Single-Image 3DGS Scene Reconstruction with Geometry-Aware Priors

Machine Visual Perception Course Project Report

December 4, 2025

Information

Authors: Kacper Michalik, Radhika Iyer, Alex Loh
Group Number: 7
Supervisor: Ahmet Canberk Baykal

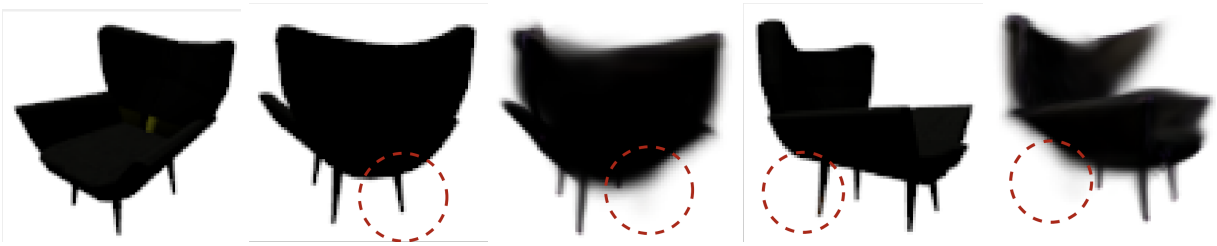
Chapter 1: Introduction and Motivation

1.1 Introduction to the problem

The advancement of 3D data acquisition, reconstruction, and rendering methods remains a fundamental and persistent open problem in computer vision. Efficient, high-quality 3D reconstruction is increasingly critical for applications ranging from augmented reality (AR/VR), autonomous devices (for example in perception or navigation in robotics or self-driving cars) to digital artistry (geometry acquisition for models in VFX, games or other digital products), driving significant research interest in this domain.
Historically 3D reconstruction has been a challenging task that requires large number of reference images and even larger amounts of compute. Advancements in computer hardware and machine-learning methods have significantly improved the efficiency and accuracy of 3D reconstruction, extending its applicability for a wider variety of tasks and hardware platforms; continuing this trend, one area of focus is made on further reducing the number of input images required whilst maintaining high quality reconstruction, thereby improving computational efficiency and reducing data acquisition hardware requirements. Namely, in 2020 Neural Radiance Fields (NeRF) [1] introduced a cutting-edge method for 3D reconstruction capable of high quality novel view synthesis. This is done by learning a continuous volumetric density and radiance function, typically using a deep learning model, which can then be queried by a ray-march for some new camera pose, enabling novel views to be synthesized. In 2023, 3D Gaussian Splatting (3DGS) [2] introduced an alternative method, offering major improvements in computational performance. Instead of an implicit function, 3DGS introduces a new explicit representation, that of a set of 3D Gaussians (called a Gaussian splat), Gaussians are volumes defined by parameters such as position, colour, opacity, rotation and scale, which can be efficiently rasterized to generate novel views. Reconstruction is typically achieved by optimizing the set of Gaussians to produce novel view with minimized loss, alternatively deep learning methods can be used to directly predict the Gaussian splat. Developments in 3D Gaussian Splatting methods have allowed for 3D scene reconstruction using few or even single RGB images. While faster than other scene reconstruction techniques and requiring only a "one-shot" pass, these approaches often suffer from challenges such as layout/scale drift, over-smooth geometry and hallucinations in occluded regions [3].

This project focuses on one recent method, Splatter Image [4], as a baseline. Splatter Image allows single or few RGB image 3DGS reconstruction. Achieved by predicting 3D Gaussians as pixels in a multichannel image; this representation reduces reconstruction to learning an image-to-image neural network, allowing the use of a 2D U-Net to form the representation. Each pixel stores the parameters for a corresponding 3D Gaussian, allowing for reconstruction in a single feed-forward pass. This overall architecture allows for a compute-efficient model. Despite its speed, Splatter Image has some issues that have been noted in related

works, particularly in reconstructing structures unseen in the input view, including for views significantly different from the source. We believe there are two reasons for this problem. The first is inherent to Splatter Images architecture, unlike methods that utilize explicit 3D feature volumes, Splatter Image's choice of 3D reconstruction as a 2D-to-2D image translation task limits its ability to learn geometric priors, as the model lacks an internal 3D representation to resolve issues like depth ambiguities. The second is that the 3DGS prediction based on only single or few RGB image features alone does not have sufficient conditional information for Splatter Image to infer appropriate geometry information or structures that are not visible in the input view [5]; shown in figure 1, Splatter Image has trouble generating the occluded chair leg in 1c and 1e.



(a) Input image (b) Ground truth (c) Prediction (d) Ground truth (e) Prediction

Figure 1: Splatter Image outputs compared with ground truth taken from [5]

This project aims to address these issues, improving reconstruction quality, by first researching inferable geometry priors (such as planes, normals, visibility cues, depth, segmentation or edge maps) which can be dynamically produced for input images by existing specialized models, then proposing a lightweight augmentation for Splatter Image, allowing predicted priors to be fed alongside the RGB images, allowing them to guide reconstruction in a more accurate manner, by providing necessary additional information and preventing Splatter Image from having to learn how to generate these geometric features itself.

1.2 Background and related work

[Include a few very relevant related works and how your work relates to those, expanding on the previous section. We do not expect you to cover all previous works.]

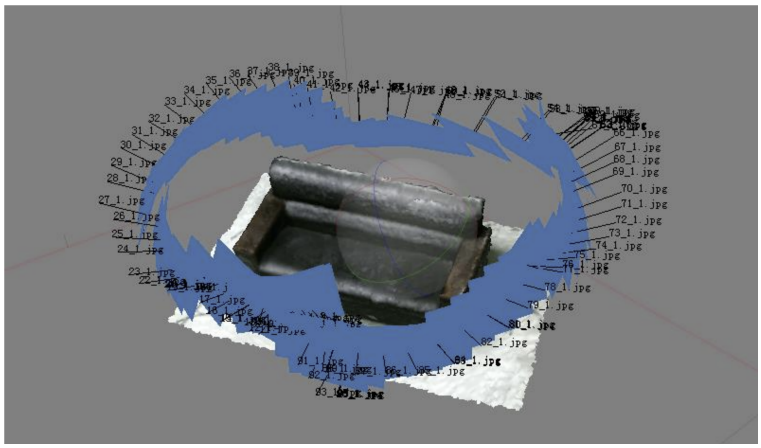


Figure 2: An example of 3D reconstruction with many overlapping images from [6]

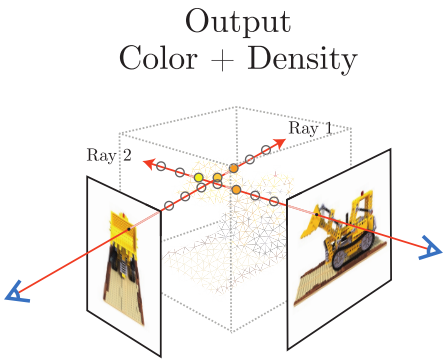


Figure 3: NeRF implicit function from [1]

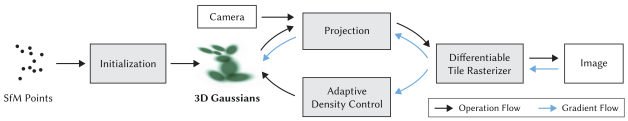


Figure 4: 3D Gaussian Splatting Process from [2]

Traditionally, 3D reconstruction has been performed by multi-stage photogrammetry pipelines, relying on explicit geometric representations such as meshes or point clouds. The industry-standard workflow begins with Structure from Motion (SfM), which matches sparse feature points across many overlapping images to estimate camera parameters and generate a sparse point cloud. This is typically followed by Multi-View Stereo (MVS) algorithms to compute dense depth maps, which are combined to generate a standard 3D mesh using techniques like Delaunay triangulation[7] or Moving Least Squares with Marching Cubes[8]. While effective for static, diffuse environments, these methods struggle significantly with surfaces such as transparent windows or reflective metals, this as they rely on strict photometric consistency and lack a mechanism to deal with view-dependent radiance. Additionally, these methods require large numbers of high-resolution images with substantial overlap to achieve high-quality reconstruction. This heavy data acquisition requirement means the methods create a computational bottleneck, requiring hours of processing time on high-end hardware, and are impractical in the first place without complex image acquisition setups, such as in Figure 2.

A paradigm shift occurred in 2020 with the introduction of Neural Radiance Fields[1]. NeRF moves away from explicit geometry representations to an implicit volumetric representation. In NeRF an underlying continuous volumetric scene function, represented using a deep learning model, is optimized using a set of input images with known camera poses; the input to the function/model is a single continuous 5D coordinate (spatial location and viewing direction) and the output is the volume density and view-dependent emitted radiance at that location. Novel views are synthesized by querying 5D coordinates along camera rays for some new camera pose (ray-marching), and output colours and densities are rendered into an image. NeRF allowed for higher quality 3D reconstruction, eventually using sparser image sets (PixelNeRF, RegNeRF), compared to traditional photogrammetry methods, achieving state-of-the-art results and becoming the gold standard for novel view synthesis[9].

In 2023 a further breakthrough in the field was achieved by the introduction of 3D Gaussian Splatting, offering a computationally high-performance alternative to NeRF. 3DGS offers a new explicit geometry representation, modelling a scene as a collection of parameterized 3D Gaussians[2]. Unlike NeRF, 3D Gaussian Splatting does not rely on a neural network to generate a scene. Instead, in the original paper, reconstruction is achieved by first initializing a set of Gaussians, either randomly or from a sparse point cloud (typically from Structure from Motion), where each Gaussian is defined by a set of parameters: position (mean), covariance (scale and rotation), opacity, and colour (represented using Spherical Harmonics to allow for view-dependent radiance). Then an optimization process is run that first adjusts the Gaussians' parameters to minimize the error between rendered images and the ground truth; and secondly performs dynamic management of the density of Gaussians within the scene, by splitting, cloning or pruning Gaussians in an interleaved manner; namely Gaussians in under-reconstructed areas are cloned, Gaussians with high variance are split, and Gaussians in areas of low opacity and of excessive size are pruned. Finally, for rendering, the Gaussians are rasterized using a custom tile-based rasterizer to produce resulting views, allowing millions of Gaussians rendered in real-time, allowing for real-time novel view synthesis. While the original method relies on per-scene optimization to generate 3D Gaussian Splats, recent works have begun using deep learning methods to predict sets of Gaussians directly, with most recent developments allowing for 3D scene reconstruction using few or even single RGB images.

Since the introduction of 3DGS, a number of deep learning architectures and processing pipelines based on the method have been developed to find the most accurate and efficient implementation capable of producing high quality 3D Gaussian Splats. Recent examples include ExScene[10], Wonderland[11], F3D-Gaus[12], Gauss VideoDreamer[13], TGS[14] and Splatter Image[4].

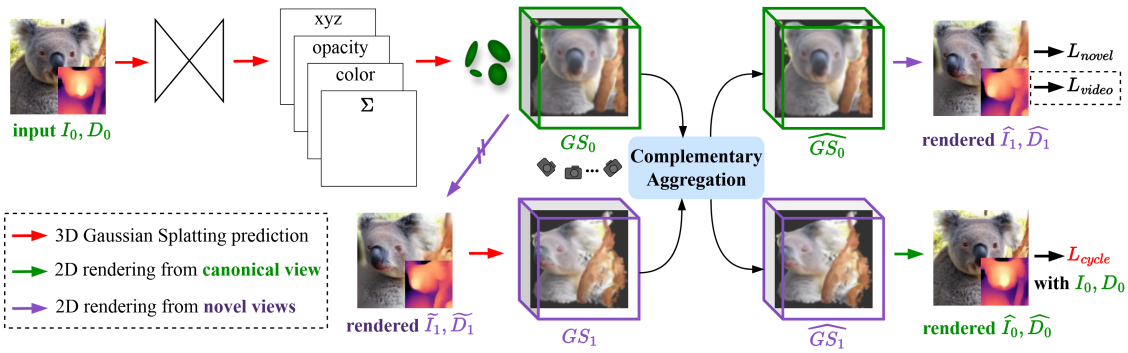


Figure 5: F3D-Gaus Framework from [12]

F3D-Gaus was proposed to help pixel-aligned Gaussian Splatting generate plausible novel viewpoints by introducing a cycle-aggregative strategy. As shown in figure 5, given a single RGB image I_0 and depth map D_0, the model directly feeds them forward to output the pixel-aligned Gaussian Splatting representation GS_0. The model then renders the image $\tilde{I}_1$ and depth map $\tilde{D}_1$ for the novel view, and then outputs its corresponding Gaussian Splatting representation GS_1. GS_0 and GS_1 are then aggregated to produce images for supervision. By enforcing this cycle consistency through aggregation, the model naturally learns through this self-supervised approach to extrapolate across views by fusing multiple representations, enhancing 3D coherence during inference. This helps to ensure that 3D representations from novel viewpoints are both aligned with and complementary to those from the original viewpoint. After complementary aggregation, F3D-Gaus then applies a video in-painting model[15] to the rendered images and depth maps to correct inconsistencies in geometry and texture caused by large viewpoint shifts, resulting in more reliable and visually consistent outputs. While the source code and a pretrained model is available, the model would require 13 days of an A100 GPU to train if any updates were made, so we deemed this unsuitable to work on.

