as the independent test set to report our final quantitative and qualitative evaluation.

Implementation Details All 3DGS models are reconstructed using PGSR (Chen et al., 2025) with 30,000 training iterations. Other hyperparameters follow the default PGSR settings. All training is conducted on an NVIDIA A40 GPU. For saliency alignment evaluation, we use the pretrained U^2-Net (X. Qin et al., 2020) to generate saliency maps on the original training images.

Evaluation Criteria We evaluate our method along three dimensions corresponding to our research questions. Feature discriminability is assessed using Cohen's d and random forest feature importance, which measure the standardized mean difference between object and

artifact distributions for each feature. 3D–2D saliency alignment is measured by projecting the retained Gaussians onto each training viewpoint and comparing the resulting binary mask against U^2-Net saliency predictions using pixel-level IoU, precision, and recall. Foreground extraction is evaluated at the Gaussian point level using precision, recall, and F1-score, where a Gaussian is considered a true positive if it is correctly classified as foreground. To assess cross-scene generalization, we report leave-one-out cross-validation (LOOCV) F1 across all training scenes and held-out test F1 on unseen scenes.

3.5.2. Feature Discriminability

To identify which intrinsic properties of 3DGS serve as meaningful saliency indicators, we employ two complementary metrics. First, Cohen's d measures the standardized mean difference between object and artifact distributions for each feature:

$$d = \frac{|\mu_{\text{obj}} - \mu_{\text{art}}|}{\sigma_{\text{pooled}}} \tag{3.10}$$

where $d > 0.8$, $0.5 < d \leq 0.8$, $0.2 < d \leq 0.5$, and $d \leq 0.2$ indicate large, medium, small, and negligible separation respectively. This metric is computed per scene and averaged across all scenes to assess cross-scene consistency.

Table 3.1 reveals a clear hierarchy of feature discriminability. Visibility achieves large separation in all 16 scenes, confirming that multi-view coverage is the most reliable indicator of capture intent. Spatial features (neighbor count and distance to subject) also show consistently large separation in 14 and 13 out of 16 scenes respectively, reflecting that the photographer's focus tends to coincide with geometrically coherent, high-density regions. Geometry features (maximum scale and anisotropy) show moderate discriminability, with 9/16 and 1/16 large separations respectively, as background Gaussians tend to be larger and artifacts tend to be needle-like. Appearance features (opacity, brightness, and SH ratio) contribute minimally across most scenes, suggesting that photometric properties are not reliable indicators of capture intent.

Second, we report the random forest feature importance, which reflects each feature's contribution to classification decisions across all training scenes. As shown in Figure 3.3, visibility and spatial features (visibility count, neighbor count, and distance to subject) account for the majority of the importance at 0.8349. Geometry-related features (maximum scale and anisotropy) contribute 0.1033, while appearance features (brightness, SH complexity ratio, and opacity logit) contribute 0.0618. The Cohen's d results and feature importance scores are consistent with each other, and jointly guide the design of our ablation study in Section 3.5.5.

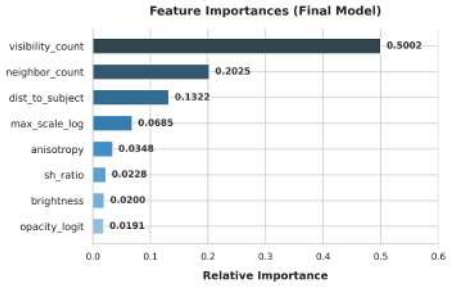


Figure 3.3: Feature importances of the final model.

3.5.3. Saliency Alignment Evaluation

To validate our hypothesis, we compare our projected 3D saliency masks against U^2-Net (X. Qin et al., 2020) predictions as a 2D reference, acknowledging that this proxy is reliable when the primary subject is visually dominant but may fail under occlusion or complex spatial layouts.

To project the 3D saliency onto 2D image planes, we render the pruned Gaussian model from each training viewpoint using the differentiable Gaussian rasterizer. All preserved Gaussians are assigned a white color ($c_i = 1$), with a black background ($C_{bg} = 0$). For each pixel x, the rendered value is computed via front-to-back alpha compositing:

$$C(x) = \sum_{i=1}^{N} \alpha_i(x) \cdot T_i(x) \tag{3.11}$$

$$T_i(x) = \prod_{j=1}^{i-1} (1 - \alpha_j(x)) \tag{3.12}$$

where $\alpha_i(x)$ is the effective alpha of the i-th Gaussian at pixel x, combining its learned opacity o_i and its 2D projected extent. The resulting grayscale image is thresholded at $\tau = 0.5$ to produce a binary 2D saliency mask M_{3D}, which is then compared against the U^2-Net prediction M_{2D} for the same viewpoint using pixel-level IoU, precision, and recall.

The qualitative results are shown in Figure 3.4. Our projected 3D saliency masks M_{3D} generally maintain strong alignment with the U^2-Net predictions M_{2D}, demonstrating that the proposed feature set successfully captures the primary subject areas. As shown in Figure 3.4a, while

Table 3.1: Cohen's d effect size for each feature across all annotated scenes. **L** = Large ($d > 0.8$), **M** = Medium ($0.5 < d \leq 0.8$), **S** = Small ($0.2 < d \leq 0.5$), **-** = Negligible ($d \leq 0.2$).

