

Closing the Loop in Scene Reconstruction: Automated Evaluation and Guided Data Collection

Joakim Lindén, joakim.linden@mdu.se ^{*†}, Ludwig Karlsson, ludwig.karlsson2@saabgroup.com [†]
Håkan Forsberg, hakan.forsberg@mdu.se ^{*}, Masoud Daneshtalab, masoud.daneshtalab@mdu.se ^{*}

^{*} Department of Computer Science & Engineering, Mälardalen University, Västerås, Sweden

[†] Saab Aeronautics, Stockholm, Sweden

Abstract—Recent advances in Neural Radiance Fields and 3D Gaussian Splatting have enabled photo-realistic rendering of complex, view-dependent scenes. However, rendering fidelity heavily depends on strategic viewpoint selection, which currently relies on manual intervention or inefficient uniform sampling. This paper introduces an automated, closed-loop framework that pairs a neural quality estimator with a multi-agent route optimization strategy for targeted, scene-dependent data collection.

Our approach integrates two core components: a reference-free perceptual quality network that predicts reconstruction defects without ground-truth images, and a multi-agent path planner formulated as a Vehicle Routing Problem to optimize autonomous resampling trajectories. In an iterative setup, the system reconstructs an initial scene, evaluates local quality to pinpoint artifacts, and dispatches an autonomous fleet to capture targeted supplementary views.

Experimental validations in simulated urban environments show a distinct reduction in Learned Perceptual Image Patch Similarity (LPIPS) error after a single refinement loop. While the route optimisation successfully minimizes operational flight costs, our analysis reveals that naive greedy viewpoint selection suffers from severe spatial redundancy. Conversely, introducing a spatial Non-Maximum Suppression (NMS) layer over predicted error distributions breaks up viewpoint clustering, decisively outperforming standard geometric coverage baselines. This framework establishes a scalable, data-efficient pathway toward autonomous, self-improving scene capture systems.

Index Terms—synthetic data, scene reconstruction, data sampling, gaussian splats

I. INTRODUCTION

Recent methods such as Neural Radiance Field (NeRF) [1] and 3D Gaussian Splats (3DGS) [2] have made 3D modelling possible through data-driven scene reconstruction, allowing view-dependent, photorealistic rendering for training, visualization, and simulation tasks.

A persistent challenge lies in determining which camera viewpoints are required to achieve a desired reconstruction quality. Current NeRF/3DGS pipelines often rely on manually selected or uniformly sampled views, leading to inconsistent coverage and suboptimal results. Recent work such as NeRF Director [3] shows that systematic, data-driven view selection significantly improves scene fidelity, motivating the need for quality-aware principled methods over manual heuristics.

Typically, NeRF/3DGS evaluations rely on full-reference metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural

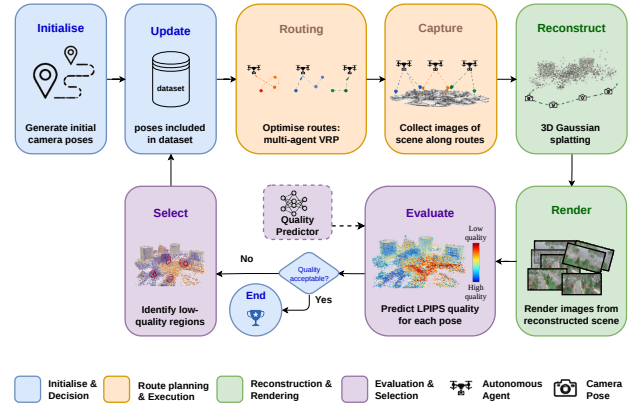


Fig. 1. Overview of the iterative scene refinement loop. The quality predictor identifies low-fidelity regions, the route optimizer plans efficient sampling trajectories, and the collected data is used to refine the 3DGS reconstruction.

Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [4] computed on held-out views [1], [5]. However, these metrics correlate imperfectly with human perception, especially for complex materials and view-dependent effects [6].

Studies show that reconstruction quality strongly depends on viewpoint selection [7]. In the novel view synthesis (NVS) domain, active view planning methods guide data acquisition using information gain objectives such as Fisher Information [8], surrogate quality metrics [9], or coverage-based strategies [10]. Collectively, these methods highlight that quality evaluation and view selection are tightly linked: an effective quality predictor not only assesses reconstruction fidelity but also identifies where additional data is most beneficial.

Recent works [11], [12] use ray entropy minimisation and uncertainty information from the NeRF training to guide view selection, while applications such as data augmentation and aerial mapping [13], [14] demonstrate the growing role of neural reconstruction in automated pipelines.

In parallel, UAV-based inspection and coverage planning has formalized viewpoint selection as a distinct optimisation problem, often decoupled from trajectory planning through two-stage formulations separating view planning from route optimisation [15], [16]. However, Papaioannou et al. [17] incorporated sensing and visibility constraints directly into the planning process, by jointly optimising UAV motion and camera orientation under field-of-view constraints. More broadly,

We acknowledge the support of the Swedish Knowledge Foundation via the RELIANT research school, grant no 20220130, via Sweden's Innovation Agency and the Swedish Foundation for Strategic Research.

perception-aware navigation couples sensing and motion for robust operation in complex environments [18], and multi-UAV systems address scalability through cooperative coverage strategies [19].