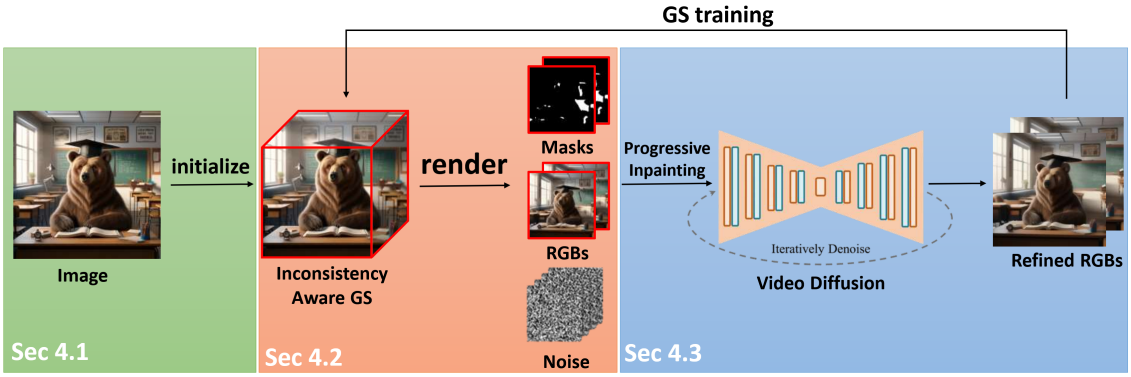


Figure 6: GaussVideoDreamer Framework from [13]

From the input image, GaussVideoDreamer first initializes a coarse video sequence depicting occluded regions of the input image under smooth novel view trajectories and inconsistency-aware Gaussian Splats. After that, at periodic optimization intervals, the model renders all viewpoint images and their corresponding inconsistency prediction masks from the Gaussian Splats. The masks and rendered images then guide a video diffusion model to perform progressive inpainting to refine the video sequence, which then in turn optimizes the Gaussian splatting representation and gradually generates better novel view images. However,

the generation quality of GaussVideoDreamer is mainly limited by the video diffusion prior (AnimateDiff[16]), which was adapted from an image diffusion model lacking dedicated video inpainting training. As such, when generating videos it may result in frame inconsistencies in longer sequences, progressive colour shifts and high-frequency detail degradation, which reduces novel view reconstruction quality. There was also no publicly available implementation of this method. In addition, since a proprietary evaluation dataset of 20 generated single-view images with corresponding text prompts was used in the paper, it would be difficult to compare it with other implementations. Therefore, we also decided not to explore this method.

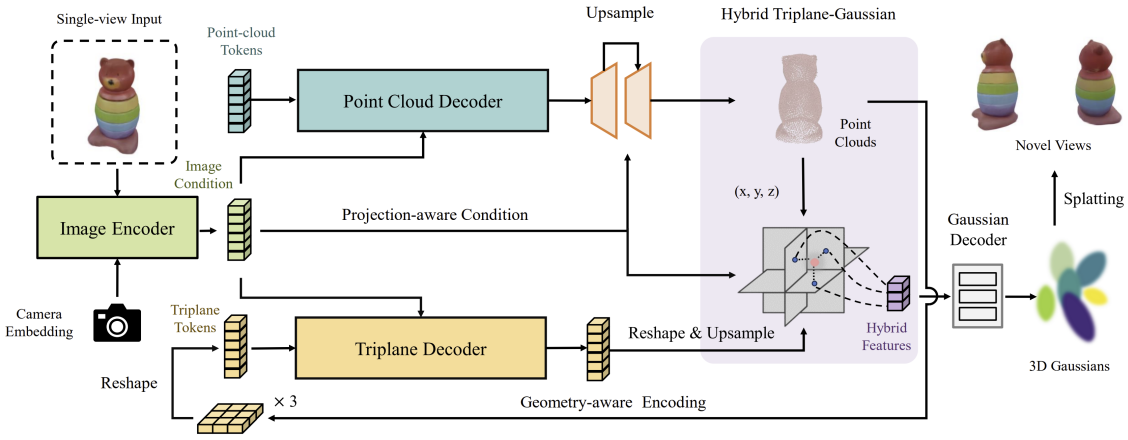


Figure 7: TGS Framework from [14]

In a triplane representation[17], explicit features of the image are aligned along three axis-aligned orthogonal feature planes. We can then query any 3D position by projecting it onto the three feature planes, retrieving the corresponding feature vector via bilinear interpolation and aggregating the three feature vectors via summation. An additional lightweight decoder network interprets the aggregated 3D features as colour and density. This representation is fast and scales efficiently with resolution by keeping the decoder small and shifting the bulk of the expressive power into the explicit features. TGS[14] uses this representation in Triplane-Gaussian, a new hybrid representation. As shown in figure 7, TGS first encodes the input image into a set of latent feature tokens, which are then passed into a point cloud decoder and a triplane decoder . A point cloud and a triplane can be de-tokenized from the output of these decoders. After the point cloud decoder, TGS adapts a point upsampling module to densify the point cloud and utilizes a geometry-aware encoding to project point cloud features into the initial positional embedding of triplane latent. Lastly, 3D Gaussians are decoded by the point cloud, the triplane features and the image features for novel view rendering through Gaussian splatting. While an official implementation of this is available at https://github.com/VAST-AI-Research/TriplaneGaussian and a pretrained model is also available, the model is only trained on the Objaverse-LVIS dataset and the training time and compute needed to adapt this implementation also made it unsuitable for our project.

1.3 Overview of the idea

[Provide an overview stating why the idea of the project makes sense and what the main motivation is.]

Splatter Images high computational performance and relative reconstruction quality makes it a highly desirable model for one-shot 3D reconstruction

the model still occasionally suffers from challenges such as layout/scale drift, over-smooth geometry and poor quality hallucinations in occluded regions, particularly in reconstructing structures unseen in the input view, including for views significantly different from the source.

We believe there are two reasons for this problem. The first is inherent to Splatter Images architecture, unlike methods that utilize explicit 3D feature volumes, Splatter Image's choice of 3D reconstruction as a 2D-to-2D image translation task limits its ability to learn geometric priors, as the model lacks an internal 3D representation to resolve issues like depth ambiguities. The second is that the 3DGS prediction based

on only single or few RGB image features alone does not have sufficient conditional information for Splatter Image to infer appropriate geometry information or structures that are not visible in the input view.

this is fundamanetally due to the lack of inforamtion the model has from a single input image

Currently there exists specialised, accureate models that can perorm on-shot prediction with hig haccuracy spefific geomtric features of images trainined on many examples so store lot of konwdleg about corectly halkucinating occluded reegions in a realsitic/more accuraet manner

We propsoe exploting thse models to gain additional sources of inormation which could be fed into the model these priros solve both of our beleived problems, namely they provide provide large amounts of additional information, tackling issue no 2, and stop the model from nedign to learn hwo to infer these geometric features itself, which it is unsuitable/poor at, tackling issue no 1. We bleive as such these priors would guide recosntruction in a more accurate manner, impriving the recosntruction quality of the model, whislt requring minimal architectural changes to Splatter Image, as only have to change inpotu channel count and do minor modifications if we want to add multimodal data, generally preserving its performance,

we also propose performing ablation study to see which priors are most effective/significant in changing the reconstruction quality

THJERBY TACKLING

Chapter 2: Method

2.1 Baseline algorithm

[Explain the baseline architecture you used to build your algorithm on. You may reproduce figures from the original papers.]

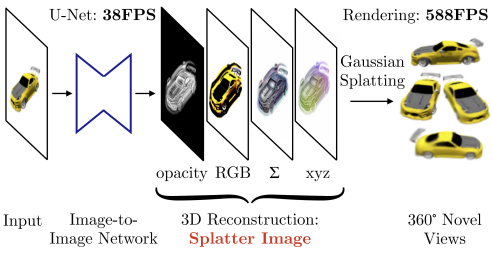


Figure 8: Overview of SplatterImage[4]

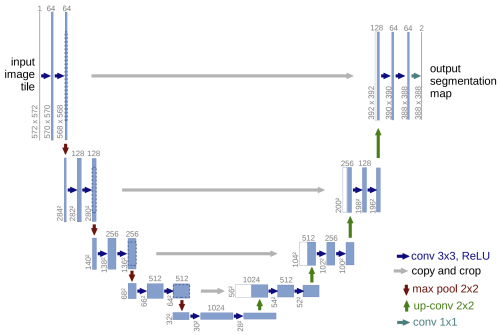


Figure 9: U-net architecture[18] that Song U-Net[19] is based on

Splatter Image uses a standard image-to-image neural network architecture to predict a Gaussian for each pixel of the input image I, generating the output image M as the Splatter Image. For this it uses a U-Net as those have demonstrated excellent performance in image generation[20], and their ability to capture small image details helps to obtain higher-quality novel view reconstructions.

As shown in figure 9, a U-Net model architecture consists of a contracting path (left side) and an expansive path (right side). The contracting path aims to capture context and follows the typical architecture of a convolutional network, which consists of the repeated application of convolutions, each followed by an activation function and a max pooling operation for downsampling. At each downsampling step we increase the number of feature channels. Afterwards, each step in the expansive path consists of an upsampling of the feature map followed by a convolution that decreases the number of feature channels, enabling precise localization. At the final layer a convolution is used to map each feature vector to the desired number of classes.

For the U-Net used in Splatter Image, learning to predict the Splatter Image can be done on a single GPU using at most 20GB of memory at training time for most single-view reconstruction experiments (except for Objaverse, where 2 GPUs were used and 26GB of memory was used on each). Most of this neural network

architecture is identical to the SongUNet of [19], but the last layer is replaced with a 1×1 convolutional layer with $12 + k_c$ output channels, where $k_c \in \{3, 12\}$ depending on the colour model. The output tensor codes for parameters that are then transformed to opacity, offset, depth, scale, rotation and colour respectively. These parameters are then activated by non-linear functions to obtain the Gaussian parameters, such as the opacity and depth.
The Gaussian Splatting implementation of [2] is used for rasterization to generate 360° views of the original input image from these parameters. As shown in figure 4, Gaussian splatting then optimizes and adaptively controls the density of this set of 3D Gaussians, using a fast tile-based renderer during optimization. The renderer allows α-blending of anisotropic splats, respecting visibility order thanks to fast sorting. It also includes a fast backwards pass by tracking accumulated α values, without a limit on the number of Gaussians that can receive gradients. This helps to quickly generate new images of novel views.

2.2 Algorithm improvements

[Explain what you implemented to improve over the baseline. You may include figures to explain the idea and logic. Focus on the ideas and not the implementation.]

2.2.1 Model Improvements

The first modification to Splatter Image's architecture is to allow the model to be initialised with a dynamic number of input channels, as opposed to the standard 3 for RGB; the required number of channels is calculated by the `GaussianSplatPredictor` module at runtime, based on the supplied training configuration. This allows models to be created that use the desired combination of additional priors.

The second modification is support for multimodal priors. While priors like depth or normal maps, which are images themselves, can be appended as additional channels to RGB images to generate the final model input, structurally different priors cannot simply be appended in the same way. For example, modern segmentation models, like some of those we researched, are capable of producing classifications (and instance IDs in the case of instance or panoptic segmentation) for identified segments. These classifications are typically in the form of strings or vector embeddings (which can be produced from strings regardless). These vector embeddings do not have matching dimensions to the RGB images and thus cannot be added as an extra channel in the model input. There are a variety of ways these multimodal priors can be provided to the model, one method is broadcasting; in broadcasting multimodal inputs are appended to an image by replicating them across every pixel. Since the multimodal input is typically a 1D vector (say a scalar or vector embedding) and the image is a 3D tensor (Height $\times$ Width $\times$ Channels), the vector is replicated at every single pixel location by being concatenated along the channel dimension. This method is poor due to its computational cost and massive data redundancy: the network is forced to process and store the exact same values millions of times, which wastes GPU memory and computation. Alternatively, Feature-wise Linear Modulation (FiLM) offers a much more efficient method for multimodal data input for U-Nets and Convolutional Layers. Instead of multimodal data being inserted into the input image, FiLM injects this data by modulating the intermediate feature maps of the network. FiLM layers are inserted into the model at specific points, for example between a convolution and a ReLU activation or at the end of a UNet block, these FiLM layers contain a generator, which is a separate, small neural network (like an MLP or RNN) which takes only the multimodal data (such as the segmentation embedding, which we will call $\mathbf{z}$) as input, and generates two output terms, γ (scale) and β (shift). The FiLM layer then takes the intercepted feature map and applies the FiLM equation to it:

$$\hat{F} = \gamma(\mathbf{z}) \cdot F + \beta(\mathbf{z}) \tag{1}$$

This modulated feature map is finally passed on within the network. The method is highly computationally efficient, only requiring a multiplication and addition be performed to terms in the feature map, and allows the models image input to remain unchanged. As such this strategy of inserting FiLM layers into Splatter Image is how we achieve multimodal data support.

An alternative to FiLM is also considered, namely cross-attention. FIXXX
Cross-attention modules can be used as slot-in replacements for FiLM layers inside the U-Net's blocks (in the bottleneck and decoder). The U-Net's image features act as the Query (Q). The multimodal input embeddings act as the Key (K) and Value (V). More expressive than FiLM as it allows for fine-grained spatial conditioning (per-pixel)., this method is consdiered state-of-the-art for high-performance conditional generation. FIXXX

2.2.2 Low-Rank Adaptation (LoRA)

To address the computational constraints of training the full U-Net architecture, we integrated Low-Rank Adaptation (LoRA) [21] directly into our GaussianPredictor, to instead allow for fine-tuning. While our implementation shares a similar class structure to the official `loralib` library [22] (utilizing mixins to wrap `Linear` and `Conv2d` layers), we manually adapted the forward pass to support the specific channel dimensions of the Splatter Image architecture (further details can be found in **Section 2.3.6: Manual LoRA Integration**).
Adapters are small modules placed after the frozen modules we wish to adapt, such as linear and convolutional layers. The adapter accepts the same input dimension as the original layer and produces the same output dimension. This allows the output from the adapter to be summed element-wise with the frozen layer's output.
Instead of learning a new large weight matrix for these adapters, the weight update is decomposed into two smaller low-rank matrices. For a weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$, the update is defined as $W + \Delta W$, where ΔW is factored into:

$$A \in \mathbb{R}^{r \times d_{in}} \quad \text{and} \quad B \in \mathbb{R}^{d_{out} \times r}$$

Here, we see that $r \ll \min(d_{in}, d_{out})$, where r is the rank hyperparameter (detailed below). During training, only the parameters of A and B are updated, while the original weights W remain frozen.
For `Conv2d` layers, we treat the kernel $W \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$ as a flattened matrix of shape $C_{out} \times (C_{in} \cdot k \cdot k)$. The flattened representation is decomposed into B and A, allowing us to apply LoRA to the full spatial kernel. By including the spatial dimensions ($k \times k$) in the decomposition, the adapter can learn spatial feature refinements, rather than being limited to channel-wise linear projections. This approach follows the implementation found in `loralib` [22].
During the forward pass, the input x is processed by both the frozen weights W and the LoRA branch:

$$h = Wx + \frac{\alpha}{r} BAx \tag{2}$$

where $\frac{\alpha}{r}$ is a scaling factor [22]. This scaling normalises the updates across different rank choices, reducing the need to re-tune the learning rate when r changes. We utilise three hyperparameters to control this adaptation:

- Rank (r): The rank of the low-rank matrices A and B. Higher ranks increase the number of parameters and the capacity of the adaptation, but also increase computational cost
- Alpha (α): A scaling factor applied to the LoRA update during the forward pass. The update is scaled relative to the weights that have been frozen from the pretrained model.
- Dropout (p): The dropout probability applied to the LoRA layers during training. This randomly disables activations, which aims to prevent overfitting.