Scene	Vis.	Neighbor	Dist.	Max Scale	Anisotropy	Opacity	Brightness	SH Ratio
Barn	**L**	**L**	**L**	M	-	S	-	-
Caterpillar	**L**	**L**	**L**	**L**	S	-	-	-
Truck	**L**	**L**	**L**	S	S	S	S	-
Ignatius	**L**	**L**	**L**	S	-	-	S	-
Family	**L**	**L**	**L**	S	-	-	S	-
Francis	**L**	**L**	**L**	**L**	S	-	-	-
Horse	**L**	**L**	**L**	M	S	-	-	-
Train	**L**	**L**	**L**	**L**	S	-	S	-
Bonsai	**L**	**L**	**L**	**L**	-	S	S	-
Kitchen	**L**	**L**	**L**	**L**	-	-	M	-
Stump	**L**	**L**	**L**	**L**	S	-	S	-
Flowers	**L**	**L**	**L**	M	**L**	S	-	-
Lighthouse	**L**	**L**	M	**L**	S	-	M	-
Courthouse	**L**	S	-	S	S	-	S	-
Bouwpub	**L**	**L**	**L**	**L**	S	-	S	-
House1	**L**	**L**	**L**	**L**	S	-	S	-
# Large	**16/16**	**14/16**	**13/16**	9/16	1/16	0/16	0/16	0/16

the core regions overlap significantly, the projected 3D mask exhibits a less smooth boundary around the rooftop compared to the U^2-Net predictions. This is a natural consequence of the 3D representation: unlike 2D image pixels that form continuous boundaries through spatial convolutions, Gaussian primitives have physical sizes and shapes, and their projection via alpha compositing produces a rougher, non-linear silhouette at object boundaries.

Courthouse (Figure 3.4b) illustrates a case where 3D saliency outperforms 2D prediction. Local illumination contrast causes U^2-Net to produce disconnected holes in the building facade, while our multi-view aggregation maintains a complete and continuous mask. These comparisons show that our 3D saliency is not a replication of 2D labels, but a geometrically consistent representation that complements single-view 2D predictions.

Ignatius (Figure 3.4c and Bonsai (Figure 3.4d) represent cases where the 3D saliency masks show strong agreement with U^2 -Net predictions. In both scenes, the primary subject is clearly separated from the background, demonstrating consistent saliency estimates across both 3D and 2D methods.

Table 3.2 reports the quantitative alignment between M_{3D} and M_{2D} across all scenes. The high mean recall of 0.9474 indicates that our 3D saliency maps successfully capture most regions deemed salient by U^2-Net. The relatively lower precision reflects a fundamental difference in scope: U^2-Net predicts saliency from a single frame's local contrast, while our method aggregates information across all training viewpoints, resulting in a broader but geometrically consistent foreground region. This is not a failure mode; it reflects that 3D saliency captures the complete subject structure, while 2D saliency responds to single-frame visual prominence.

Scene-level variation in IoU reflects the clarity of capture intent. Scenes with a single dominant subject clearly separated from the background (e.g., House1, IoU=0.7214; Bonsai, IoU=0.5582) achieve strong alignment. Stump yields the lowest IoU (0.1065) and recall (0.7181), as the scene contains multiple visually and geometrically similar elements: the primary stump is surrounded by other stumps and vegetation with comparable appearance and spatial distribution, making it difficult for both methods to agree on a single salient region.

Beyond quantitative alignment, the comparison also reveals a fundamental limitation of 2D saliency detection. As shown in Figure 3.5, U^2-Net is biased by local contrast, focusing on the foreground picnic table while missing the main building behind it. In contrast, our method leverages multi-view consistency to successfully capture the complete 3D structure of the primary building. This demonstrates that 3D saliency derived from Gaussian properties provides a more