Despite these advances, most existing approaches rely on geometric coverage, visibility constraints, or implicitly learned uncertainty measures, and do not explicitly model the perceptual, view-dependent quality of the reconstructed scene. In contrast, as seen in Figure 1, we introduce an automated, closed-loop framework for quality-driven 3DGS scene refinement that directly links reconstruction, evaluation, and data acquisition. The main contributions of this work are:

- 1) a reference-free perceptual quality estimator that predicts quality in reconstructed 3DGS scenes;
- 2) a multi-agent data acquisition route optimiser, formulated as a Vehicle Routing Problem (VRP), for efficient supplementary image capture;
- 3) a closed-loop framework that links 3DGS reconstruction, reference-free quality estimation, pose selection, and multi-agent data collection;
- 4) an empirical analysis demonstrating that purely quality-guided, greedy viewpoint selection suffers from spatial redundancy, and establishing that integrating a 3D Spatial Non-Maximum Suppression (NMS) layer yields reconstruction updates that consistently outperform both random allocation and geometric coverage baselines.

Experiments show that supplementary data collection improves reconstruction quality and that the proposed route optimisation enables efficient multi-agent acquisition. At the same time, our results reveal an important limitation of purely greedy perceptual pose selection: low-quality predictions often cluster around the same scene regions, leading to redundant viewpoints and diminishing returns. Applying NMS in the spatial domain shows significant improvements in scene quality.

This paper is organized as follows: Section II presents our quality-guided scene refinement method and multi-agent route optimisation process. In section III we define the parameters and details of our experiments and section IV reports experimental results and quantitative improvements. Section V discusses the findings, and Section VI concludes with future directions.

II. METHOD

Figure 1 illustrates the proposed closed-loop framework for quality-driven 3D scene refinement. The system operates iteratively, linking reconstruction, evaluation, and data acquisition to progressively enhance scene fidelity. The core of our approach consists of two integrated components: a reference-free perceptual quality estimator and a multi-agent route optimisation strategy. The iterative process is formalized in Algorithm 1.

A. Quality-Guided Scene Refinement

The primary objective is to identify and acquire supplementary views that maximize the improvement of the reconstructed

Algorithm 1 Iterative Quality-Guided Pose Selection and Scene Refinement

Require: Initial poses $\mathcal{P}_0$, candidate poses $\mathcal{C}$, quality predictor Q , number of added poses N , fleet $\mathcal{A}$

Ensure: Refined reconstruction $\mathcal{S}$

```

1:  $\mathcal{P} \leftarrow \mathcal{P}_0$  are the poses to sample scene at,
2:  $\mathcal{D} \leftarrow \emptyset$  is the dataset of sampled images,
3:  $\mathcal{E} \leftarrow \mathcal{C}$  is the set of evaluation poses.
4: repeat
5:   Plan capture routes for  $\mathcal{P}$  using multi-agent VRP
6:   Capture images  $\mathcal{I}(\mathcal{P})$ 
7:   Extend dataset  $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{I}(\mathcal{P})$ 
8:   Reconstruct 3DGS scene  $\mathcal{S}$  from  $\mathcal{D}$ 
9:   Render candidate views from  $\mathcal{S}$  at poses in  $\mathcal{E}$ 
10:  Predict quality for all rendered views using  $Q$ 
11:  Compute scene-level quality estimate
12:  if quality is acceptable then
13:    return  $\mathcal{S}$ 
14:  end if
15:  Sort remaining candidate pool  $\mathcal{C}$  in descending order of
    predicted error scores  $s_c$  and filter with 3D spatial NMS.
16:   $\mathcal{P}_{\text{add}} \leftarrow \text{NMS}(\mathcal{C}, s_c, N, D_{\text{min}})$ 
17:   $\mathcal{P} \leftarrow \mathcal{P}_{\text{add}}$   $\triangleright$  Assign diverse high-error poses to next
    fleet routing cycle
18:   $\mathcal{C} \leftarrow \mathcal{C} \setminus \mathcal{P}_{\text{add}}$ 
19: until quality is acceptable or budget is exhausted
20: return  $\mathcal{S}$ 

```

scene. Unlike traditional pipelines that rely on uniform sampling or manual selection, our method employs an active view planning strategy.

1) *Quality Prediction:* Given an initial 3DGS reconstruction, the system generates a set of novel virtual viewpoints to evaluate the current scene quality. A perceptual quality predictor, trained to estimate reconstruction quality without requiring ground-truth reference images, scores these views. The algorithm ranks the evaluated poses based on predicted quality, identifying the regions of the scene with the highest estimated perceptual error.

2) *Spatial 3D Non-Maximum Suppression:* These found low-quality poses represent regions where the current reconstruction lacks sufficient geometric or textural information. By prioritising the acquisition of images of these specific regions, the framework ensures that new data is collected where it is most needed. Because a single reconstruction artifact yields high rendering errors across many adjacent camera positions, a naive greedy pose selection (e.g., picking the top- N highest error poses) will likely waste the data collection budget on redundant poses. To mitigate this redundancy defect, we introduce a Spatial 3D NMS layer over the quality prediction distribution. Let $\mathcal{C}$ be the pool of candidate camera viewpoints, where each viewpoint $c_i \in \mathcal{C}$ is parameterized by its 3D position $\mathbf{p}_i = [x_i, y_i, z_i]^T \in \mathbb{R}^3$ and its predicted LPIPS error score s_i . The viewpoints are sorted in descending order of error severity ($s_0 \geq s_1 \geq \dots \geq s_M$). We iteratively select

the worst-performing frame c_{max} and suppress all remaining candidate viewpoints c_j whose Euclidean distance falls below a minimum spatial threshold D_{min} :

$$\|\mathbf{p}_{max} - \mathbf{p}_j\|_2 < D_{min}$$

This process is repeated until our target allocation budget of N viewpoints is satisfied. If high spatial density causes the NMS to exhaust the candidate pool of poses before satisfying the target sampling budget (N), a greedy fallback selector gathers the highest-error suppressed poses until the target budget is reached. The selected poses are then added to the training dataset, and the 3DGS model is retrained, resulting in a refined scene with reduced perceptual error. This process can be repeated iteratively until a target quality threshold is met or resources are exhausted.