Following [22], A uses Kaiming uniform intialisation, and B is initialised to zero. This ensures that $\Delta W = 0$ at the start of training, which preserves the behavior of the pre-trained model initially.

2.3 Implementation details

[Explain how you implemented the improvements. You may include code snippets with the corresponding explanations.]

All code associated with the project can be found in the following repositories:

3DGS-priors (Top level repository for the project): `https://github.com/Kacper-M-Michalik/3dgs-priors`
Splatter Image Fork: `https://github.com/Kacper-M-Michalik/splatter-image`

Generated datasets and model weights can be found in the following repositories:

Datasets with Predicted Priors: `https://huggingface.co/datasets/MVP-Group-Project/srn_cars_priors`
Pretrained Models: `https://huggingface.co/MVP-Group-Project/splatter-image-priors`

2.3.1 Planes and Normal Maps Exploration

We considered providing the model with structural information to be one of the most likely avenues of improvement. These structural priors were considered in two flavours, in the form of predicted scene planes and scene surface normal maps.

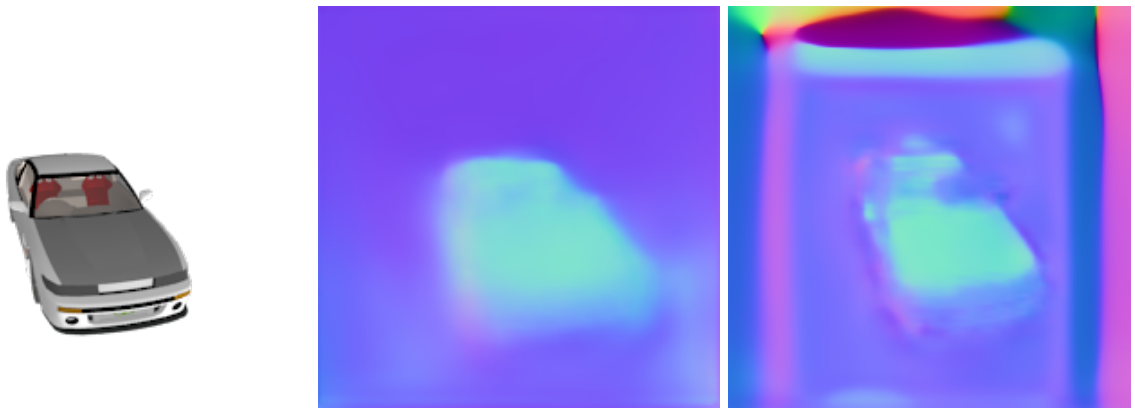
When researching plane prediction, we reached the conclusion that this flavour would in fact be unlikely to help guide reconstruction. For example teddy bears (as seen on the CO3D[23] Teddybears dataset) have complex, convex shapes as shown in 10, lacking dominant planes on their surface, using a planar prior might confuse the network, causing it to flatten the bear's features or causing poor quality hallucination; as such we decided against using planes as a prior.

Figure 10: Example of CO3D Teddybears dataset from [23]

A much more favourable option were normal maps. Normal maps store surface normal data as RGB colour information, showing the orientation of a surface on a per-pixel level, we considered this to be an excellent prior as it supports both complex shapes such as teddy bears, but can equally well describe a planar surface. Hence, we selected this prior as a prime candidate that could improve the models' 3D surface reconstruction and help guide accurate hallucination in occluded regions. We take our ground truth images from a dataset (`https://github.com/Xharlie/ShapenetRender_more_variation`) provided by [24] which contain higher resolution images of ShapeNet[25] models. Each RGB image of a ShapeNet model is paired with its corresponding depth map, normal map and albedo map as shown in figure 11. We feed these images into the normal map generation models and compare against the ground truth normal maps, using Pixel Based Visual Information Fidelity as a metric to evaluate their performance.

Figure 11: Example of image with maps used as ground truth taken from [26]

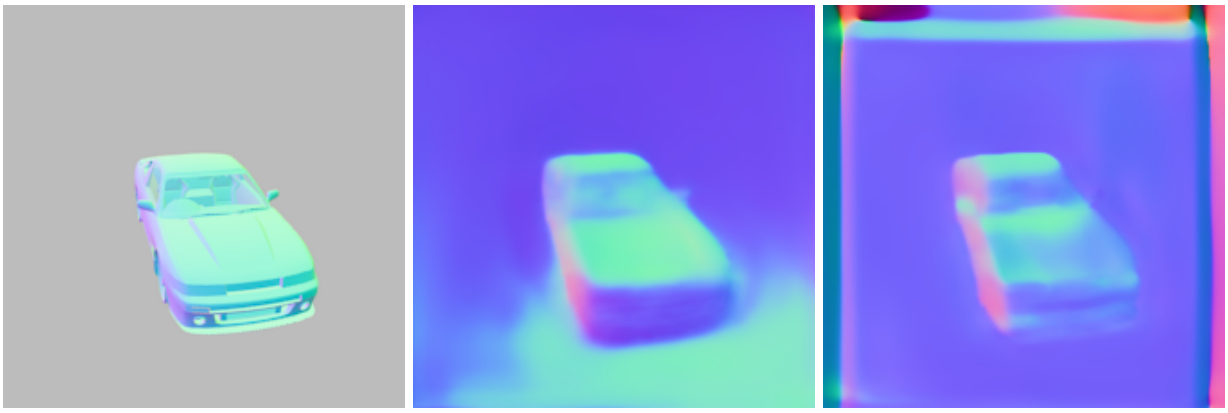
For normal map generation we used models from [26] which implements a network which estimates the per-pixel surface normal probability distribution and uses uncertainty-guided sampling to improve the quality of prediction of surface normals. The paper provided code at `https://github.com/baegwangbin/surface_normal_uncertainty` that implemented this method on a network trained on ScanNet [27], with the ground truth and data split provided by FrameNet [28], and another trained on NYUv2 [29], with the ground truth and data split provided by GeoNet [30] [31]. Both models take in the original image and dimensions of the image as input and return a corresponding normal map with the same dimensions as the given input dimensions. We run both pretrained models on the dataset.



(a) Original (b) ScanNet output (224x224) (c) NYUv2 output (224x224)

Figure 12: Comparison of original input and two model outputs

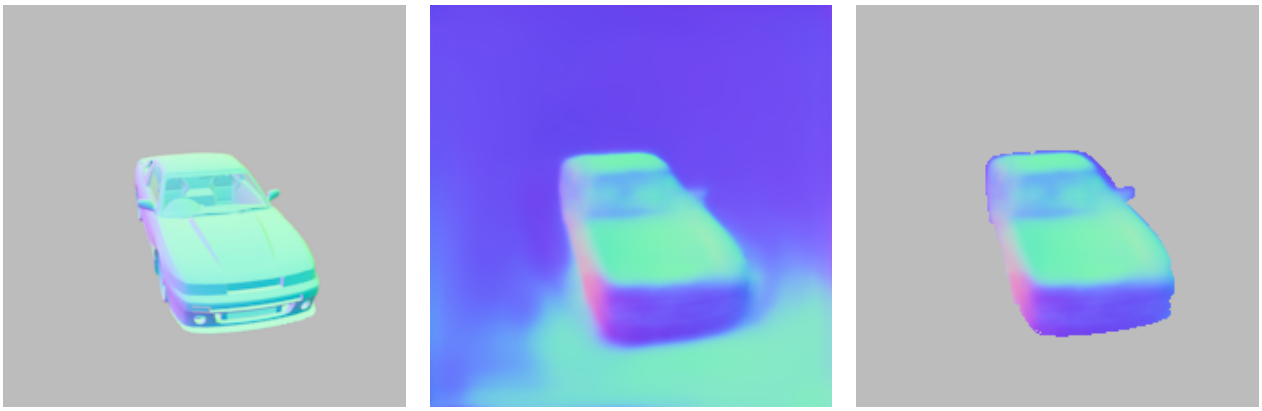
We then pass in input dimensions larger than the actual ones into the models, such that a normal map larger than the original input is produced. We then resize the image to the original input dimensions.



(a) Ground truth from [24] (b) ScanNet output (448x448) (c) NYUv2 output (448x448)

Figure 13: Comparison of model outputs when setting input dimensions as 448x448 instead of 224x224 alongside ground truth

The normal map generated for images when given larger input dimensions seem to have more clearly defined edges and surface contouring as shown in figures 13b and 13c as compared to figures 12b and 12c. It is also important to note that the ground truth for NYUv2 is only defined for the centre crop of the image and the prediction is therefore not accurate outside the centre. This is shown in figures 12c and 13c where noise is generated around the borders of the normal maps.
To compare our generated normal maps to the ground truth normal maps provided in [24], we first mask out the background of the generated normal maps such that the difference in background colour does not contribute to the evaluation metrics for normal map generation.



(a) Ground truth from [24] (b) ScanNet output (448x448) (c) Output with background masking

Figure 14: Example of masking out background for model evaluation against ground truth

We then use Pixel Based Visual Information Fidelity to compare the normal maps generated by the two models to the ground truth. Visual Information Fidelity is a reference image quality metric that quantifies the amount of visual information preserved after image processing [32] and can be used to measure various image quality attributes such as noise level and sharpness [33].

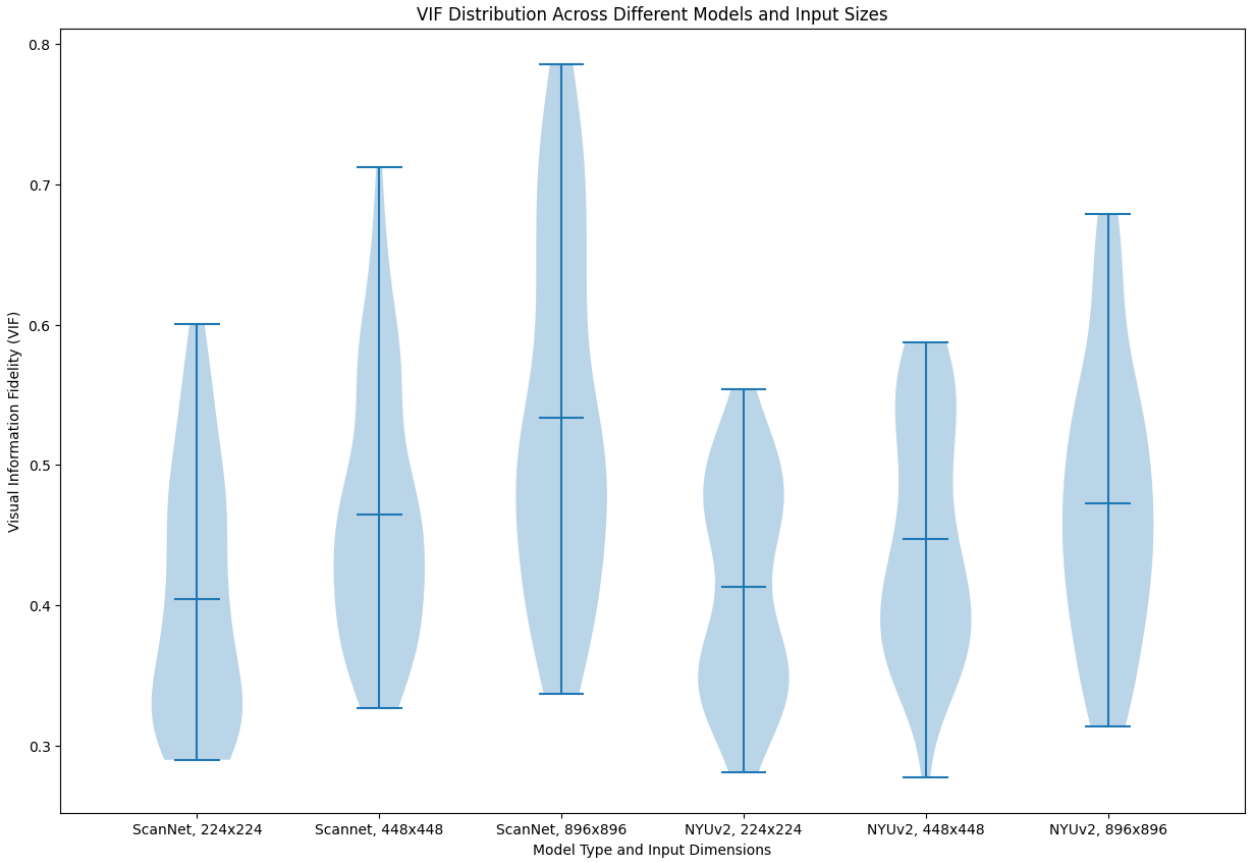
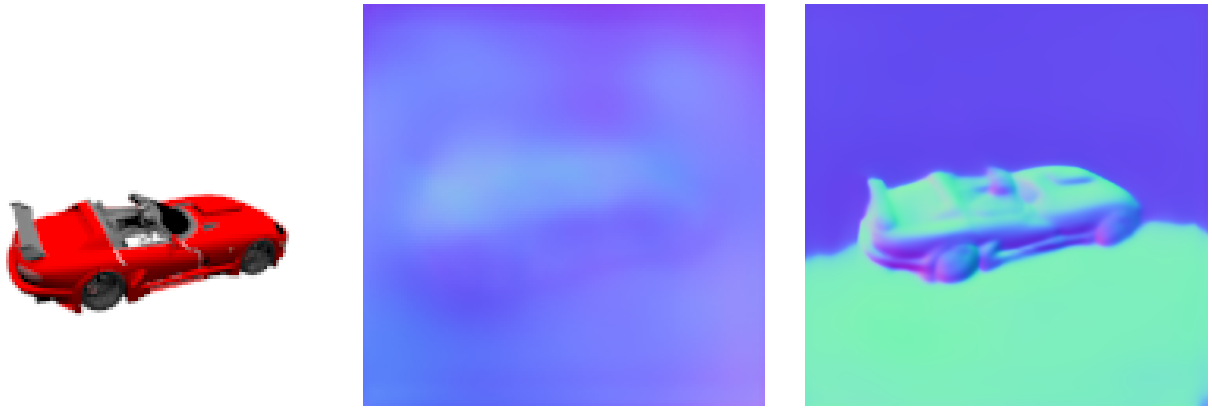


Figure 15: Comparison of VIF between ground truth and different models

From Figure 15 we see that that normal map generation quality increases when passing in larger input arguments and that the model trained on ScanNet generates normal maps that are closer to the ground truth compared to that trained on NYUv2 on average. Hence, in the final model we decided to use the model trained on ScanNet on the ShapeNet database in [25].