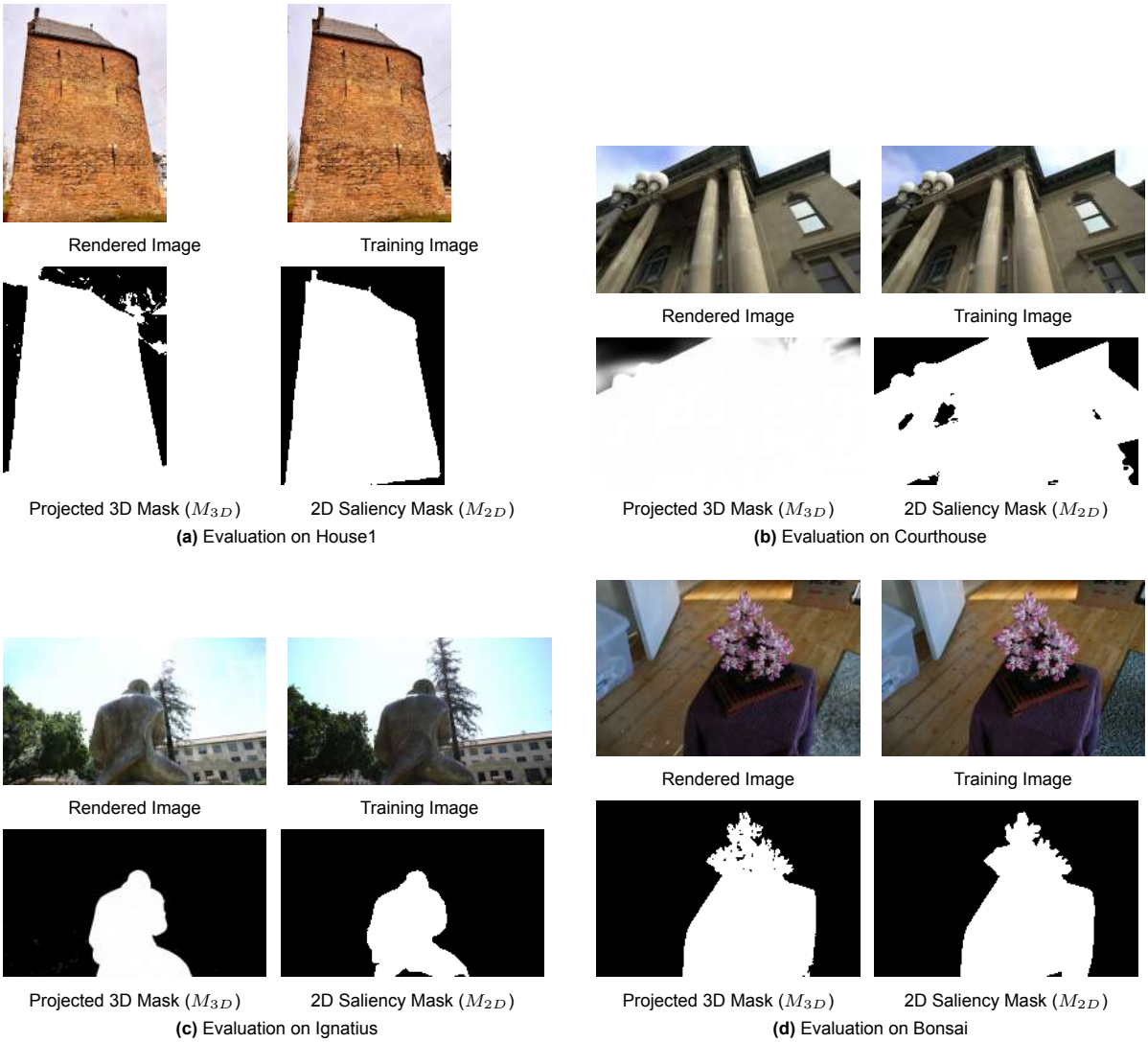


Figure 3.4: Qualitative comparison of saliency alignment across four representative scenes. Within each subfigure, the layout displays: the 3DGS rendered image from our pruned model (top-left), the original input training image (top-right), our projected 3D saliency mask M_{3D} (bottom-left), and the corresponding U^2-Net 2D saliency mask prediction M_{2D} (bottom-right).

complete and view-consistent representation of capture intent than single-frame 2D saliency detection.

3.5.4. Foreground Extraction

The leave-one-out cross-validation (LOOCV) results demonstrate strong cross-scene generalization, achieving a mean F1 of 95.7% across 12 scenes (Table 3.3). The low standard deviation (3.5%) indicates that the classifier generalizes consistently rather than overfitting to specific scene characteristics. Bonsai achieves the highest F1 (99.7%), benefiting from a clearly defined subject with strong background contrast. Train yields the lowest F1 (86.9%) with notably low recall (78.7%), which we attribute to two factors: the metallic surface introduces strong view-dependent appearance variation that destabilizes appearance-based features, while the surrounding environment exhibits geometric and spatial statistics similar to the foreground subject, reducing feature discriminability.

Table 3.4 reports performance on three held-out test scenes unseen during training, achieving a mean F1 of 92.9%. House1 achieves the highest F1 (94.1%) with balanced precision and recall, reflecting clear foreground-background separation in this architectural scene. Bouwpub achieves near-perfect precision (99.5%) but lower recall (84.3%), suggesting that the classifier correctly identifies foreground when confident but misses some boundary regions. Courthouse, despite its complex architectural structure spanning a large area, still achieves a competitive F1 of 93.5%,

(a) Training image **(b)** 2D Saliency Mask (M_{2D}) **(c)** Projected 3D Mask (M_{3D})

Figure 3.5: Comparison of 2D and 3D saliency under single-view ambiguity. U^2-Net is biased by local contrast, focusing on the foreground picnic table while missing the main building. Our method leverages multi-view consistency and successfully captures the complete 3D structure of the primary building.

Table 3.2: Per-scene IoU, Precision, and Recall between projected 3D saliency mask M_{3D} and U^2-Net prediction M_{2D} (mean over all training viewpoints).

Scene	IoU	Precision	Recall
Barn	0.4822	0.4848	0.9851
Bonsai	0.5582	0.5819	0.8972
Bouwpub	0.3608	0.3654	0.9689
Caterpillar	0.3463	0.3477	0.9900
Courthouse	0.2991	0.3022	0.9541
Family	0.4405	0.4430	0.9887
Francis	0.4373	0.4395	0.9871
Horse	0.3787	0.3862	0.9546
House1	0.7214	0.7242	0.9871
Ignatius	0.2800	0.2809	0.9896
Kitchen	0.1724	0.1730	0.9873
Train	0.3905	0.3962	0.9057
Truck	0.3315	0.3342	0.9803
Flowers	0.4643	0.4866	0.9174
Stump	0.1065	0.1079	0.7181
Average	0.3847	0.3902	0.9474

Table 3.3: Leave-One-Out Cross-Validation (LOOCV) results across all 12 training scenes. Each row reports the performance when that scene is held out as the test scene. Metrics (%) are computed on the *object* class. **Bold** denotes the best value per column.

Scene	F1 (%)	Precision (%)	Recall (%)
Barn	98.0	99.4	96.6
Caterpillar	98.1	96.8	99.4
Truck	97.6	95.4	**100.0**
Ignatius	95.6	91.6	**100.0**
Family	93.5	87.9	**100.0**
Francis	94.9	90.7	99.5
Horse	97.5	96.8	98.2
Train	86.9	97.0	78.7
Bonsai	**99.7**	**100.0**	99.5
Kitchen	90.8	84.3	98.3
Stump	98.1	97.5	98.8
Flowers	97.5	95.2	99.9
Mean	95.7	94.4	97.4
Std	3.5	4.6	5.7

consistent with its large Cohen's d for visibility count, which provides strong discriminative signal even in complex scenes.

Table 3.4: Test-set performance on three held-out scenes unseen during training or LOOCV. Metrics (%) are computed on the *object* class. **Bold** denotes the best value per column.