B. Multi-Agent Route Optimisation

To efficiently collect the identified supplementary images, we formulate the data acquisition problem as a multi-agent Vehicle Routing Problem (VRP). This is particularly relevant for large-scale scenes where a single agent would be inefficient.

Let $\mathcal{P} = \{P_1, \dots, P_M\}$ be the set of target camera poses identified by the quality-based selection process, where each pose P_j is defined by its 3D position and orientation (pitch and yaw). We consider a fleet of K autonomous agents (e.g., drones or robotic cameras), each with specific kinematic constraints, including maximum translational velocity v_i and rotational velocity ω_i .

We construct a weighted, directed graph $G_i = (V, E, C_i)$ for each agent i . The node set V includes the target poses $\mathcal{P}$ and a depot pose P_0 representing the start and end location. The edge set E represents all possible transitions between poses. The cost function $C_i : E \rightarrow \mathbb{R}_{\geq 0}$ quantifies the time required for agent i to transition from pose P_k to P_ℓ , accounting for:

- 1) **Travel Time:** The Euclidean distance between positions divided by the agent's translational speed.
- 2) **Maneuvering Time:** The time required to reorient the camera (calculated based on the angular difference in pitch and yaw using spherical geometry) divided by the rotational speed.
- 3) **Acquisition Time:** A fixed overhead T_a required to capture and process an image at a given pose.

The objective is to find a set of routes for the fleet such that every target pose in $\mathcal{P}$ is visited exactly once, while minimizing the makespan (the total time until the last agent completes its route). We solve this optimisation problem using a guided local search metaheuristic, which relatively quickly provides near-optimal flight plans that balance the workload across agents and minimize total travel time. This ensures that the data collection phase is executed efficiently, reducing the operational cost of scene refinement.

III. EXPERIMENTS

To validate the proposed framework, we conducted several experiments using simulated environments and specific implementation configurations.

For each combination of scene, viewpoint-selection strategy, and sampling budget, a separate 3DGS model was trained from scratch. Thus, the evaluation comprises more than 30 independently trained reconstructions across the three scenes, selection strategies, and sampling budgets.

This subsection details the scene generation, model parameters, and optimisation settings used to generate the results presented in the subsequent sections. Note that even though we perform our experiments in simulator environments, the method proposed here does not in principal change when moving to a real world setting, apart from the fact that naturally we have to train or fine-tune the quality predictor on images of relevant real world scenery.

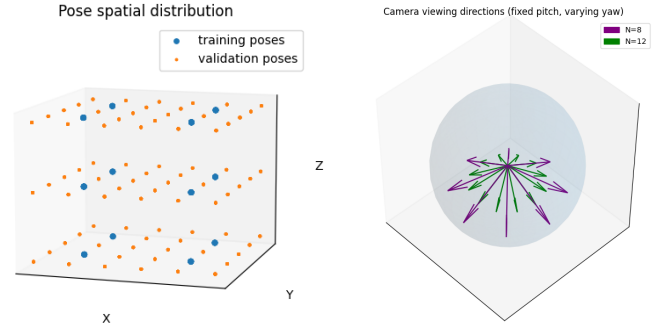


Fig. 2. Within the scene volume of interest, a grid of camera poses is laid out. Some were selected for the initial scene reconstruction (blue). The remaining (orange) poses were used for quality evaluation and selectable pose additions. Each point in the figure shows the spatial position, and hold 8 or 12 camera poses with uniformly distributed (around yaw axis) viewing angles, as shown in the figure on the right. Best viewed in colour.

A. Scene Generation and Data Acquisition

We utilized the CARLA simulator [20] to define three distinct urban scenes, denoted scene 11, 12, and 13, defining specific volumetric boundaries for each. The scene reconstructions were performed using the Splatfacto implementation within the NeRF Studio framework [21], with camera poses estimated using COLMAP [22].

For the initial training phase, virtual cameras were positioned on a 3D grid within the scene volume, see Fig. 2. Two initial sampling densities were tested:

- **Dense Sampling:** A $2 \times 2 \times 3$ spatial grid with 30° yaw increments and a 30° downward tilt, resulting in 12 poses per spatial point (144 total views for Scene 11).
- **Sparse Sampling:** A $2 \times 2 \times 2$ spatial grid with 30° yaw increments and a 30° downward tilt, resulting in 12 poses per spatial point (96 total views for Scenes 12 and 13).

Denser validation grids ($6 \times 5 \times 3$ or $6 \times 5 \times 2$) were established to generate 936 or 624 novel viewpoints for quality evaluation, excluding the initial training poses.