(a) Original image from [25] of size 128x128

(b) Output from ScanNet model with 128x128 passed in as input dimensions

(c) Output from ScanNet model with 512x512 passed in as input dimensions

Figure 16: Original ShapeNet image and normal map outputs

Without passing in dimensions larger than the input image into the model, we can see from comparing Figures 12a and 12b to Figures 16a and 16b that the quality of the normal map generated decreases as the resolution of the original input image decreases. Hence, we pass in much larger input dimensions (512x512) to generate a normal map of higher quality, as shown in Figure 16c.

2.3.2 Depth Map Exploration

Depth maps store the distance of a surface from the camera per-pixel. These distances vary in type, such as metric, which considers the physical distance from the camera to the observed point, and relative (such as those produced by the models below). Monocular depth estimation (MDE) models input just a singular image, and produce a depth map (relative distance).
Produced depth maps were compared against the "ground truths" produced by `https://github.com/Xharlie/ShapenetRender_more_variation`, as was done in the normal priors exploration. An example of the depth map produced by them is visible in Figure 11. However, it is important to note that these depth map "ground truths" were not always perfect, as can be seen in the following example:

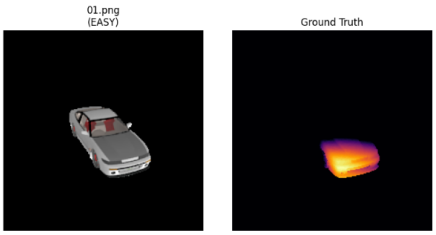


Figure 17: Example of poor depth ground truth data

This inclined us to take the quantitative results produced by comparing MDE models tested against these ground truths with a pinch of salt. For each produced depth map, the following metrics were used to compare against the ground truths.

1. **Absolute Relative Error**: Measures the average difference between the predicted depth and the ground truth, normalised by the ground truth depth.
2. **Root Mean Squared Error (RMSE)**: Calculates the standard deviation of the residual errors.
3. **Scale-invariant RMSE (SI-RMSE)**: Computes the RMSE while ignoring the unknown absolute scale and shift between the prediction and ground truth.
4. δ at **1.25** ($\delta_{1.25}$): Represents the percentage of predicted pixels p that satisfy the condition $\max(\frac{p}{p^{gt}}, \frac{p^{gt}}{p}) < 1.25$, which takes into account close pixel-wise agreement.

The following table summarises the metrics across the MiDaS models tested.

Table 1: Comparison of MiDaS models on set of easy and hard images.

Difficulty	**Model**	**AbsRel** $\downarrow$	**RMSE** $\downarrow$	**SI-RMSE** $\downarrow$	$\delta < 1.25 \uparrow$
Easy	DPT_Hybrid	0.089 ± 0.12	20.38 ± 19.39	0.123	0.909
Easy	DPT_Large	0.091 ± 0.12	20.56 ± 19.74	0.124	0.909
Easy	MiDaS_small	0.096 ± 0.13	21.54 ± 20.44	0.129	0.918
Hard	DPT_Hybrid	0.101 ± 0.15	19.65 ± 17.99	0.128	0.907
Hard	DPT_Large	0.170 ± 0.41	22.40 ± 22.12	0.151	0.906
Hard	MiDaS_small	0.190 ± 0.45	24.45 ± 23.56	0.164	0.900

The quantitative degradation of `DPT_Large` on the Hard set contradicts visual inspection. This discrepancy can be attributed to the quality of the available Ground Truth (GT) depth maps (as discussed earlier).

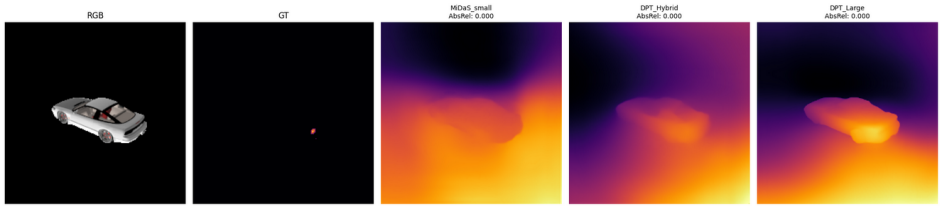


Figure 18: Depth Maps produced by MiDaS models on image with poor GT depth map.

Although `DPT_Hybrid` appears to align more closely with the GT depth map, `DPT_Large` is what is used in the final depth prior generation (as described in **2.3.4 Selected Prior Integration**). One reason is that the model produces depth maps with cleaner edges along the object boundaries (unlike `DPT_Hybrid`), which can be seen to have closer sections of the object blend into the foreground pixels.

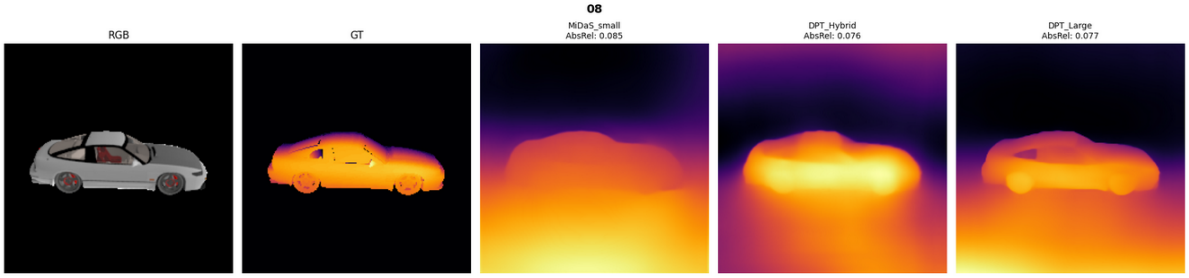


Figure 19: Depth Maps produced by the MiDaS models

Exploration of Depth Discontinuities

We also explored the utility of explicit depth discontinuity features, or depth edges. These reflect sudden changes in depth values (when the first spatial derivative of the depth function is large) within a depth map. These can serve as a proxy for occlusion boundaries, and we were investigated as they potentially aid in separating foreground objects from background geometry.

We looked four classical edge detection operators applied directly to the estimated depth maps:

1. **Canny Edge Detector:** Tends to produce sparse, crisp boundaries, but with noisy depth maps, it can form edges that are disjoint. Available within the OpenCV (`cv2.Canny`) library.
2. **Sobel Operator:** Approximates using weighted convolution kernels, acting as a low-pass differentiator.
3. **Scharr Operator:** Similarly to Sobel, uses weights, works to improve the Sobel operator to provide better rotational symmetry.
4. **Prewitt Operator:** Similar to Sobel, but uses a simpler integer-valued kernel (uniform weighting across rows and columns).

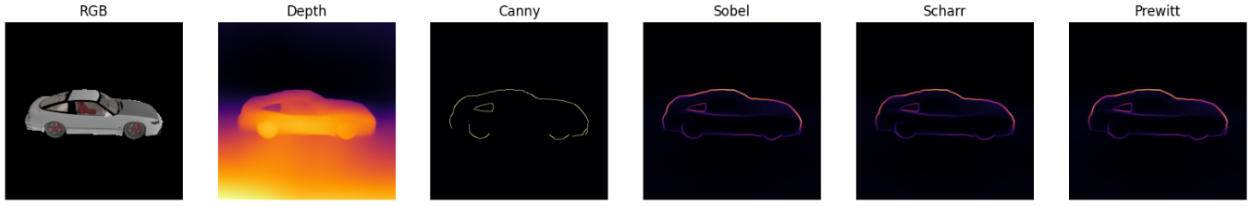
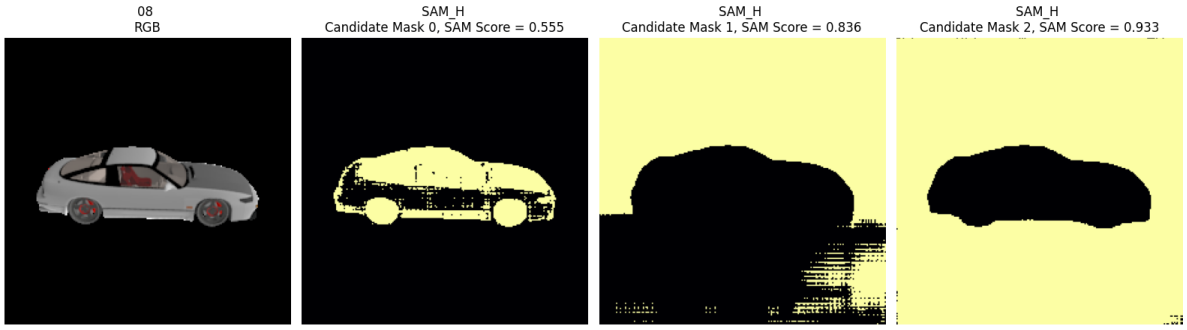


Figure 20: Edge operators applied to estimated depth maps. From left to right: Reference RGB, Estimated Depth, Canny, Sobel, Scharr, and Prewitt outputs.

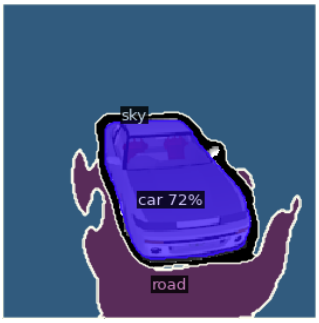
Canny produces binary edges that often fragment under the noise inherent in monocular depth estimation. While these operators successfully identify major structural outlines, we found that they are at times sensitive to artifacts in the estimated depth maps. Consequently, we determined that explicit edge channels added insufficient benefits compared to other priors, and were excluded.

2.3.3 Segmentation and Salient Object Detection Exploration

Separating pixels belonging to the foreground object, through a segmentation mask or, as will be detailed below, using a salient object detection (SOD) model, can be another prior. This involves producing a binary mask that separates an object from its background.
Initially, we explored standard semantic and panoptic segmentation models, such as those found in the Detectron2 [34] model zoo, and the Segment Anything Model (SAM) [35]. These models are often used for segmentation, but as illustrated in Figure 21, these produced non-contiguous masks that often had sections that included more background pixels. Segmentation models are also limited on their training classes, and despite being tested on categories in this set, their masks were improved on by salient object detection models.



(a) SAM Results (3 candidate masks produced per input)



(b) Panoptic Segmentation (Detectron2 Model) Result

Figure 21: Sample outputs from standard segmentation approaches.

SOD models identify the most visually distinct object in a scene, which allows producing a binary mask that tightly hugs the object boundary. To quantitatively evaluate SOD models, we noted that the ShapeNet images used (the same as in the normal and depth priors section) had transparent backgrounds, allowing using the alpha channel to be used as the 'ground truth' for the object silhouette.
We tested three SOD architectures: `rembg` (based on the U-2-Net architecture) [36], `InSPyReNet` [37], and `BiRefNet` [38]. We evaluated performance using Mean Absolute Error (MAE), Intersection over Union (IoU), and F_β-Measure. F_β is a weighted harmonic mean of precision and recall, defined as:

$$F_\beta = \frac{(1+\beta^2)\,\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{3}$$

It is commonly used is salient object detection to assess the quality of binary masks produced. When $\beta = 1$, precision and recall are equally weighted. Greater values of β prioritise recall, while lower ones prioritise pre-

cision. We set β^2 to 0.3, following conventional practice in SOD literature, to emphasise precision over recall [39], [40], [41]. When identifying a single salient object, false positives (background pixels incorrectly classified as foreground) are often considered worse than small false negative sections along the boundary of the object.

Note that the ShapeNet images used are split into 'Easy' and 'Hard' as categories, as before.

Table 2: Comparison of Salient Object Detection models on ShapeNet renders. ↑ indicates higher is better, ↓ indicates lower is better.

Difficulty	**Model**	**IoU ↑**	F_β ↑	**MAE ↓**
Easy	rembg (U-2-Net)	**0.986**	0.991	**0.004**
	InSPyReNet	0.983	**0.996**	**0.004**
	BiRefNet	0.966	0.979	0.006
Hard	rembg (U-2-Net)	**0.980**	0.988	**0.005**
	InSPyReNet	0.966	**0.991**	0.006
	BiRefNet	0.952	0.973	0.007

Figure 22 illustrates the distribution of F_β scores across both "Easy" and "Hard" datasets. `InSPyReNet` has the tightest interquartile range, particularly on the Easy set. `rembg` demonstrates similar stability but with a slightly broader spread on the 'Hard ' dataset. `BiRefNet` is competitive (note the scale of the y-axis, with all achieving scores greater than 0.95), but comparatively shows a lower median score and higher variance, suggesting it is more sensitive to specific geometries.