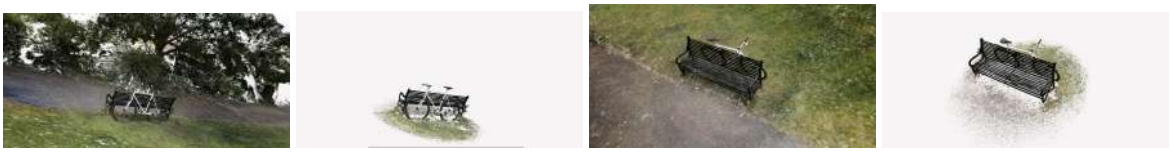
Scene	F1 (%)	Precision (%)	Recall (%)
Courthouse	93.5	93.9	93.1
Bouwpub	91.2	**99.5**	84.3
House1	**94.1**	93.0	**95.2**
Mean	92.9	95.5	90.9

The qualitative results in Figure 3.6 further illustrate the classifier's ability to isolate the primary subject across diverse scene types, including outdoor scenes (Figures 3.6a and 3.6c) and an indoor scene (Figure 3.6b). Our method accurately preserves fine details of the target objects, and the extracted foregrounds exhibit strong multi-view consistency across viewpoints.

3.5.5. Ablation Study

To evaluate the contribution of each feature group, we conduct an incremental ablation study across three configurations. The three configurations are designed to reflect the four feature categories identified in Section 3.4.2, grouped by their expected discriminability as suggested by the Cohen's d analysis: visibility and spatial features form the base as the strongest indicators, geometry and basic appearance are added next as moderate contributors, and high-order appearance features are added last as the weakest group.

- **Base (3 features):** Visibility count, neighbor count, and distance to subject.
- **+ Geometry & Brightness (6 features):** Extends the base set with maximum scale, anisotropy, and brightness.



(a) Bicycle (MipNeRF360)

(b) Bonsai (MipNeRF360)

(c) Barn (Tanks and Temples)

Figure 3.6: Qualitative foreground extraction results across three diverse scenes. For each scene, we show the original 3DGS point cloud and the extracted foreground from two viewpoints. Our method successfully isolates the primary subject across both indoor and outdoor scenes.

- **+ High-order Appearance (8 features, Full Model):** Further adds SH complexity ratio and opacity logit.

The results are summarized in Table 3.5. The base setting already yields a mean F1 of 0.899, confirming that visibility and spatial features provide a solid foundation. Adding geometry and brightness boosts the mean F1 to 0.924, indicating that geometric structure is important for separating the primary subject from background artifacts. The full 8-feature model achieves the highest mean F1 of 0.930, demonstrating that SH complexity ratio and opacity logit offer complementary information that further improves performance despite their modest individual contribution.

Table 3.5: Ablation study on feature set size. Results report F1-score on the three test scenes for each feature configuration.

Scene	3 features	6 features	8 features
Courthouse	0.903	0.928	**0.935**
Bouwpub	0.857	0.906	**0.912**
House1	0.937	0.939	**0.941**
Mean F1	0.899	0.924	**0.930**

3.6. Conclusion and Future Work

3.6.1. Conclusion

This work investigated whether the intrinsic properties of 3DGS primitives encode the photographer's visual saliency and capture intent. We designed an 8-dimensional feature vector covering visibility, spatial, geometric, and appearance properties of each Gaussian primitive, and trained a Random Forest classifier on annotated scenes. Evaluated on Tanks and Temples, MipNeRF360, and a self-captured dataset, our method achieves a mean LOOCV F1 of 0.957 and generalizes to unseen scenes with a mean test F1 of 0.929. Projected 3D saliency masks show strong alignment with U^2-Net predictions, confirming that multi-view Gaussian statistics capture a geometrically consistent, view-stable notion of saliency that single-frame 2D methods cannot provide. These results demonstrate that 3DGS primitives encode information beyond what is essential for rendering, opening new possibilities for capture intent analysis, automatic foreground extraction, and perceptual quality-driven compression.

3.6.2. Limitations and Future Work

Several limitations of this work point towards directions for future research. First, the classifier was trained and evaluated on 16 scenes of relatively small scale. Validating performance on city-

scale datasets would test whether these features generalise to more complex urban environments. Second, the definition of foreground is inherently subjective and may vary across annotators, particularly in scenes with multiple competing subjects. Future work could explore probabilistic labelling approaches that model annotation uncertainty explicitly. Third, our method relies on the quality of the PGSR reconstruction, scenes with significant floaters or incomplete surfaces may yield unreliable feature values, degrading classification performance. Incorporating depth priors could partially address this dependency.

References

Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., & Luo, J. (2025). 3D Gaussian Splatting: Survey, Technologies, Challenges, and Opportunities. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(7), 6832–6852. https://doi.org/10.1109/TCSVT.2025.3538684

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2021). Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. https://doi.org/10.48550/arXiv.2111.12077

Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., & Tian, Q. (2023). Segment Any 3D Gaussians. https://doi.org/10.48550/arXiv.2312.00860

Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., & Zhang, G. (2025). PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, *31*(9), 6100–6111. https://doi.org/10.1109/TVCG.2024.3494046

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual Attention: Bottom-Up Versus Top-Down. *Current Biology*, *14*(19), R850–R852. https://doi.org/10.1016/j.cub.2004.09.041

Dhara, G., & Kumar, R. K. (2025). A survey on visual saliency detection approaches and attention models. *Multimedia Tools and Applications*, *84*(35), 44183–44225. https://doi.org/10.1007/s11042-025-20888-x

Diener, M. J. (2010). Cohen's d. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0200

Fan, L., Yang, Y., Li, M., Li, H., & Zhang, Z. (2024). Trim 3D Gaussian Splatting for Accurate Geometry Representation. https://doi.org/10.48550/arXiv.2406.07499

Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., & Wang, Z. (2023). LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS. https://doi.org/10.48550/arXiv.2311.17245

Fathi, H., Dai, F., & Lourakis, M. (2015). Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics*, *29*(2), 149–161. https://doi.org/10.1016/j.aei.2015.01.012

Feng, Z., Zhan, H., Chen, Z., Yan, Q., Xu, X., Cai, C., Li, B., Zhu, Q., & Xu, Y. (2024). NARUTO: Neural Active Reconstruction from Uncertain Target Observations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21572–21583.