B. Quality Predictor Configuration

The reference-free quality predictor is a ResNet-18-based regression model trained to predict perceptual quality scores at the image tile level. The data used in the construction and

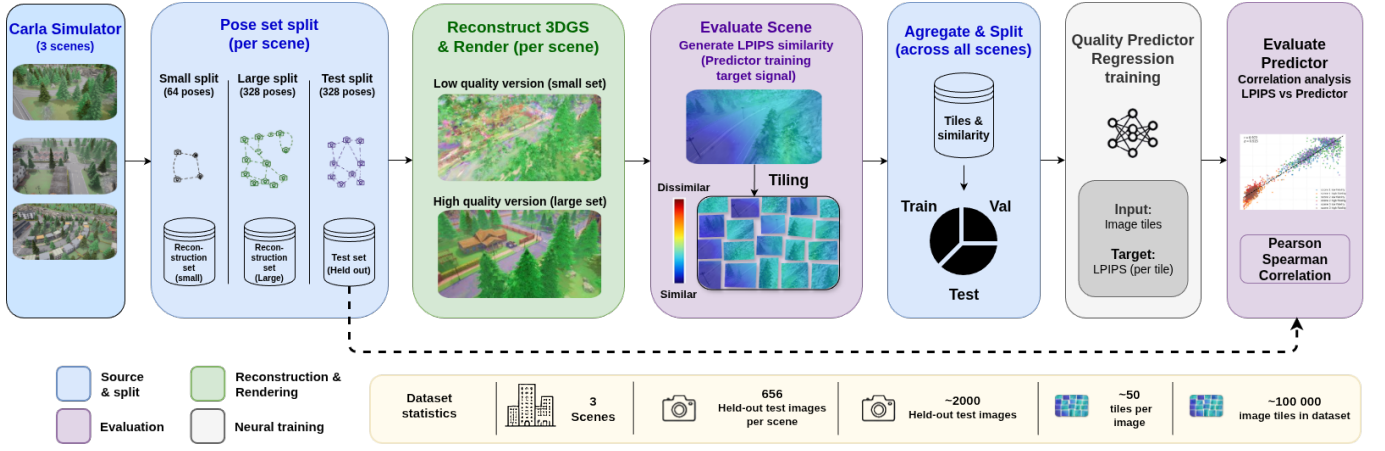


Fig. 3. Quality predictor training data generation process. About 2000 images resulting in around 100k image tiles being generated in total.

evaluation of our quality prediction model, which was applied on image tile level, were generated in accordance with Fig. 3.

To assure scene diversity, we selected three different training scenes and defined their volumetric boundaries. Note that these scenes were independent from those used in the scene reconstruction above. All scenes originated from the same larger map, but defined distinct spots. Within each scene volume we asserted a grid of camera poses (see Fig. 2) which was later used for predictor model training and evaluation.

- **Training Data:** Generated from three independent training scenes (Scene 1-3). This was used to generate two 3DGS reconstructions at different levels of quality (green box in Fig. 3), from which synthesized views were rendered. Each such scene yielded 720 high-definition images, split into 256×256 pixel sized tiles. These were compared with the simulator ground truth tiles, measuring LPIPS similarity which established a ground truth quality value for each tile. Combining tiles from all training scenes resulted in approximately 100,000 tiles sized dataset.
- **Ground Truth:** LPIPS scores were computed between rendered views and simulator ground truth to serve as regression targets.
- **Training Protocol:** The dataset was split into training, validation, and test sets (80%/10%/10%). Training was stopped after 10 epochs to avoid over-fitting to the training split of the tiled image data.
- **Inference:** During evaluation, the predictor operated at three different image scales (downsizing by a factor of 2 at each step). The final quality score for a frame was determined by the scale yielding the highest predicted mean quality.
- **Correlation:** Post-training, the predictor model was validated using correlation analysis, comparing the predicted quality value to the actual calculated LPIPS value. This was done on another set of 3×328 test poses unseen by the training process (the test split in Fig 3).

C. Route Optimisation Setup

The multi-agent route optimisation was implemented using Google *OR-Tools* [23].

To efficiently record supplementary images of the scene, we formulate a VRP by considering a list of vehicles with speeds and turn speeds $U_i = (v_i, \omega_i)$, and a list of camera poses $P_j = (\vec{x}_j, \theta_j, \psi_j)$ with a 3D-position, pitch, and yaw. The weighted, directed vehicle-specific graph is denoted $G_i = (V, E, C_i)$, where $V = \{P_j\}_{j=1}^M \cup \{P_0\}$ is the set of poses, with the pose P_0 denoting the start and end positions for the vehicles, $E = V \times V$ is the set of all tuples of poses, and $C_i : E \mapsto \mathbb{R}_0^+$ is a vehicle dependent cost function defined as

$$C_i((P_k, P_\ell)) = \frac{|\vec{x}_\ell - \vec{x}_k|}{v_i} + \frac{\arccos(\cos(\theta_\ell - \theta_k) \cos(\psi_\ell - \psi_k))}{\omega_i} + T_a \quad (1)$$

The terms represent euclidean distance travel time, turning time based on the angular distance using the spherical Pythagorean theorem and T_a is image acquisition time.

The graphs G_i constitute a VRP problem, where the goal is to find a set of edges, a route, for each vehicle to traverse such that each pose (except P_0) is visited by exactly one vehicle. Note that poses may share the same spatial position but differ in attitude. If T_i is the total schedule time for vehicle i , determined by cumulatively adding the weights C_i over the route of the vehicle, then the objective function is configured to minimize

$$100 \max_i (T_i) + \sum_i T_i, \quad (2)$$

approximating the makespan. Since *OR-Tools* does not allow for optimizing the makespan directly, a weight of 100 is used to make the makespan term dominant. *OR-Tools'* implementation of guided local search for VRP is used as a metaheuristic, and the best solution found after searching for a predetermined time is used as the resulting flight plan. Finding an optimal or near-optimal route for the drones allows us to efficiently dispatch drones to collect the missing data which we can include in the scene generation's next iteration.

Parameters of the optimisation setup:

- **Solver Configuration:** We employed the guided local search metaheuristic for the VRP.
- **Problem Size:** We used $k = 3$ and $k = 5$ for the number of drones used for the sampling. The number of camera poses collected was $N = 30$ or $N = 100$.
- **Search Time:** Convergence studies indicated that a search time of 200 seconds provided a suitable balance between solution quality and computational cost for the problem size considered. See chapter IV-D.