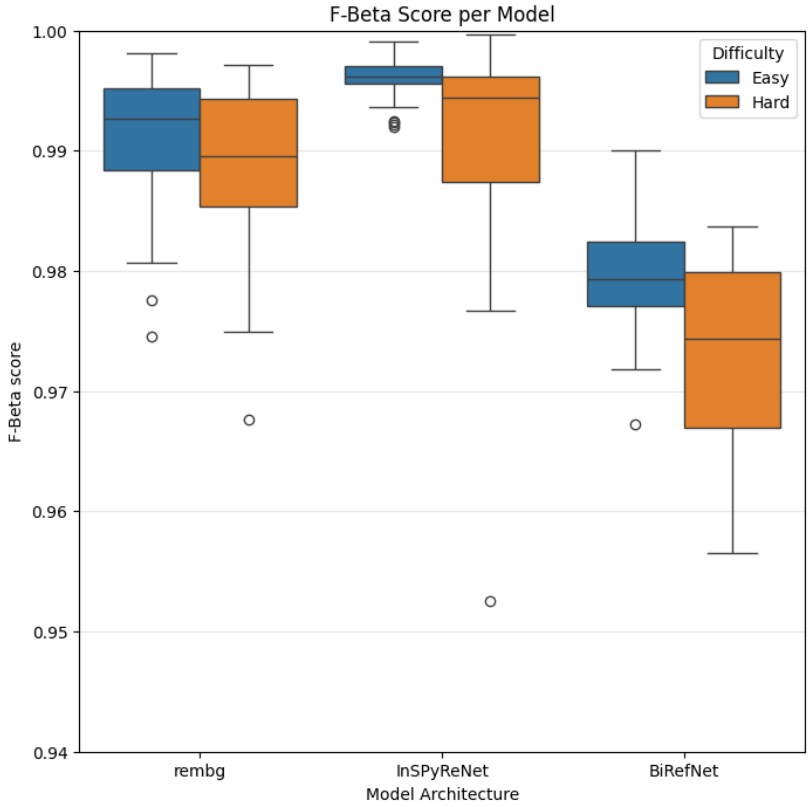


Figure 22: Distribution of F_β scores for each model across the two difficulty levels.

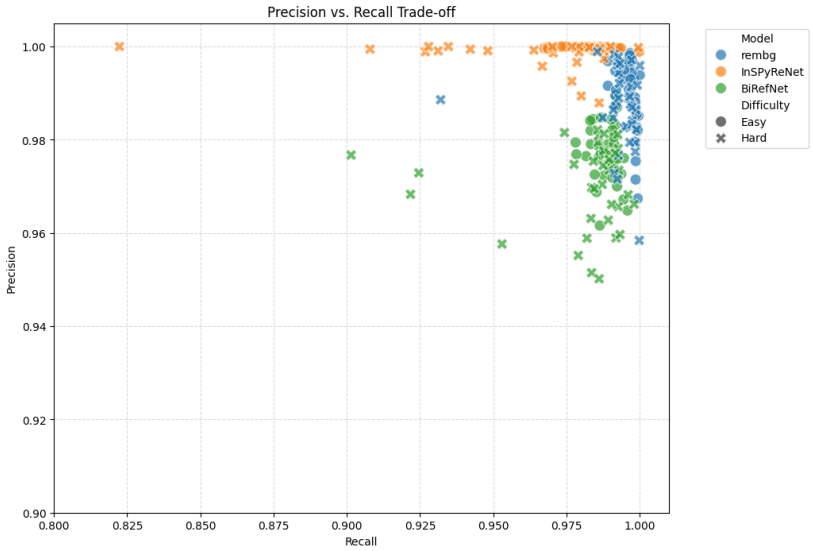


Figure 23: Precision vs. Recall trade-off

The scatter plot in Figure 23 shows how the models operate with different aims; `InSPyReNet` (in orange) clusters tightly towards the top of the high-performance region (top right), suggesting a priority of precision (often very close to 1.0). In practice, this may mean eliminating some background noise, but this may mean missing parts of the object. `rembg` (blue) also has high recall and precision, and it could also be used as a robust general-purpose model for segmentation (that does not clip parts of the object of interest).
To identify the most frequent "winner," we counted the number of images where each model achieved the highest F_β score. `InSPyReNet` achieved the highest score on the majority (28 vs `rembg`'s 8).
However, to contextualise these results, all three models achieved very high scores (with $F_\beta > 0.97$ and IOU > 0.95), largely as the ShapeNet images have a clear foreground object against a uniform background. This reduces the number of possible cases of background vs foreground confusion, which means the models differ meaningfully here on fine structures at the object boundary. In practice, these can include wing mirrors, antennae, etc. Therefore, the differences in the models come from how well they capture these fine details.

2.3.4 Selected Prior Integration

During model selection, basic notebooks were written to operate and evaluate the relevant models. Once a model was selected, the corresponding notebook would be refactored into a high-performance script designed to process full datasets, producing ready-processed priors for every RGB image within the given dataset. Specifically, inference code would be rewritten to ensure it operates on high performance hardware such as the GPU, and that all operations were executed in a batched manner.

```
with torch.no_grad():
    for batch_imgs, batch_filenames in loader:
        # Prediction, B C H W format
        batch_imgs = batch_imgs.to(device)
        preds = model(batch_imgs)

        # Batched downsize
        preds = torch.nn.functional.interpolate(
            preds.unsqueeze(1),
            size=target_size,
            mode="bicubic",
            align_corners=False
        ).squeeze(1)

        # Batched 0 to 1 normalization
        batch_flat = preds.flatten(start_dim=1)
```

```
        min_val, max_val = torch.aminmax(batch_flat, dim=1, keepdim=True)
        min_val = min_val.view(preds.size(0), 1, 1)
        max_val = max_val.view(preds.size(0), 1, 1)
        preds_normalized = (preds - min_val) / (max_val - min_val + 1e-8)
        preds_uint8 = preds_normalized.mul(255).byte().cpu().numpy()

        # H W format -> 128 * 128
        for j, file_id in enumerate(batch_filenames):
            full_batch.append(ProcessedImage(
                uuid=uuid,
                file_id=file_id.item(),
                image=preds_uint8[j].tobytes()
            ))
```

Listing 1: Batched processing of depths

An example of a performant rewrite is that of the depth generation python script. As shown in the excerpt, inference operates on batches supplied by a DataLoader, with all other operations also being executed batch-wise on the device (the GPU in our case); only once the batch is fully processed is it moved back to regular (host) memory, and the individual priors extracted with relevant metadata for use/saving. We also note that predicted depths are quantized to 8-bit unsigned integers, all predicted priors are quantized this way; for example normals are quantized from 3 * 32-bit floats to 3 * 8-bit unsigned integers; this is due to compute and storage limitations. For example the SRN cars dataset contain 387,956 images, each 128 * 128 pixels, storing only priors such as depths and normals as 32-bit floats for each image would require:

$$\begin{aligned} \textbf{Depths:} \quad & 387,956 \times 128 \times 128 \times 4 \text{ bytes} \\ & = 25,425,084,416 \text{ bytes} \approx \textbf{25.43} \text{ GB} \end{aligned}$$

$$\begin{aligned} \textbf{Normals:} \quad & 387,956 \times 128 \times 128 \times 3 \times 4 \text{ bytes} \\ & = 76,275,253,248 \text{ bytes} \approx \textbf{76.28} \text{ GB} \end{aligned}$$

This makes storing the dataset with calculated priors impractical, be it in memory or disk, quantization allows us to cut these requirements down to a quarter of the original size.

Prior generation scripts were succesfully implemented for:

- Depth
- Surface Normals
- Segmentation

All prior generation scripts can be found in the `/geometry-priors` folder within the 3DGS-priors repository.

As to improve training and evaluation performance, we chose to generate priors in advance for selected datasets. As such we developed a pipeline that executes the prior generation scripts and constructs a ready-to-use dataset, alongside a custom Torch Dataset class that can read said dataset. Implementation details can be found in **Section 2.4: Data pipelines**.

2.3.5 Model Changes

The first change to the model was to have the top-level `GaussianSplatPredictor` class dynamically calculate the number of input channels required based on the training configuration, this information was then passed to the underlying Convolutional layers and UNet blocks during initialisation.

```
def calc_channels(cfg):
    # Base RGB channels
    in_channels = 3
```

```
    # Older configs may not have relvant options, select() returns None if the option is
    missing
    if OmegaConf.select(cfg, "data.use_pred_depth") is True:
        in_channels += 1
    if OmegaConf.select(cfg, "data.use_pred_normal") is True:
        in_channels += 3

    return in_channels
```

Listing 2: Channel calculation code

```
# Calculate number of input channels
self.in_channels = calc_channels(cfg)

# Initialise correct model depending on if Gaussian mean offsets are to be calculated
if cfg.model.network_with_offset:
    split_dimensions, scale_inits, bias_inits = self.get_splits_and_inits(True, cfg)
    self.network_with_offset = networkCallBack(cfg,
                                cfg.model.name,
                                self.in_channels,
                                split_dimensions,
                                scale = scale_inits,
                                bias = bias_inits)
    assert not cfg.model.network_without_offset, "Can only have one network"
if cfg.model.network_without_offset:
    split_dimensions, scale_inits, bias_inits = self.get_splits_and_inits(False, cfg)
    self.network_wo_offset = networkCallBack(cfg,
                                cfg.model.name,
                                split_dimensions,
                                scale = scale_inits,
                                bias = bias_inits)
    assert not cfg.model.network_with_offset, "Can only have one network"
```

Listing 3: New model initialisation code

added film layer for text input

Due to time and compute constraints, combined with the selected segmentation model not producing classification or instance labels, we did not implement the cross-attention mechanism for multimodal data input into the model, leaving it as an option for future development.

One of our desired features was to allow a model with new priors to be fine-tuned using existing weights, namely those from the pretrained Splatter Image models. This was achieved by grafting the old weights of a pretrained model onto a new instance of a GaussianSplatPredictor. This grafting mechanism was implemented externally to the model, as such its implementation details are covered in **Section 2.3.7: Training and Evaluation Changes**.

2.3.6 Manual LoRA Integration

To adapt Splatter Image for fine-tuning on the ShapeNet-SRN Cars dataset (with depth and normal priors), we added Low-Rank Adaptation (LoRA) [21] into the GaussianPredictor module of Splatter Image. LoRA adds low-rank trainable matrices into frozen, pretrained layers. This preserves the base model, while further fine-tuning it to the new data.
Existing libraries such as Hugging Face's PEFT [42] provide out-of-the-box LoRA integration, we found them incompatible with the weight-grafting mechanism required by our SplatterImage training pipeline. Due to compute limitations, we rely on Splatter Image grafting weights from pretrained checkpoints where input channel dimensions change (e.g., when adding depth or normal map channels). This leads to shape mismatches during initialisation steps (such as for PEFT's LoRA layers) since the pre-trained weights cannot

be directly loaded into layers with modified input dimensions.
We added LoRA manually within the `GaussianPredictor` and `train_network` modules, using code from Microsoft's `loralib` [22]. We modified the underlying `Linear` and `Conv2d` layers of the U-Net architecture to include a `LoRALayer` mixin. This allows the freezing of pre-trained base weights while injecting the LoRA matrices (A and B) directly into the forward pass. As mentioned earlier, this preserves the frozen pretrained weights from the original 3-channel RGB model, but is compatible with the new channels.

2.3.7 Training and Evaluation Changes

The existing Splatter Image evaluation script needed minimal changes, thanks to how we integrated our priors into the project. The training script initially needed minimal changes, but then was split into two versions, one with minimal changes for standard training, and one with additional changes required for LoRA support.