Fischer, T., Kulhanek, J., Bulò, S. R., Porzi, L., Pollefeys, M., & Kontschieder, P. (2024). Dynamic 3D Gaussian Fields for Urban Areas. https://doi.org/10.48550/arXiv.2406.03175

Gilitschenski, I., Jain, U., & Mirzaei, A. (2024). GaussianCut: Interactive segmentation via graph cut for 3D Gaussian Splatting. *Advances in Neural Information Processing Systems 37*, *37*, 89184–89212. https://doi.org/10.52202/079017-2830

Gomes, L., Regina Pereira Bellon, O., & Silva, L. (2014). 3D reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, *50*, 3–14. https://doi.org/10.1016/j.patrec.2014.03.023

Guédon, A., & Lepetit, V. (2024). SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5354–5363.

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 545–552.

Herold, H., & Hecht, R. (2018). 3D Reconstruction of Urban History Based on Old

Maps. In S. Münster, K. Friedrichs, F. Niebling, & A. Seidel-Grzesińska (Eds.), *Digital Research and Education in Architectural Heritage* (pp. 63–79, Vol. 817). Springer International Publishing. https://doi.org/10.1007/978-3-319-76992-9_5

Hou, X., & Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383267

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jiang, W., Lei, B., & Daniilidis, K. (2023). FisherRF: Active View Selection and Uncertainty Quantification for Radiance Fields using Fisher Information. https://doi.org/10.48550/arXiv.2311.17874

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. https://doi.org/10.48550/arXiv.2308.04079

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. https://doi.org/10.48550/arXiv.2304.02643

Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, *36*(4), 1–13. https://doi.org/10.1145/3072959.3073599

Lee, S., Chen, L., Wang, J., Liniger, A., Kumar, S., & Yu, F. (2022). Uncertainty Guided Policy for Active Robotic 3D Reconstruction Using Neural Radiance Fields. *IEEE Robotics and Automation Letters*, *7*(4), 12070–12077. https://doi.org/10.1109/LRA.2022.3212668

Li, Y., Ran, X., Xu, L., Lu, T., Yu, M., Wang, Z., Xiangli, Y., Lin, D., & Dai, B. (2024). ProcGS: Procedural Building Generation for City Assembly with 3D Gaussians. https://doi.org/10.48550/arXiv.2412.07660

Liu, N., & Han, J. (2016). DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 678–686. https://doi.org/10.1109/CVPR.2016.80

Liu, Y., Guan, H., Luo, C., Fan, L., Wang, N., Peng, J., & Zhang, Z. (2024). CityGaussian: Real-time High-quality Large-Scale Scene Rendering with Gaussians. https://doi.org/10.48550/arXiv.2404.01133

Liu, Y., Luo, C., Mao, Z., Peng, J., & Zhang, Z. (2024). CityGaussianV2: Efficient and Geometrically Accurate Reconstruction for Large-Scale Scenes. https://doi.org/10.48550/arXiv.2411.00771

Llull, C., Baloian, N., Bustos, B., Kupczik, K., Sipiran, I., & Baloian, A. (2023). Evaluation of 3D Reconstruction for Cultural Heritage Applications. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1642–1651.

Miao, S., Huang, J., Bai, D., Yan, X., Zhou, H., Wang, Y., Liu, B., Geiger, A., & Liao, Y. (2025). EVolSplat: Efficient Volume-based Gaussian Splatting for Urban View Synthesis. https://doi.org/10.48550/arXiv.2503.20168

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. https://doi.org/10.48550/arXiv.2003.08934

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. https://doi.org/10.48550/arXiv.2304.07193

Qin, M., Li, W., Zhou, J., Wang, H., & Pfister, H. (2023). LangSplat: 3D Language Gaussian Splatting. https://doi.org/10.48550/arXiv.2312.16084

Qin, X., Fan, D.-P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A. C., Suàrez, A., Jagersand, M., & Shao, L. (2021). Boundary-Aware Segmentation Network for Mobile and Web Applications. https://doi.org/10.48550/arXiv.2101.04704

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, 107404. https://doi.org/10.1016/j.patcog.2020.107404

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017). Learning to Detect Salient Objects with Image-Level Supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3796–3805. https://doi.org/10.1109/CVPR.2017.404

Wu, T., Yuan, Y.-J., Zhang, L.-X., Yang, J., Cao, Y.-P., Yan, L.-Q., & Gao, L. (2024). Recent Advances in 3D Gaussian Splatting. https://doi.org/10.48550/arXiv.2403.11134

Xiao, F., Xu, H., Liang, W., & Kang, W. (2025). GSsplat: Generalizable Semantic Gaussian Splatting for Novel-view Synthesis in 3D Scenes. https://doi.org/10.48550/arXiv.2505.04659

Yang, S., Im, C., Lee, J. W., & Choi, J. B. (2026). TIDI-GS: Floater Suppression in 3D Gaussian Splatting for Enhanced Indoor Scene Fidelity. https://doi.org/10.48550/arXiv.2601.09291

Ye, M., Danelljan, M., Yu, F., & Ke, L. (2023). Gaussian Grouping: Segment and Edit Anything in 3D Scenes. https://doi.org/10.48550/arXiv.2312.00732

Zhang, Q., Li, W., Zhang, T., Xiong, R., Zhang, J., Jin, Z., & Li, L. (2025). Representation of top-down versus bottom-up attention in the right dorsolateral prefrontal cortex and superior parietal lobule. *Behavioral and Brain Functions*, *21*(1), 31. https://doi.org/10.1186/s12993-025-00297-8

Zhu, G., Wang, Q., & Yuan, Y. (2014). TagSaliency: Combining bottom-up and top-down information for saliency detection. *Computer Vision and Image Understanding*, *118*, 40–49. https://doi.org/10.1016/j.cviu.2013.07.011

Zuo, X., Samangouei, P., Zhou, Y., Di, Y., & Li, M. (2024). FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding. https://doi.org/10.48550/arXiv.2401.01970

4

Conclusion

4.1. Reflection on the Research

This thesis investigated whether the intrinsic properties of 3D Gaussian Splatting primitives encode the photographer's visual saliency. This problem was investigated from three perspectives. First, regarding which intrinsic properties encode visual saliency, visibility count, neighbour count, and distance to subject emerged as the three most discriminative indicators, jointly accounting for 83.49% of the Random Forest feature importance. Second, when projected onto 2D image planes, the recovered 3D saliency masks show strong alignment with U^2-Net predictions, confirming consistency with human perceptual saliency. Third, when applied to downstream tasks such as foreground extraction, the method achieves a mean LOOCV F1 of 0.957 and generalizes to unseen scenes with a mean test F1 of 0.929, demonstrating strong cross-scene applicability.