D. Refinement Protocols

For the scene refinement experiments, we compared three selection strategies for adding N new poses:

- 1) **Structured:** Selecting the top- N worst poses identified by the quality predictor.
- 2) **Random:** Selecting N poses uniformly at random from the available validation grid. This was repeated for an ensemble of 7 randomized selections.
- 3) **Mixed:** A combination of $N/2$ worst poses and $N/2$ random poses.
- 4) **Farthest Point Sampling (FPS):** To evaluate our framework against established active vision and Next-Best-View (NBV) planning approaches, we implemented a pure coverage-based FPS method. Without utilizing semantic or reconstruction quality information, starting from an initial seed pose, it iteratively selects the next camera viewpoint that maximizes the minimum Euclidean distance to the already selected set.
- 5) **Structured with 3D Spatial NMS:** Applying NMS to our Structured quality-guided pose set, reducing redundant poses.

Experiments were conducted for $N = 30$ and $N = 100$, with additional ablation study varying N to analyse the impact of added data volume on reconstruction quality.

IV. RESULTS

A. Quality Predictor

The predictor training process is shown in Fig. 3 and the progress in Fig. 4. The final test loss at 10 epochs was somewhat elevated at 0.04 compared to validation losses, however predictor effectiveness is primarily assessed through correlation with the reference LPIPS values, which is the relevant criterion for its intended use in viewpoint selection. These quality predictions were collected across the aggregated test splits, along with the LPIPS measured values. A scatter plot of predicted vs true LPIPS values on image level is shown in Fig. 5. Correlation (Pearson) and rank correlation (Spearman) measures are also calculated. We visualise in Fig. 10 the predicted scene quality, comparing our reference-free predictor’s output to the measured LPIPS quality heatmap. The Pearson correlation between predicted and LPIPS values was measured at 0.972 and the Spearman rank correlation was measured at 0.925. We used 656 held-out rendered test images per scene (328 per quality version, see Fig 3) to measure the

correlation, bringing the total to 1968 images, which allowed us to see a clear correlation between the predictor and LPIPS.

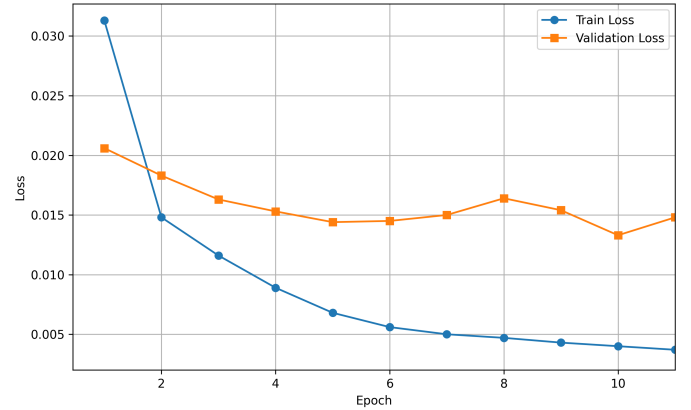


Fig. 4. Quality predictor training process. 80% of tiles were used for the training, 10% for validation during training and 10% for the final test. Training results after 10 epochs were discarded due to overfit tendencies. Apart from these tests we perform a correlation analysis on unseen data as shown in Fig. 5.

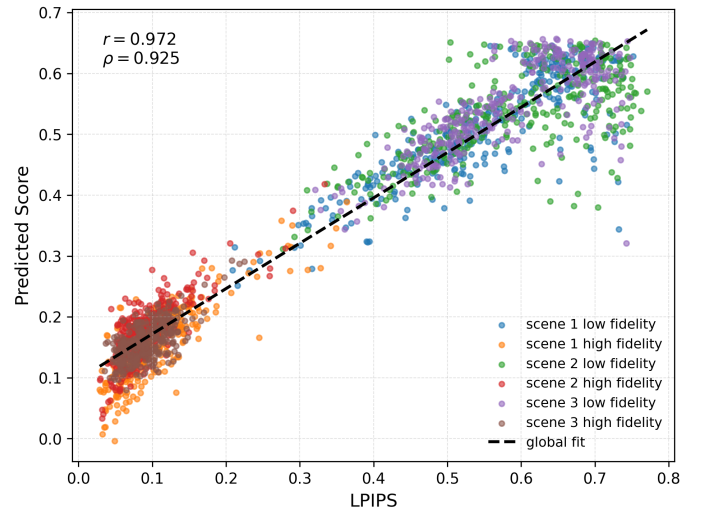


Fig. 5. Quality predictor image-based evaluation. The correlation between the predicted value and the ground truth LPIPS is characterized with a Pearson correlation of 0.972 and a Spearman correlation of 0.925. We note that the predictor acted slightly pessimistic in the high quality region (predicted higher values than LPIPS close to 0) and conversely slightly optimistic (predicted lower values than LPIPS around 0.7).

B. Scene Refinement

We selected N views for values of $N = \{30, 100\}$ using the methods listed in section III-D, sampled new images from the pool of candidates, see Fig. 2, and trained 3DGS scenes on these extended splits. We then evaluated the LPIPS quality of the scenes before and after the added images by rendering a full set of validation pose images and comparing to the ground truth test split simulator images.

As demonstrated in Table I as well as in Fig. 7, our original naive Structured method struggled to outperform a purely Random baseline across the test configurations for $N = 30$. This

TABLE I
SELECTION STRATEGY COMPARISON AT $N = 30$ AND $N = 100$ FOR THE RECONSTRUCTED SCENES. SCENE LPIPS MEDIAN, LOWER IS BETTER.