For evaluation code, the first change was adding support for loading our pretrained models available on HuggingFace.

```
# Load pretrained model from HuggingFace if no local model specified
if args.experiment_path is None:
    # Eval run on the our new dataset with priors
    if dataset_name in ["cars_priors"]:
        cfg_path = hf_hub_download(repo_id="MVP-Group-Project/splatter-image-priors",
                                   filename="model-depth-normal/config.yaml")
        model_path = hf_hub_download(repo_id="MVP-Group-Project/splatter-image-priors",
                                filename="model-depth-normal/model_best.pth")

    # Eval run on previous Splatter Image datasets
    else:
        cfg_path = hf_hub_download(repo_id="szymanowiczs/splatter-image-v1",
                                   filename="config_{}.yaml".format(dataset_name))
        if dataset_name in ["gso", "objaverse"]:
            model_name = "latest"
        else:
            model_name = dataset_name
        model_path = hf_hub_download(repo_id="szymanowiczs/splatter-image-v1",
                                filename="model_{}.pth".format(model_name))

else:
    cfg_path = os.path.join(experiment_path, ".hydra", "config.yaml")
    model_path = os.path.join(experiment_path, "model_latest.pth")

# load cfg
training_cfg = OmegaConf.load(cfg_path)
```

Listing 4: New model loading code

Input preparation was also changed, namely priors are concatenated (if enabled) as additional channels to the input RGB images, before being fed into the network. Splatter Image uses the PyTorch DataLoader alongside custom Dataset classes to load batches of RGB images to perform inference on. The Dataset classes do not directly return a tensor of images, but a dictionary containing relevant batch data:

```
images_and_camera_poses = {
    "gt_images": self.all_rgbs[example_id][frame_idxs].clone(),
    "world_view_transforms": self.all_world_view_transforms[example_id][frame_idxs],
    "view_to_world_transforms": self.all_view_to_world_transforms[example_id][frame_idxs],
    "full_proj_transforms": self.all_full_proj_transforms[example_id][frame_idxs],
    "camera_centers": self.all_camera_centers[example_id][frame_idxs]
}

images_and_camera_poses = self.make_poses_relative_to_first(images_and_camera_poses)
```

```
images_and_camera_poses["source_cv2wT_quat"] = self.get_source_cw2wT(images_and_camera_poses
    ["view_to_world_transforms"])

return images_and_camera_poses
```

Listing 5: Excerpt of srn.py Dataset code

This allowed our priors to be introduced into the evaluation code with ease, by simply creating a new custom Torch Dataset class in `srn_priors.py` that provides priors as new key/value pairs in the returned dictionary. Priors were specifically returned as PyTorch tensors matching the RGB image batch. The priors can then be accessed and concatenated with the RGB images in a specific order to generate the final input tensor for Splatter Image to perform inference on. Assertions are also performed to verify that priors are indeed available for the currently active Dataset.

```
# Concatenate selected priors
input_images = data["gt_images"][:, :model_cfg.data.input_images, ...]
if "use_pred_depth" in model_cfg.data and model_cfg.data.use_pred_depth:
    assert model_cfg.data.category == "cars_priors", "Dataset does not have predicated maps!
    "
    input_images = torch.cat([input_images,
                    data["pred_depths"][:, :model_cfg.data.input_images, ...]],
                    dim=2)
if "use_pred_normal" in model_cfg.data and model_cfg.data.use_pred_normal:
    assert model_cfg.data.category == "cars_priors", "Dataset does not have predicated maps!
    "
    input_images = torch.cat([input_images,
                    data["pred_normals"][:, :model_cfg.data.input_images, ...]],
                    dim=2)

# Get camera to center depth
if model_cfg.data.origin_distances:
    input_images = torch.cat([input_images,
                    data["origin_distances"][:, :model_cfg.data.input_images, ...]],
                    dim=2)
```

Listing 6: Splatter Image input tensor preparation code

For the training code, two changes had to be made. The first change was to model loading, adding support for training using existing HuggingFace model weights from the base Splatter Image project.

```
# Resume from HuggingFace pretrained weights, perform weight graft if now training with
    additional priors
elif cfg.opt.pretrained_hf:
    category = cfg.data.category
    if category == "cars_priors":
        category = "cars"

    model_name = category
    if cfg.data.category in ["gso", "objaverse"]:
        model_name = "latest"

    cfg_path = hf_hub_download(repo_id="szymanowiczs/splatter-image-v1", filename="config_
    {}.yaml".format(category))
    model_path = hf_hub_download(repo_id="szymanowiczs/splatter-image-v1", filename="model_
    {}.pth".format(model_name))
    old_cfg = OmegaConf.load(cfg_path)
    assert is_base_model(old_cfg)

    checkpoint = torch.load(model_path, map_location=device, weights_only=False)

    # Check if new model uses priors
    if is_base_model(cfg):
        try:
```

```
            gaussian_predictor.load_state_dict(checkpoint["model_state_dict"])
        except RuntimeError:
            gaussian_predictor.load_state_dict(checkpoint["model_state_dict"], strict=False)
    else:
        gaussian_predictor = graft_weights_with_channel_expansion(checkpoint["model_state_dict"], gaussian_predictor, old_cfg, cfg)
        print("Grafting performed successfully")

    best_PSNR = checkpoint["best_PSNR"]
    print('Loaded model from a pretrained Huggingface checkpoint')
    OmegaConf.save(config=cfg, f=os.path.join(vis_dir, "config.yaml"))
```

Listing 7: New training setup code

The program first downloads the appropriate base splatter image configuration and model files. The current training configuration is then checked to see if it uses any priors, with `is_base_model` returning true if no priors are to be used. In the case priors have been requested, the new gaussian_predictor instance will have mismatched layer dimensions compared to the pretrained Splatter Image model, meaning the pretrained weights cannot be directly loaded into gaussian_predictor, in this case weights are copied manually into a state dictionary with layers of correct dimensions, we refer to this process as grafting, and is achieved using the `graft_weights_with_channel_expansion` function in `prior_utils.py`.

```
def graft_weights_with_channel_expansion(old_state_dict, new_model, old_cfg, new_cfg):
    new_state_dict = new_model.state_dict()

    # Iterate over all layers
    for name, new_param in new_state_dict.items():
        if name not in old_state_dict:
            # New LoRA parameters not in base model checkpoint. Skip to avoid KeyError.
            if "lora_" in name:
                continue
            print("Failed to find layer {} in HuggingFace model state_dict".format(name))
            raise Exception("Mismatched source model for graft")

        old_param = old_state_dict[name]

        # Directly copy tensors if matching in size (handles most layers)
        if (new_param.shape == old_param.shape):
            new_state_dict[name] = old_param.clone()
            continue

        # In theory we should only reach here for Conv2D layers, as such only need to handle weights, and these should only have extra channels in shape[1]
        if ('weight' in name):
            # Dimension check for Conv2D weights
            if new_param.dim() == 4 and old_param.dim() == 4:
                assert new_param.shape[0] == old_param.shape[0], "Grafting only supported for adding channels, not changing resolution"
                assert new_param.shape[1] > old_param.shape[1], "Cannot truncate channels during graft, can only add channels"

                new_weights = new_param.clone()
                new_weights[:, :old_param.shape[1], :, :] = old_param
                new_weights[:, old_param.shape[1]:, :, :] = 0.0

                new_state_dict[name] = new_weights
            else:
                 print(f"Warning: Skipping graft for {name} due to dimension mismatch")
        else:
            raise Exception("Failed layer graft")

    new_model.load_state_dict(new_state_dict)
    return new_model
```

Listing 8: Grafting code

Grafting works by iterating over the newly configured model's state dictionary, looking up the equivalent layers in the pretrained model's state dictionary. If equivalent layers match in shape, that means no changes have been made and the tensor from the pretrained model's layer is copied directly into the new state dictionary. If there is a mismatch in shape however, we create a new tensor matching the new model layer's shape and manually copy the values from the pretrained layer into the appropriate lower dimensions of the tensor, then zero-initialize all remaining elements in the tensor. By zero-initializing the newly added parameters, the expanded layer remains mathematically equivalent to the original. This ensures the new model's output is identical to that of the pretrained one, despite the change in architecture. The newly configured model is reloaded with the newly generated state dictionary and finally the ready model is returned. This grafting mechanism works despite the behaviour of the newly configured model remaining the same, as back-propagation will compute non-zero gradients for the new inputs channels, allowing the model to update the zero-initialized parameters and gradually integrate the new channels, making them useful.

The second change to training code was similar to that of the evaluation script, updating input preparation, the resulting updated code block is identical to that in **Listing 6**.

For the LoRA-specific training script, within the modified `GaussianPredictor`, `requires_grad=False` is explicitly set for all pre-trained backbone weights (see `gaussian_predictor_lora.py`), leaving only the low-rank matrices A and B as trainable. This ensures that optimizer updates are restricted to the adapter layers, with instances defined as `LoRALayer`.

CODE BLOCK

2.4 Data pipelines

[Explain your data format, how you consume the data in your algorithms, and data augmentation.]

DIAGRAM: SRN FILE STRUCTURE -¿ ORCHESTRATOR -¿ PARALLEL MODELS -¿ PARQUET DIAGRAM: PARQUET -¿ HUGGINGFACE STRUCTURE -¿ HUGGINGFACE

For training and evaluation performance reasons, we chose to generate priors in advance for selected datasets. As such we developed a pipeline that follows an orchestrator pattern; a central notebook initializes a Virtual Machine (VM) for each prior generation model, then installs the necessary dependencies into them, defined in the relevant `PriorName_requirements.txt` files located in the `/geometry-priors` folder. Once the environment is configured, the orchestrator executes the prior generation scripts to produce a complete, ready-to-use dataset.

```
# Create venv for each prior model
!python3 -m venv /content/models/depth --without-pip
!python3 -m venv /content/models/normal --without-pip

# Have to manually install pip to correctly build venvs on colab
!curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
!/content/models/depth/bin/python3 get-pip.py
!/content/models/normal/bin/python3 get-pip.py

# Verify venv's work
!/content/models/depth/bin/pip --version
!/content/models/normal/bin/pip --version

!ls -l /content/models/
```

Listing 9: VM setup

```
# Setup models

```

```
# Depth
!/content/models/depth/bin/pip install -r /content/3dgs-priors/geometry-priors/
    depth_requirements.txt

# Normals
!git clone https://github.com/baegwangbin/surface_normal_uncertainty.git /content/3dgs-
    priors/geometry-priors/surface_normal_uncertainty/
!mkdir /content/3dgs-priors/geometry-priors/surface_normal_uncertainty/checkpoints/
!gdown 1lOgY9sbMRW73qNdJze9bPkM2cmfA8Re- -O /content/3dgs-priors/geometry-priors/
    surface_normal_uncertainty/checkpoints/scannet.pt
!/content/models/normal/bin/pip install -r /content/3dgs-priors/geometry-priors/
    normal_requirements.txt
```

Listing 10: Package setup

```
# Launch models to process dataset
!/content/models/normal/bin/python /content/3dgs-priors/geometry-priors/generate_normal.py
    --in_folder="{in_folder}" --out_folder="{out_folder}" --save_iter=250

!/content/models/depth/bin/python /content/3dgs-priors/geometry-priors/generate_depth.py --
    in_folder="{in_folder}" --out_folder="{out_folder}" --save_iter=250

!python /content/3dgs-priors/geometry-priors/generate_base.py --in_folder="{in_folder}" --
    out_folder="{out_folder}" --save_iter=500
```

Listing 11: Dataset and prior generation script execution

The processing scripts generate `parquet` format files containing the dataset or prior information alongside relevant metadata, such as the image frame_id or depicted model's UUID. Namely, the 5 parquet files currently generated are:

srn_cars_intrins	srn_cars_poses	srn_cars_rgbs	srn_cars_depths	srn_cars_normals
id : int	oid : uuid	sku : str	ts : time	ver : float
name : string	uid : int	cost : float	err : str	auth : str
email : string	amt : float	qty : int	code : int	hash : hex

These files now provide a ready-to-use dataset. To upload the dataset to HuggingFace, additional processing must be performed to transform the files into a file structure that's recognisable by HF for purposes of correctly indexing splits and subsets. Specifically, we transform the dataset to be compatible with this HuggingFace YAML configuration:

```
---
configs:
- config_name: srn_cars_intrins
  data_files:
  - split: train
    path: intrins/train/*.parquet
  - split: test
    path: intrins/test/*.parquet
  - split: val
    path: intrins/val/*.parquet

- config_name: srn_cars_poses
  data_files:
  - split: train
    path: poses/train/*.parquet
  - split: test
    path: poses/test/*.parquet
  - split: val
    path: poses/val/*.parquet

- config_name: srn_cars_rgbs
  data_files:
```

```
  - split: train
    path: rgbs/train/*.parquet
  - split: test
    path: rgbs/test/*.parquet
  - split: val
    path: rgbs/val/*.parquet

- config_name: srn_cars_depths
  data_files:
  - split: train
    path: depths/train/*.parquet
  - split: test
    path: depths/test/*.parquet
  - split: val
    path: depths/val/*.parquet

- config_name: srn_cars_normals
  data_files:
  - split: train
    path: normals/train/*.parquet
  - split: test
    path: normals/test/*.parquet
  - split: val
    path: normals/val/*.parquet
---
```

Listing 12: srn_cars_priors HuggingFace file structure config

```
# Split generated parquets into correct folder/file structure for HuggingFace upload
%cd /content/
!mkdir upload
%cd upload

datatypes = ["rgbs", "intrins", "poses", "depths", "normals"]
splits = ["train", "test", "val"]

for datatype in datatypes:
    # Make a directory for the current dataset file
    !mkdir {datatype}
    %cd {datatype}
    load_path = os.path.join(out_folder, datatype + ".parquet")
    dataset = load_dataset("parquet", data_files=load_path)['train']
    for split in splits:
        # Generate split specific directory
        !mkdir {split}
        save_path = "/content/upload/{}/{}/{}.parquet".format(datatype, split, datatype)
        filter_split = lambda data: data['split'] == split
        filtered = dataset.filter(filter_split).to_pandas().drop(columns=['split'])
        filtered.to_parquet(save_path)
    %cd ..
```

Listing 13: srn_cars_priors HuggingFace file structure config

A diagram depicting this pipeline can be seen at Figure X.

Take a concretne example, we offer one ready pregenerated dataset on HF: We took our input data from the ShapeNet-SRN dataset from [43] at 128×128 resolution. transform pipeline standard, the depth, rgb, normal columns store images as uint8s in C H W format, this is as there is X images, do math, thus this manmy GB of data, we must un fortunately for compute reasons take this loss of accuracy.
Models use the pregenerated datasets via a custom laoder we implemented, taht slots in into all of the exisitng training/evlaution code. AS STATED EARILR SECTION, WE HAVE CUSTOM DATASET FOR LAODER, THAT DOWNLOADS HF DATASET, AND PROCESSES TO STORE AS CORRECT PANDAS DATAFRAME TO RETURN DATA

DURING TRAINING/ATT EMD ARE SAVED FOLLOWING FILES:

THESE ARE UPLAODED THESE WEIGHTS ARE UPLOADED TO HF WITH THIS NAMING SCHEME:

NAMING CODE BLOCK

UPLAOD CODE BLOCK

2.5 Training procedures

[Explain which framework and optimizers you use, how you implemented the training logic.]

THE ENTIRE MODEL IS PROGRAMMED USING TORCH WE USE PEFT OR WHATEVRE TO ADD LORA Layers

OUR TRAINING IS SET BY CONFIG FROM WADB?