An important turning point in the research was an empirical observation made during data exploration. The initial direction of this project focused on leveraging multi-view semantic consistency for geometry-aware 3DGS pruning, following existing approaches that inject external semantic supervision DINOv2 model (Oquab et al., 2023) into the representation. However, while visualizing reconstructed scenes in SuperSplat, a different pattern emerged: spurious Gaussians consistently exhibited needle-like shapes and sparse spatial distributions, in contrast to the compact, planar Gaussians that formed the primary subject surfaces. This observation suggested that the intrinsic geometric properties of the Gaussians themselves already carried discriminative information about the scene structure, without requiring any external semantic input. This insight led to a fundamental shift in research direction, from semantic feature embedding to intrinsic property analysis, and ultimately to the core contribution of this thesis.

Reflecting on the research process, the most challenging aspect was feature selection, as no existing work has systematically investigated Gaussian attributes as saliency indicators. The high-dimensional nature of 3DGS also means the feature combination space is substantially larger than for traditional LiDAR point clouds. We addressed this by evaluating each feature using Cohen's d effect size before training the classifier, which provided a solid foundation for feature design and ablation studies.

4.2. Relation to MSc Geomatics for the Built Environment

This research connects to the MSc Geomatics for the Built Environment programme in several ways.

Data Acquisition and Capture Planning Traditional terrestrial laser scanning workflows follow a manual paradigm: an operator performs an initial scan at fixed positions, identifies regions that are insufficiently mapped, and manually decides where additional scans are needed. This problem is closely related to active reconstruction, a subfield of robotic perception in which an agent autonomously determines where to move next in order to maximise the quality or completeness of an ongoing 3D reconstruction (Feng et al., 2024). Rather than following a pre-planned trajectory, an active system continuously updates its acquisition strategy based on what has already been observed, selecting the

next best view according to an information gain criterion. Information gain is typically defined in terms of reconstruction uncertainty — for instance, by computing the entropy of volumetric weight distributions along rays (Lee et al., 2022), modelling depth rendering variance within neural implicit representations (Feng et al., 2024), or quantifying Fisher Information over radiance field parameters (Jiang et al., 2023). However, these approaches treat all scene regions as equally worthy of reconstruction effort, ignoring any notion of perceptual importance.

Our findings suggest that the intrinsic properties of 3DGS primitives can serve as a perceptual prior, automatically identifying regions that align with the operator's capture intent. This allows active reconstruction systems to deprioritise geometrically incomplete but perceptually unimportant regions, concentrating the acquisition budget where it matters most. Crucially, unlike 2D saliency methods which produce view-dependent predictions that vary with camera pose, our 3DGS-derived saliency is grounded in 3D space and remains stable across viewpoints. This view-stability is essential in practice: saliency estimates derived from an initial coarse scan must remain consistent when predicting the importance of regions observed from entirely new scan positions.

Since our classifier assigns each Gaussian primitive a binary label, the saliency map is represented as a sparse set of flagged primitives in 3D space. For a given view candidate v, the saliency score $S(v)$ can be computed as the proportion of visible Gaussians classified as foreground:

$$S(v) = \frac{1}{|\mathcal{G}_v|} \sum_{g \in \mathcal{G}_v} \mathbf{1}[\hat{y}_g = \texttt{fg}], \tag{4.1}$$

where $\mathcal{G}_v$ denotes the set of Gaussians visible from view v. This score can be directly integrated into any uncertainty-based next-best-view utility function:

$$g(v) = U_{\text{uncertainty}}(v) \cdot S(v), \tag{4.2}$$

where $U_{\text{uncertainty}}(v)$ denotes any conventional uncertainty-based information gain (Feng et al., 2024; Jiang et al., 2023; Lee et al., 2022). Under this formulation, only view candidates that observe geometrically uncertain and perceptually important regions are prioritised, with no additional hyperparameters required.

In practice, a lightweight image-based 3DGS reconstruction could serve as a low-cost pre-survey step, producing a saliency prior that informs a subsequent high-accuracy terrestrial laser scanning. This hybrid workflow combines the accessibility of photogrammetric reconstruction with the geometric accuracy of laser scanning, concentrating the expensive LiDAR capture budget on the regions that matter most.

Bandwidth-Efficient 3D Data Delivery Large-scale 3D scene representations pose significant challenges for storage and streaming in geomatics applications. In traditional urban data formats such as CityJSON and CityGML, Level-of-Detail (LoD) is defined geometrically: coarser levels represent buildings as simple block shapes, while finer levels progressively add architectural details such as rooftop structures and facade elements. This approach treats all regions of equal geometric complexity as equally important.

Recent work on 3DGS-based urban reconstruction has explored similar strategies for real-time rendering, generating different levels of detail through Gaussian compression and selecting the appropriate level based on viewing distance (Y. Liu, Luo, et al., 2024). Our method offers a complementary perceptual dimension: rather than treating all Gaussians within a LoD level uniformly, saliency scores could guide non-uniform compression. Primitives with low saliency scores could be aggressively compressed or discarded, while high-saliency regions are preserved at full quality regardless of viewing distance. For applications such as web-based urban visualization, where bandwidth is limited, a saliency-aware streaming pipeline could prioritize the transmission of perceptually important primitives first, ensuring that the primary structures of interest are always rendered at the highest fidelity even under constrained network conditions.