Method	Scene 11	Scene 12	Scene 13
+30 (Structured)	0.153	0.112	0.103
+30 (Random)	0.150	0.104	0.095
+30 (Mixed)	0.146	0.108	0.098
+30 (FPS)	0.149	0.123	0.110
+30 (Structured NMS)	0.133	0.095	0.087
+100 (Structured)	0.124	0.093	0.073
+100 (Random)	0.125	0.088	0.067
+100 (Mixed)	0.121	0.085	0.072
+100 (FPS)	0.150	0.112	0.113
+100* (Structured NMS)	0.120	0.086	0.066
*actual added with NMS	38	42	78

TABLE II
LPIPS QUALITY IMPROVEMENT WITH INCREASING NUMBER OF ADDED POSES N (STRUCTURED METHOD) FOR THE RECONSTRUCTED SCENES. SCENE MEDIAN, LOWER IS BETTER.

N	Scene 11		Scene 12		Scene 13	
	LPIPS	Δ (%)	LPIPS	Δ (%)	LPIPS	Δ (%)
0	0.203	0.0	0.143	0.0	0.140	0.0
15	0.167	-17.7	-	-	-	-
30	0.153	-24.6	0.112	-21.7	0.103	-26.4
100	0.124	-38.9	0.093	-35.0	0.073	-47.9

phenomenon is deeply rooted in spatial clustering: packing supplementary viewpoints together adds marginal information to the underlying 3D Gaussian Splat. For $N = 100$, the NMS exhausted the candidate pose pool. The number of poses at which this happened is shown in Table I for relevant cases.

For the random selection method, an ensemble of 7 random scene reconstructions were made and the average across those was used to represent the random pose selection case. We show in Fig. 11 an example of the visual result of refinement of scene 13, including the LPIPS heatmap indicating the measured quality.

C. Ablation study

We studied how the quality gain varied with the number of poses N selected from the list of worst quality poses defined by the predictor. We selected the top- N worst views for a range of N , sampled new images and trained 3DGS scenes on these extended splits (one scene was generated for each value of N) and then evaluated the LPIPS quality of the scenes by rendering a full set of validation pose images and comparing to the ground truth test split simulator images. It can be seen in Fig. 6 and Table II that the number of added image poses increase the LPIPS quality.

D. Route Optimisation

An empirical experiment was conducted to determine a good search time for the metaheuristic. Randomized problem instances were solved with increasing search time for different problem sizes relevant to the overall study, and the best makespan found was plotted as a function of the solve time. Fig. 8 shows that the makespan converged to a fairly stable

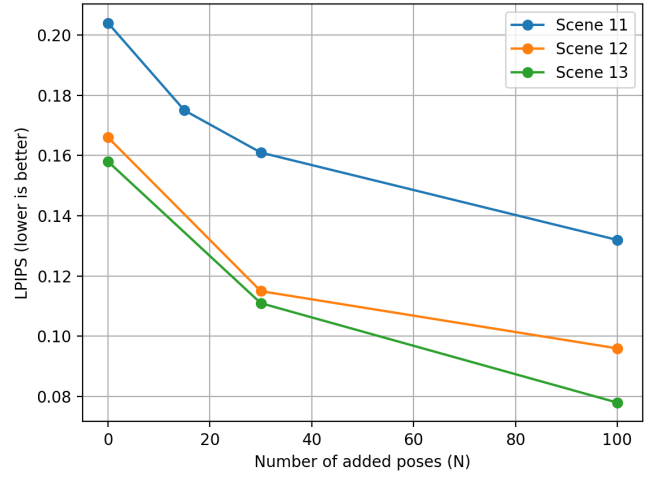


Fig. 6. The effect of adding increasing amounts of top- N predicted worst quality poses to the base set. We see diminishing returns but no plateau even at 100 added images.

value within 100 s for all problem instances, whereafter further search only gave small improvements in solution quality. Experiments were run on an Lenovo Legion laptop with an 8-core AMD Ryzen 7 5800H processor utilizing a single core.

Based on the results of this experiment, a solve time of 200 s was deemed sufficient. In Fig. 9 we exemplify the routing solution for capturing 100 additional images for using three or five drones.

V. DISCUSSION

While the reference-free quality predictor correlates strongly with ground-truth LPIPS scores ($\rho = 0.972$), the downstream impact of selecting viewpoints based solely on predicted perceptual quality reveals a critical limitation in purely greedy strategies.

A key factor underlying these results can be interpreted through the lens of exploration versus exploitation. Contrary to the expectation that targeting the "worst" views would yield the highest returns, our results show that the Structured selection strategy often comes out on par with the Random strategy. We attribute this to spatial correlation of perceptual errors. In 3D scene reconstruction, low-fidelity regions are rarely isolated to a single viewpoint; rather, they manifest across clusters of adjacent poses.

However, we showed that applying a 3D Spatial NMS to the quality-guided pose selection yields a profound performance leap. By enforcing a spatial diversity constraint ($D_{min} = 5.0$ m in this case), Structured NMS achieves the lowest LPIPS error scores across all evaluated scenes for $N = 30$, outperforming both the Random baseline and the geometric FPS baseline.