SHWO CONFGI

WHEN MODEL TRAINIGN OCCURS CONFIG OPTIOSN ARE VERFIEID TO BE COMPATIABLE OPTIONS INCLUIDE USING LORA vs NOT, WHICH PRIORS TO USE, AND WHETEHR YOU WNAT TO START THE MODEL WITH EXISITNG WEIGHTS (if so grafted) WEIGHTS FUSED AT SAVE

TRAINGING IS RUN USING DEFAULT SPALTTER IAMGE APRAMSTERS, Namely: LIST OPTIMSIER OtHER OPTIOSM

THE AVAIABLE MODCEL WEIGHST ON HF WERE TRAINIG NNTOEBOOK RUSN ON CLOAB WITH A100 ITERS: 60K

In order to isolate the parameter-efficient adaptation due to LoRA, we used a dedicated training script (`train_network_lora.py`). While preserving the data pipeline and training loop architecture as the standard framework, it is necessary to enforce strict freezing where the gradients for the U-Net backbone are disabled. Optimisation is therefore restricted solely to the injected low-rank matrices (as described in **2.2.1 Low-Rank Adaptation (LoRA)**).

2.6 Testing and validation procedures

[Explain which framework you use, how you implemented the testing/ validation logic.]

As discussed in previous sections, our changes often include a variety of runtime assertions to verify correctness, an example can be found in our input preparation changes, where assertions are performed to check if every prior requested for training is available in the training/test/validation datasets requested.

```
if "use_pred_depth" in model_cfg.data and model_cfg.data.use_pred_depth:
    assert model_cfg.data.category == "cars_priors", "Dataset does not have predicated maps!
    "
    input_images = torch.cat([input_images,
                    data["pred_depths"][:, :model_cfg.data.input_images, ...]],
                    dim=2)
if "use_pred_normal" in model_cfg.data and model_cfg.data.use_pred_normal:
    assert model_cfg.data.category == "cars_priors", "Dataset does not have predicated maps!
    "
    input_images = torch.cat([input_images,
                    data["pred_normals"][:, :model_cfg.data.input_images, ...]],
                    dim=2)
```

Listing 14: Excerpt from Splatter Image input tensor preparation code

Notebooks performing unit and integration tests were additionally written. Namely, 3 such notebooks can be found in the `testing` folder of the `3DGS-priors` repository.

The `eval_test.ipynb` notebook performs tests to verify if the modified evaluation script successfully performs evaluation both on previous datasets (such as ShapeNet cars) and our new dataset `cars_priors`. It does so by running correctly configured evaluations and checking whether the resulting `scortes.txt` files

exist.

The `graft_test.ipynb` notebook performs tests checking whether the `graft_weights_with_channel_expansion` function returns successfully, for a variety of model configurations. It additionally verifies whether an automatically grafted model's state dictionary exactly matches the state dictionary of a manually grafted reference model.

The `dataloader_test.ipynb` notebook tests our custom Dataset and ready generated `cars_priors` data. It does so by loading both the previous reference `srn.py` Dataset and our new `srn_priors.py` Dataset, walking both in an ordered mannered using Torch DataLoaders with shuffling disabled, and then comparing the resulting batches, which in our case means comparing common dictionary entries, where common key/values entries should be identical.

All tests in these notebooks were passed successfully; each notebook should have visible cell outputs showing this.

Chapter 3: Experiments and Evaluation

3.1 Datasets

[Explain the datasets utilized: what they contain, why they are utilized, assumptions, limitations, possible extensions.]

The standard benchmark for evaluating single-view 3D reconstruction is ShapeNet-SRN [43], hence we used this to test and evaluate our main model implementation. For this dataset, we specifically use the "Car" class, which used the "car" class of ShapeNet v2 [25] with 2.5k 3D CAD model instances. The SRN dataset was generated by disabling transparencies and specularities and training on 50 observations of each instance at a resolution of 128×128 pixels, with camera poses being randomly generated on a sphere with the object at the origin. A limitation of this dataset is the lack of subject variety in the dataset as the model may end up overfitting to cars. A possible extension to address this limitation could be to include other classes in the ShapeNet-SRN database to make sure that the model can still generalise to other types of objects.

An extension of this dataset is implemented in [24], which presents a Deep Implicit Surface Network to generate a 3D mesh from a 2D image by predicting the underlying signed distance fields. In the paper, they generated a 2D dataset composed of renderings of the models in ShapeNet Core [25]. For each mesh model, the dataset provides 36 renderings with smaller variation and 36 views with larger variation (bigger yaw angle range and larger distance variation). The object is allowed to move away from the origin, which provides more degrees of freedom in terms of camera parameters, and the "roll" angle of the camera is ignored since it was deemed very rare in real-world scenarios. The images were rendered at a higher resolution of 224×224 pixels and were paired with a depth image, a normal map and an albedo image as shown in figure 11. This dataset was mainly used as a ground truth to evaluate the generation of geometry priors (e.g. normal map and depth map). A limitation of this dataset would be its small size since only 72 samples are available for us to use, such that the performance of geometry prior generation may not be evaluated correctly. However, in the same GitHub repository, the script to generate these images from the ShapeNet Core dataset is provided, so a possible extension given more time could be to include more images by running the script on other objects in the ShapeNet Core dataset. Another limitation is that some of the depth maps provided were flawed as shown in figure 17, so using them as ground truths to evaluate the performance of the depth map generator model was not ideal. In the future, evaluation could benefit from regenerating these depth maps to be more accurate, perhaps by exploring a different method of depth map generation from the 3D model provided in the ShapeNet Core dataset.

3.2 Training and testing results

[Explain the training and testing results with graphs and elaborating on why they make sense, what could be improved.]

3.3 Qualitative results

[Show in figures and explain visual results. Include different interesting cases covering different aspects/ limitations/ dataset diversity. If not converged, explain what we can expect once converged. Include any other didactic examples here.]

(a) Ground Truth (b) LoRA Prediction)

Figure 24: Comparison of novel view case. (a) Ground truth rendering of the target view. (b) Reconstruction by a LoRA-adapted model.

One example of evaluation output can be seen in 24, where LoRA adapters were trained with a rank of 2, α of 2 (both low compared to standards) and dropout of 0.05. The LoRA-adapted model is able to reconstruct the car from a novel view while preserving details such as headlights, despite the low parameter overhead.

3.4 [Optional] Quantitative results

[A table and associated explanations for quantitative results.]

For novel view synthesis evaluation, PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index) and LPIPS (Learned Perceptual Image Patch Similarity) are standard metrics used to evaluate image quality from different perspectives[9]. PSNR is calculated using the mean squared error and is commonly used to quantify reconstruction quality, such that a high PSNR suggests higher reconstruction quality. Meanwhile, SSIM depends on 3 metrics mimicking the human visual perception system: luminance, contrast and structure, where an SSIM score close to 1 indicates high similarity while a score closer to -1 indicates low similarity. Lastly, unlike pixel-wise metrics like PSNR and SSIM that assume pixel independence, LPIPS measures the perceptual similarity between images by comparing their features extracted from a deep neural network, where a low LPIPS score indicates that the compared images are perceptually similar to humans.

To evaluate whether integrating different geometry priors into the model is effective, we performed a set of experiments testing different combinations of geometry priors. We measured the performance using PSNR, SSIM and LPIPS.

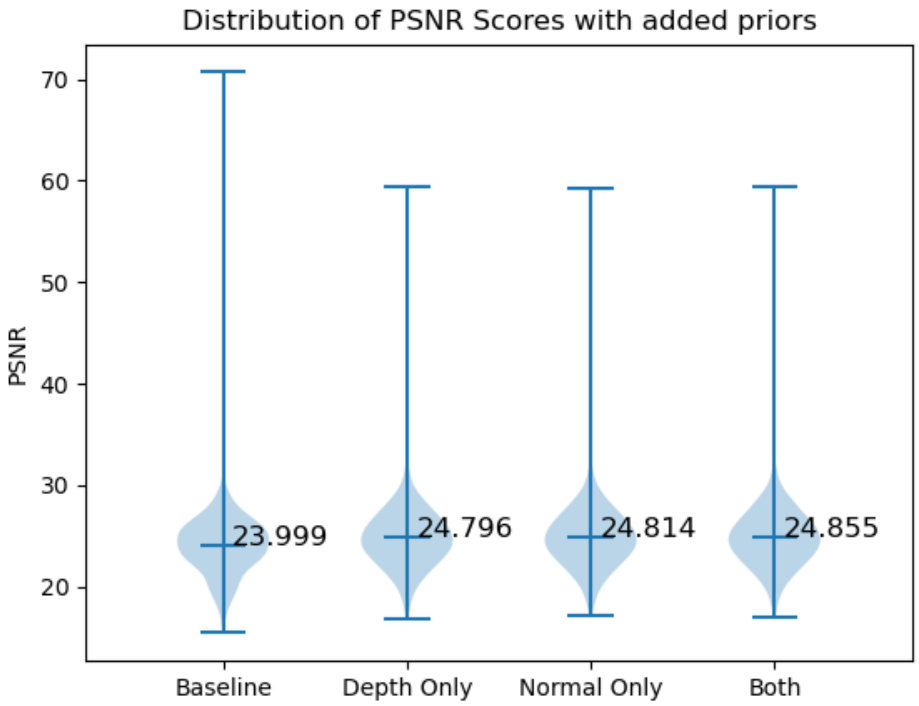


Figure 25: PSNR distribution for different combinations of added geometry priors

Figure 26: SSIM distribution for different combinations of added geometry priors

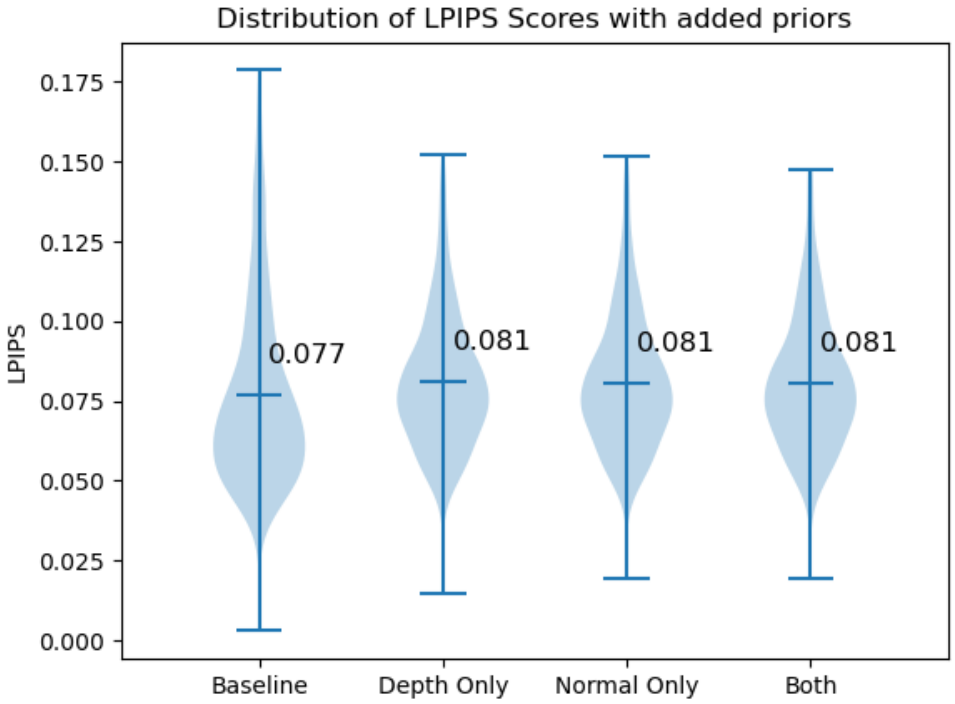


Figure 27: LPIPS distribution for different combinations of added geometry priors

Table 3: Significance testing for metrics of augmented models vs baseline (to 3 s.f.)

Added Geometry Priors	**Metric**	**P-value**
Depth Only	PSNR	0.000247
	SSIM	0.000273
	LPIPS	0.0311
Normals Only	PSNR	0.000247
	SSIM	0.000187
	LPIPS	0.0369
Both	PSNR	0.000100
	SSIM	0.000100
	LPIPS	0.0385

We conduct a two-sided significance test with non-parametric resampling to calculate the P-value, which is the probability that the metrics of the augmented model are not significantly different from that of the original model and any difference is only due to random variation. We select the power of the test to be 0.05. Since all P-values are less than 0.05 from table 3, we conclude that the metrics of augmented model significantly differ from that of the original model.

Adding combinations of depth and normal priors to augment the model with geometry priors seems to make the model perform more consistently, as shown by the lower variance in all metrics as compared to the baseline in figures 25, 26 and 27. All augmented models have a higher mean PSNR and SSIM from figures 25 and 26 respectively which suggests better image reconstruction quality, with the model adding both depth and normal priors performing the best. However, all augmented models have a higher LPIPS score than the baseline 27, suggesting that the difference between the ground truth and the images generated by the model

perceived by humans are larger for augmented models than the baseline. We conclude that integrating these priors may not be beneficial in some cases, for example cases where successful image reconstruction is defined as being similar to the ground truth based on a human's perception.

3.4.1 LoRA experiment results

To evaluate whether Low-Rank Adaptation (as described in **Section 2.3.5: Model Changes**) is effective, we performed a set of finetuning experiments. LoRA behaviour can be adjusted through three hyperparameters: rank, alpha and dropout (as mentioned in section 2.2.2).

We tested multiple configurations by varying these hyperparameters, while keeping all other components. We measured the performance using PSNR, SSIM and LPIPS.