4.3. Limitations Revisited

The scientific article identified several limitations at a high level. Here we expand on these in more detail.

Annotation Subjectivity The definition of foreground is inherently subjective and dependent on the annotator's interpretation of the photographer's intent. In our dataset, foreground was consistently defined as geometrically coherent surfaces of the primary subject, excluding distant background and floaters. However, in scenes with ambiguous capture intent, where multiple visually similar objects compete for salience, different annotators might produce conflicting labels. This was observed in scenes such as Stump, where the primary subject is surrounded by visually similar objects, resulting in the lowest IoU in our saliency alignment evaluation. Future work could investigate inter-annotator agreement studies or probabilistic labeling that model annotation uncertainty explicitly, rather than enforcing hard binary labels.

Dependence on Reconstruction Quality Our method operates on a trained 3DGS model and its performance is therefore coupled to the quality of the underlying reconstruction. In scenes where PGSR produces floaters or incomplete surfaces, such as regions near the boundary between the sky and the primary subject, visibility counts may be unreliable for the affected primitives. This was reflected in the Train scene, where the metallic surface introduced strong view-dependent appearance variation that destabilized appearance-based features. This dependency could be partially addressed by incorporating additional depth priors or using more robust reconstruction backbones that better handle reflective and transparent surfaces.

Scale of the Dataset The classifier was trained and evaluated on 16 annotated scenes. While the LOOCV results demonstrate consistent generalization, a larger and more diverse dataset would be needed to validate performance across different scene types, capture conditions, and camera setups.

4.4. Future Work

Several directions are worth addressing in future research.

Integration with Semantic Understanding The current method operates without any semantic labels, classifying Gaussians purely based on geometric and photometric attributes. Combining 3D saliency with semantic scene understanding could enable object-level capture intent analysis, where the system not only identifies where the photographer focused but also what category of object it belongs to. This would allow more fine-grained applications, such as automatically tagging salient object by building type.

Extension to Dynamic Scenes The current method assumes a static scene, as 3DGS models are optimized on a fixed set of training images. Extending the approach to dynamic urban scenes (Fischer et al., 2024) would introduce new challenges: a salient object in one frame may be absent in others, and the notion of capture intent becomes time-dependent. Addressing this would require adapting the feature extraction pipeline to account for temporal consistency across frames.

Extension to Larger and More Diverse Datasets Validating the method on city-scale datasets (Y. Liu, Guan, et al., 2024) would test whether these features generalize beyond the relatively small-scale scenes used in this thesis. Collecting a larger annotated dataset with diverse scene types, capture conditions, and camera setups would also further ensure the robustness of the results.

References

Bao, Y., Ding, T., Huo, J., Liu, Y., Li, Y., Li, W., Gao, Y., & Luo, J. (2025). 3D Gaussian Splatting: Survey, Technologies, Challenges, and Opportunities. *IEEE Transactions on Circuits and Systems for Video Technology*, *35*(7), 6832–6852. https://doi.org/10.1109/TCSVT.2025.3538684

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2021). Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. https://doi.org/10.48550/arXiv.2111.12077

Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., & Tian, Q. (2023). Segment Any 3D Gaussians. https://doi.org/10.48550/arXiv.2312.00860

Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., & Zhang, G. (2025). PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, *31*(9), 6100–6111. https://doi.org/10.1109/TVCG.2024.3494046

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual Attention: Bottom-Up Versus Top-Down. *Current Biology*, *14*(19), R850–R852. https://doi.org/10.1016/j.cub.2004.09.041

Dhara, G., & Kumar, R. K. (2025). A survey on visual saliency detection approaches and attention models. *Multimedia Tools and Applications*, *84*(35), 44183–44225. https://doi.org/10.1007/s11042-025-20888-x

Diener, M. J. (2010). Cohen's d. In *The Corsini Encyclopedia of Psychology* (pp. 1–1). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0200

Fan, L., Yang, Y., Li, M., Li, H., & Zhang, Z. (2024). Trim 3D Gaussian Splatting for Accurate Geometry Representation. https://doi.org/10.48550/arXiv.2406.07499

Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., & Wang, Z. (2023). LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS. https://doi.org/10.48550/arXiv.2311.17245

Fathi, H., Dai, F., & Lourakis, M. (2015). Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics*, *29*(2), 149–161. https://doi.org/10.1016/j.aei.2015.01.012

Feng, Z., Zhan, H., Chen, Z., Yan, Q., Xu, X., Cai, C., Li, B., Zhu, Q., & Xu, Y. (2024). NARUTO: Neural Active Reconstruction from Uncertain Target Observations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21572–21583.

Fischer, T., Kulhanek, J., Bulò, S. R., Porzi, L., Pollefeys, M., & Kontschieder, P. (2024). Dynamic 3D Gaussian Fields for Urban Areas. https://doi.org/10.48550/arXiv.2406.03175

Gilitschenski, I., Jain, U., & Mirzaei, A. (2024). GaussianCut: Interactive segmentation via graph cut for 3D Gaussian Splatting. *Advances in Neural Information Processing Systems 37*, *37*, 89184–89212. https://doi.org/10.52202/079017-2830

Gomes, L., Regina Pereira Bellon, O., & Silva, L. (2014). 3D reconstruction methods for digital preservation of cultural heritage: A survey. *Pattern Recognition Letters*, *50*, 3–14. https://doi.org/10.1016/j.patrec.2014.03.023

Guédon, A., & Lepetit, V. (2024). SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5354–5363.

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 545–552.