This confirms a crucial scientific insight: active data collection frameworks cannot rely on quality metrics alone, nor can they rely purely on geometric spacing such as FPS. Rather, the optimal loop requires quality-guided localization filtered by geometric diversity. Although orientation is not explicitly considered in the spatial NMS, its purpose is to prevent

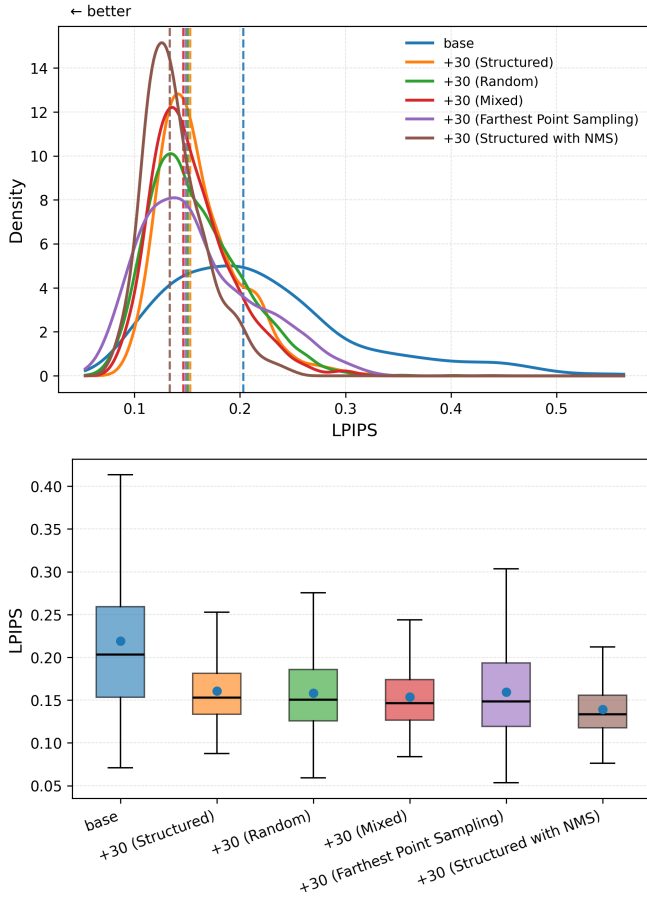


Fig. 7. Comparison of refinement strategies for Scene 11 using $N = 30$ added poses: Top: Quality distribution histograms. Bottom: Box plots of LPIPS scores. While medians are comparable across methods, the Structured and Mixed strategies yield a higher density of high-quality (low LPIPS) poses. It is clear that the addition of NMS to the quality-based Structured selection method resolves the pose clustering giving it an edge over both FPS and Random methods.

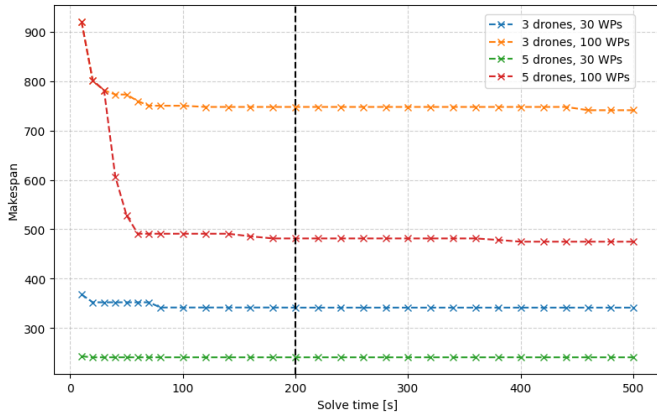


Fig. 8. Convergence test results. Solve time of 200 s was selected for our experimental setup.

excessive concentration of selected viewpoints in regions with low reconstruction quality. Nearby candidates therefore compete for selection, promoting a more spatially distributed acquisition. The results show that this simplified position-only criterion is nevertheless effective in practice.

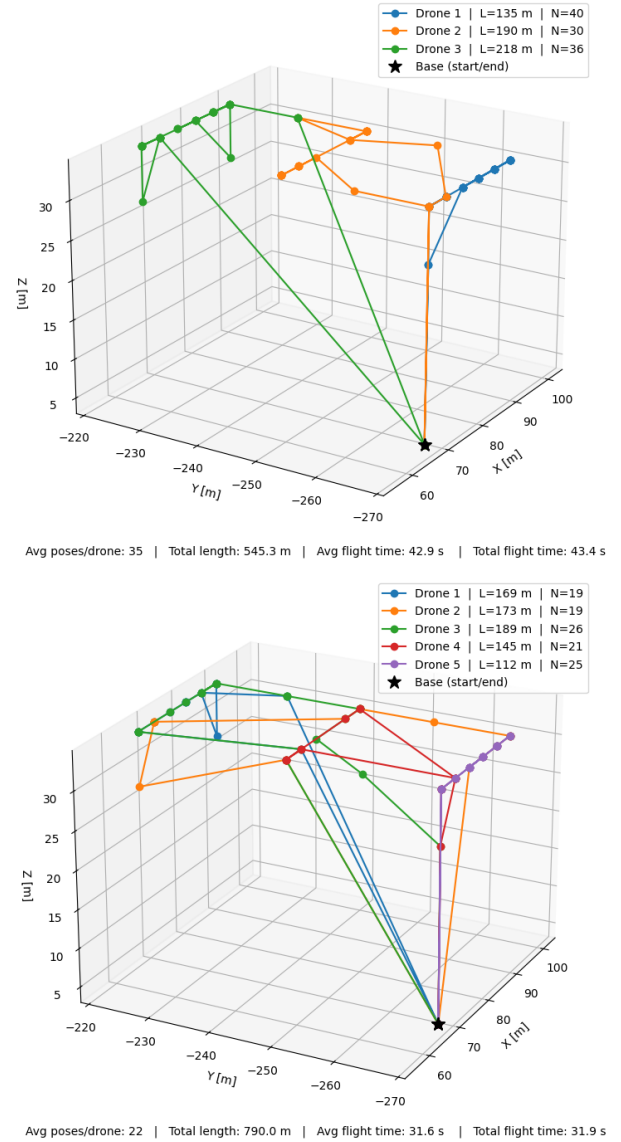


Fig. 9. Sampling drone routes for collecting added images of scene 11. $N = 100$ (number of poses to visit). Each coloured graph shows the optimised route for one drone and dots indicate camera poses to be visited for sampling the scene. The black star denotes the take-off and landing location (base). Note that several poses co-locate spatially, differing only in attitude. Top: Routing of three data sampling drones. Bottom: Routing of five data sampling drones. Both solutions visit the exact same poses, using different resources.