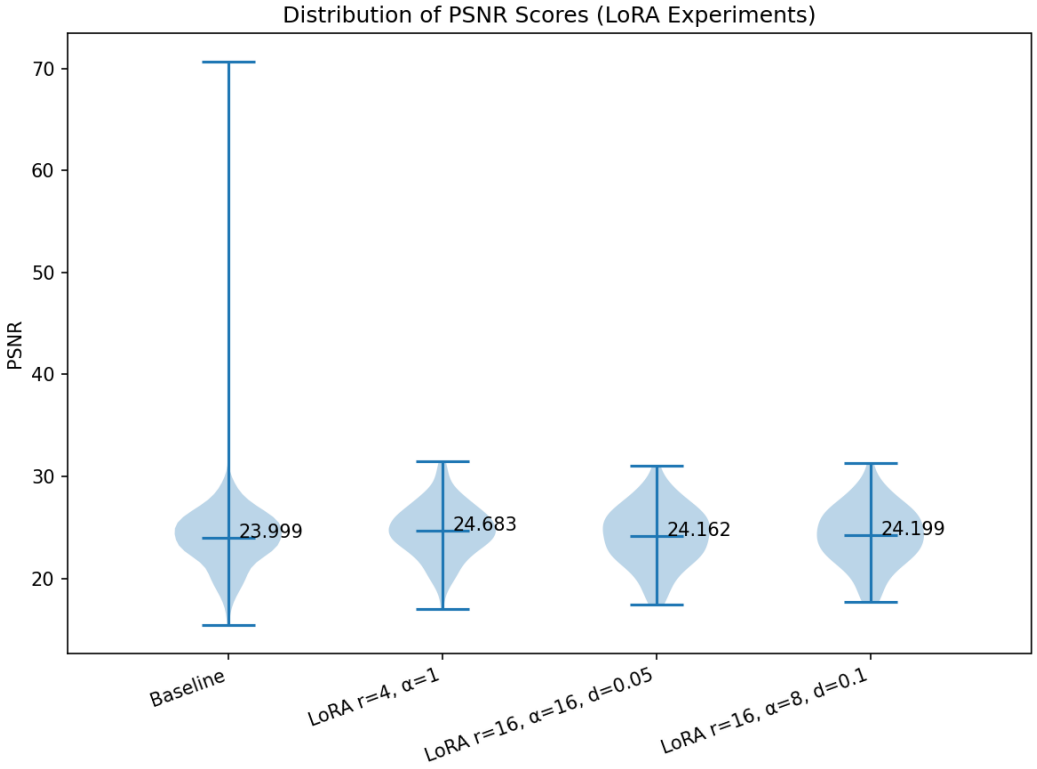


Figure 28: PSNR distribution for different LoRA configurations

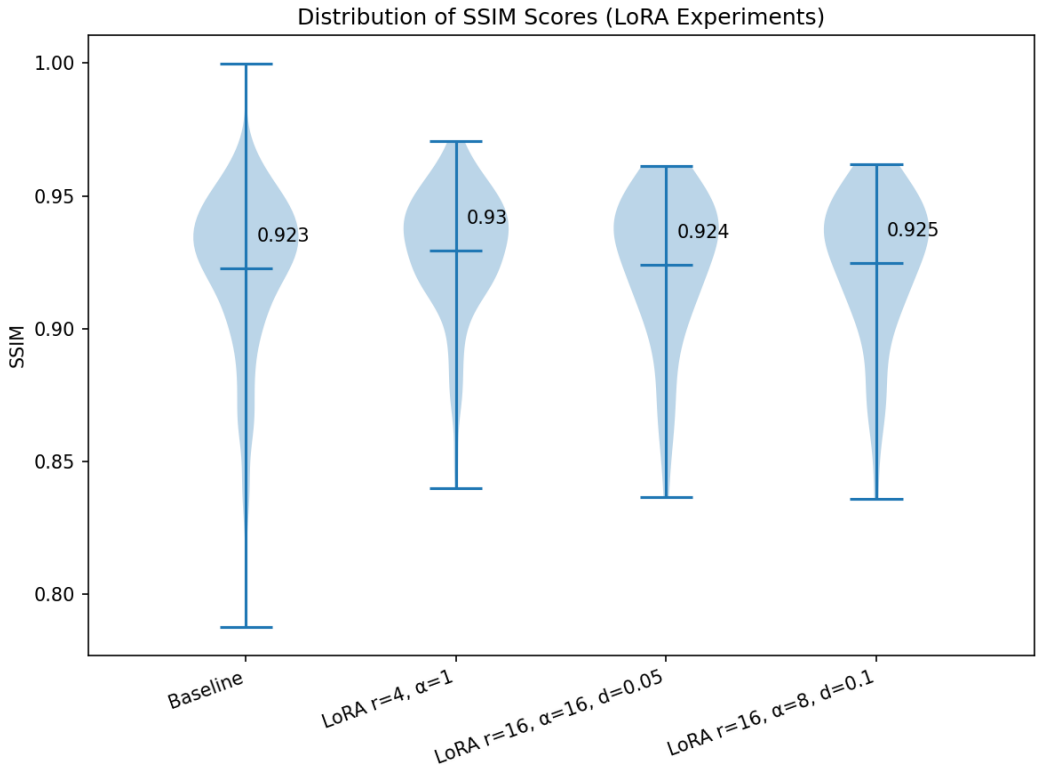


Figure 29: SSIM distribution for different LoRA configurations

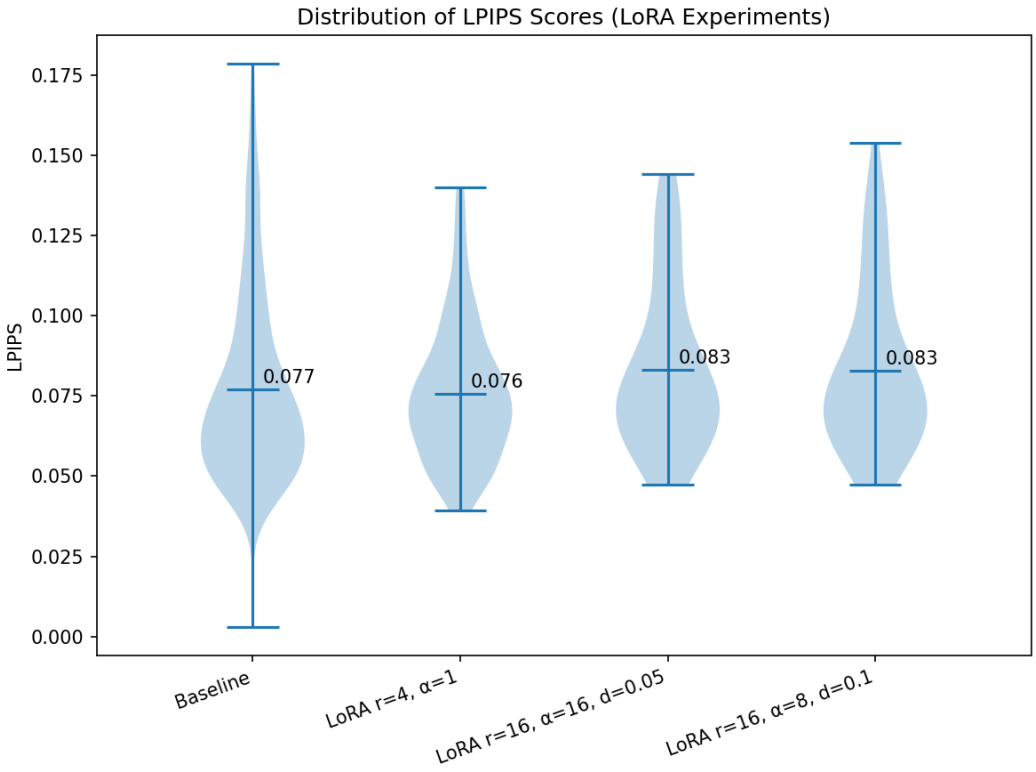


Figure 30: LPIPS distribution for different LoRA configurations

From figures 28, 29 and 30, we can see that all changed models now perform more consistently, as shown by the lower variance in all metrics as compared to the baseline. The PSNR metrics in figure 28 and the SSIM metrics in figure 29 also suggest better image reconstruction overall due to having higher mean PSNRs and SSIMs than the baseline, with the configuration $r = 4, \alpha = 1$ performing the best with the highest mean PSNR and SSIM. This configuration also has a lower mean LPIPS score than the baseline, further supporting our claim that this configuration is best for augmenting the model, while the other 2 configurations have higher mean LPIPS scores.

(Placeholder LoRA results comparison with SplatterImage) We can compare our LoRA results with the original Splatter Image paper [4].

Chapter 4: Conclusions and Future Directions

4.1 Conclusions

[Summarize what the project was about and the main conclusions.]

4.2 Discussion of limitations

[Explain the limitations of your technique. You may want to refer to previous sections or show figures on the limitations.]

halluciantion in hidden areas still a problem, as at end of the day you dont know what there, only doign best guess basd on alrge datasets lot of data cnd comptue needed to perofrm accurate reweults we lacked compute to generate and test more models priors

4.3 Future directions

[State a few future directions for research and development. These typically follow from the discussion on limitations.]

generating and testing mroe priors (espoecially more priors aoriented to hidden areas, could bring up planes again) implementing and comapring cross-attention vs FiLM for multimodal data trying longer training times trying alternative datasets Trying alternatigve metrics for 3D reconstruction, like Chaumfeer distance, for exampel could sample off bunch of points of SRN cars meshes to genrate point cloud, then treat gaussians as point cloud and do comaprison

As mentioned in last supervisor meeting: explicit loss function/rendering loss: e.g. depth edges for tinier Gaussians (based on the size of the covariance matrix), could help with fine details.

4.4 Project Contributions

Report Writing Contributions:
Section 1.1: Kacper and Alex
Section 1.2: Kacper and Alex
Section 1.3: Kacper
Section 2.1: Alex
Section 2.2: Kacper and Radhika
Section 2.3: Kacper and Radhika and Alex
Section 2.4: Kacper
Section 2.5: Kacper and Radhika
Section 2.6: Kacper
Section 3.1: Alex
Section 3.2: Radhika
Section 3.3: Radhika
Section 3.4: Radhika and Alex
Section 4.1: FILL IN
Section 4.2: FILL IN
Section 4.3: FILL IN
Image and citation collection: Kacper and Radhika and Alex

Presentation Contributions:
Slides: Alex
Recording: Kacper and Radhika and Alex

Technical Contributions:
Depths exploration: Radhika
Edge operators exploration: Radhika
Segmentation exploration: Radhika
Normals exploration: Alex
Planes exploration: Alex
Splatter Image setup and bugfixes: Radhika and Alex
Splatter Image architectural modification (Grafting, Channel changes, FiLM, Cross-Attention): Kacper
Splatter Image LoRA integration: Radhika
Splatter Image Training Modification: Kacper and Radhika
Splatter Image Eval Modification: Kacper
Optimised depth generation: Kacper
Optimised normals generation: Kacper and Alex
Optimised segmentation generation: Radhika
Splatter Image `cars_priors` custom Dataset: Kacper
HuggingFace dataset data pipeline: Kacper
Results processing code: Alex
Graphing code: Alex

Testing custom Dataset: Kacper
Testing evaluation: Kacper and Radhika
Testing prior configurations: Kacper
Testing LoRA configurations: Radhika

References

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

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023.

[3] Shengji Tang, Weicai Ye, Peng Ye, Weihao Lin, Yang Zhou, Tao Chen, and Wanli Ouyang. Hisplat: Hierarchical 3d gaussian splatting for generalizable sparse-view reconstruction, 2024.

[4] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction, 2024.

[5] Jianghao Shen, Nan Xue, and Tianfu Wu. A pixel is worth more than one 3d gaussians in single-view 3d reconstruction, 2024.

[6] Zhizhong Kang, Juntao Yang, Zhou Yang, and Sai Cheng. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5), 2020.

[7] Andrea Romanoni, Amaël Delaunoy, Marc Pollefeys, and Matteo Matteucci. Automatic 3d reconstruction of manifold meshes via delaunay triangulation and mesh sweeping, 2016.

[8] Timothy Newman and Hong Yi. A survey of the marching cubes algorithm. *Computers & Graphics*, 30:854–879, 10 2006.

[9] Jiahui Zhang, Yuelei Li, Anpei Chen, Muyu Xu, Kunhao Liu, Jianyuan Wang, Xiao-Xiao Long, Hanxue Liang, Zexiang Xu, Hao Su, Christian Theobalt, Christian Rupprecht, Andrea Vedaldi, Kaichen Zhou, Paul Pu Liang, Shijian Lu, and Fangneng Zhan. Advances in feed-forward 3d reconstruction and view synthesis: A survey, 2025.

[10] Tianyi Gong, Boyan Li, Yifei Zhong, and Fangxin Wang. Exscene: Free-view 3d scene reconstruction with gaussian splatting from a single image, 2025.

[11] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N. Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image, 2025.

[12] Yuxin Wang, Qianyi Wu, and Dan Xu. F3d-gaus: Feed-forward 3d-aware generation on imagenet with cycle-aggregative gaussian splatting, 2025.

[13] Junlin Hao, Peiheng Wang, Haoyang Wang, Xinggong Zhang, and Zongming Guo. Gaussvideodreamer: 3d scene generation with video diffusion and inconsistency-aware gaussian splatting, 2025.

[14] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers, 2023.

[15] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting, 2023.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2024.

[17] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks, 2022.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[23] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. *CoRR*, abs/2109.00512, 2021.

[24] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.

[25] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

[26] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *International Conference on Computer Vision (ICCV)*, 2021.

[27] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[28] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. *arXiv preprint arXiv:1903.12305*, 2019.

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[30] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018.

[31] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[32] Saeed Mahmoudpour and Manbae Kim. Chapter 10 - a study on the relationship between depth map quality and stereoscopic image quality using upsampled depth maps. In Leonidas Deligiannidis and Hamid R. Arabnia, editors, *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*, pages 149–160. Morgan Kaufmann, Boston, 2015.

[33] Xinwei Liu, Marius Pedersen, and Renfang Wang. Survey of natural image enhancement techniques: Classification, evaluation, challenges, and perspectives. *Digital Signal Processing*, 127:103547, 2022.

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

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

[36] Daniel Gatis. rembg. `https://github.com/danielgatis/rembg`, 2025. Version 2.0.66.

[37] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124, 2022.

[38] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

[39] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[40] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[41] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, December 2015.

[42] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

[43] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations, 2020.