Herold, H., & Hecht, R. (2018). 3D Reconstruction of Urban History Based on Old Maps. In S. Münster, K. Friedrichs, F. Niebling, & A. Seidel-Grzesińska (Eds.), *Digital Research and Education in Architectural Heritage* (pp. 63–79, Vol. 817). Springer International Publishing. https://doi.org/10.1007/978-3-319-76992-9_5

Hou, X., & Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383267

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jiang, W., Lei, B., & Daniilidis, K. (2023). FisherRF: Active View Selection and Uncertainty Quantification for Radiance Fields using Fisher Information. https://doi.org/10.48550/arXiv.2311.17874

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. https://doi.org/10.48550/arXiv.2308.04079

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. https://doi.org/10.48550/arXiv.2304.02643

Knapitsch, A., Park, J., Zhou, Q.-Y., & Koltun, V. (2017). Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, *36*(4), 1–13. https://doi.org/10.1145/3072959.3073599

Lee, S., Chen, L., Wang, J., Liniger, A., Kumar, S., & Yu, F. (2022). Uncertainty Guided Policy for Active Robotic 3D Reconstruction Using Neural Radiance Fields. *IEEE Robotics and Automation Letters*, *7*(4), 12070–12077. https://doi.org/10.1109/LRA.2022.3212668

Li, Y., Ran, X., Xu, L., Lu, T., Yu, M., Wang, Z., Xiangli, Y., Lin, D., & Dai, B. (2024). Proc-GS: Procedural Building Generation for City Assembly with 3D Gaussians. https://doi.org/10.48550/arXiv.2412.07660

Liu, N., & Han, J. (2016). DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 678–686. https://doi.org/10.1109/CVPR.2016.80

Liu, Y., Guan, H., Luo, C., Fan, L., Wang, N., Peng, J., & Zhang, Z. (2024). CityGaussian: Real-time High-quality Large-Scale Scene Rendering with Gaussians. https://doi.org/10.48550/arXiv.2404.01133

Liu, Y., Luo, C., Mao, Z., Peng, J., & Zhang, Z. (2024). CityGaussianV2: Efficient and Geometrically Accurate Reconstruction for Large-Scale Scenes. https://doi.org/10.48550/arXiv.2411.00771

Llull, C., Baloian, N., Bustos, B., Kupczik, K., Sipiran, I., & Baloian, A. (2023). Evaluation of 3D Reconstruction for Cultural Heritage Applications. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1642–1651.

Miao, S., Huang, J., Bai, D., Yan, X., Zhou, H., Wang, Y., Liu, B., Geiger, A., & Liao, Y. (2025). EVolSplat: Efficient Volume-based Gaussian Splatting for Urban View Synthesis. https://doi.org/10.48550/arXiv.2503.20168

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. https://doi.org/10.48550/arXiv.2003.08934

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. https://doi.org/10.48550/arXiv.2304.07193

Qin, M., Li, W., Zhou, J., Wang, H., & Pfister, H. (2023). LangSplat: 3D Language Gaussian Splatting. https://doi.org/10.48550/arXiv.2312.16084

Qin, X., Fan, D.-P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A. C., Suàrez, A., Jagersand, M., & Shao, L. (2021). Boundary-Aware Segmentation Network for Mobile and Web Applications. https://doi.org/10.48550/arXiv.2101.04704

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, *106*, 107404. https://doi.org/10.1016/j.patcog.2020.107404

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017). Learning to Detect Salient Objects with Image-Level Supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3796–3805. https://doi.org/10.1109/CVPR.2017.404

Wu, T., Yuan, Y.-J., Zhang, L.-X., Yang, J., Cao, Y.-P., Yan, L.-Q., & Gao, L. (2024). Recent Advances in 3D Gaussian Splatting. https://doi.org/10.48550/arXiv.2403.11134

Xiao, F., Xu, H., Liang, W., & Kang, W. (2025). GSsplat: Generalizable Semantic Gaussian Splatting for Novel-view Synthesis in 3D Scenes. https://doi.org/10.48550/arXiv.2505.04659

Yang, S., Im, C., Lee, J. W., & Choi, J. B. (2026). TIDI-GS: Floater Suppression in 3D Gaussian Splatting for Enhanced Indoor Scene Fidelity. https://doi.org/10.48550/arXiv.2601.09291

Ye, M., Danelljan, M., Yu, F., & Ke, L. (2023). Gaussian Grouping: Segment and Edit Anything in 3D Scenes. https://doi.org/10.48550/arXiv.2312.00732

Zhang, Q., Li, W., Zhang, T., Xiong, R., Zhang, J., Jin, Z., & Li, L. (2025). Representation of top-down versus bottom-up attention in the right dorsolateral prefrontal cortex and superior parietal lobule. *Behavioral and Brain Functions*, *21*(1), 31. https://doi.org/10.1186/s12993-025-00297-8

Zhu, G., Wang, Q., & Yuan, Y. (2014). Tag-Saliency: Combining bottom-up and top-down information for saliency detection. *Computer Vision and Image Understanding*, *118*, 40–49. https://doi.org/10.1016/j.cviu.2013.07.011

Zuo, X., Samangouei, P., Zhou, Y., Di, Y., & Li, M. (2024). FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding. https://doi.org/10.48550/arXiv.2401.01970

A

AI Declaration

In this research, LLM tools(Gemini and Claude) are used for data visualisation, paper outline planning, code analysis, and grammar checking. All literature review, research design, mathematical formulations, and scientific contributions are the original work of the author.

B

Reproducibility self-assessment

All data related to this research are publicly available datasets.

- Tanks and Temple: https://www.tanksandtemples.org/
- MipNeRF360: https://jonbarron.info/mipnerf360/

Self captured dataset is only used to evaluate the model performance.

All tools and code used in this research is publicly available.

- PGSR: https://github.com/zju3dv/PGSR?tab=readme-ov-file
- SuperSplat: https://superspl.at/editor
- Repository for the code: https://github.com/MaxineXinyaBi/3DGS_Pruning

I rate the reproducibility of this thesis as **High** according to the provided scale.