At a larger budget of $N = 100$, the analytical edge of Structured NMS becomes less pronounced. Because the strict spatial constraint ($D_{min} = 5.0$ m) depletes the optimal candidate subsets, the system is forced to substitute fallback greedy poses to meet the quota. We hypothesize that expanding the initial candidate sampling space would preserve this performance margin at higher sampling scales.

VI. CONCLUSIONS

We introduced an automated closed-loop framework for quality-driven refinement of 3DGS reconstructions. By combining a reference-free perceptual quality predictor with a multi-agent VRP route optimiser, our system automatically

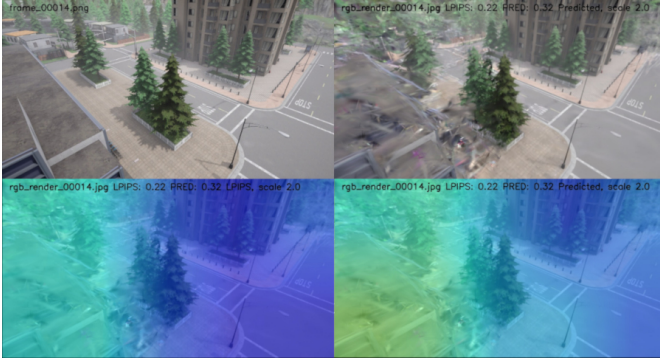


Fig. 10. Scene 13 reconstruction and quality prediction. Top left: Ground truth scene example. Top right: Rendered scene view from reconstructed base scene (reconstruction based on 96 images). Bottom left: LPIPS perceptual distance heatmap visualisation for reconstructed scene (distance between the two top images, blue is more similar). Bottom right: Estimated quality heatmap using the reference-free trained predictor (only seeing the top right rendered scene image). Best viewed in colour.

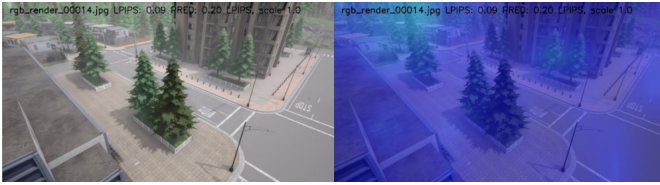


Fig. 11. Reconstruction enhancement result for scene 13. Left: Rendered scene view from reconstructed refined scene (Structured, i.e. based on adding 30 worst quality images to base scene set). Right: LPIPS perceptual distance heatmap visualisation for said reconstructed scene (blue is more similar to ground truth). Best viewed in colour.

identifies and samples low-quality viewpoints to extend existing dataset.

Our main finding is that greedy viewpoint selection based solely on perceptual error suffers from severe spatial redundancy, wasting budget on clustered observations. Conversely, maximizing coverage without quality guidance is also sub-optimal. The best performance is achieved by pairing our perceptual predictor with an NMS-based selection strategy that balances quality cues with spatial diversity.

Future work will explore hybrid quality metrics, optimisation-based pose selection for real-world UAV operations, as well as better simulated dynamics, enabling a more automated scene reconstruction improvement system.

REFERENCES

- [1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
- [3] W. Xiao, R. S. Cruz, D. Ahméd-Aristizabal, O. Salvado, C. Fookes, and L. Lebrat, “Nerf director: Revisiting view selection in neural volume rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 742–20 751.
- [4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [5] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, “Nerf in the wild: Neural radiance fields for unconstrained photo collections,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7210–7219.
- [6] H. Liang, T. Wu, P. Hanji, F. Banterle, H. Gao, R. Mantiuk, and C. Öztireli, “Perceptual quality assessment of nerf and neural view synthesis methods for front-facing views,” in *Computer Graphics Forum*, vol. 43, no. 2. Wiley Online Library, 2024, p. e15036.
- [7] O. Mendez, S. Hadfield, N. Pugeault, and R. Bowden, “Taking the scenic route to 3d: Optimising reconstruction from moving cameras,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4677–4685.
- [8] W. Jiang, B. Lei, and K. Daniilidis, “Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information,” *arXiv preprint arXiv:2311.17874*, 2023.
- [9] K. Lee, S. Gupta, S. Kim, B. Makwana, C. Chen, and C. Feng, “Sonnerf: Active view planning for nerf using surrogate objectives,” *arXiv preprint arXiv:2312.03266*, 2023.
- [10] X. Chen, Q. Li, T. Wang, T. Xue, and J. Pang, “Gennbv: Generalizable next-best-view policy for active 3d reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 436–16 445.
- [11] X. Pan, Z. Lai, S. Song, and G. Huang, “Activenerf: Learning where to see with uncertainty estimation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 230–246.
- [12] M. Kim, S. Seo, and B. Han, “Infonerf: Ray entropy minimization for few-shot neural volume rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 912–12 921.
- [13] S. Moon, J. Lee, J. Lee, D. Nam, and W. Yoo, “A study on nerf-based synthetic image generation and post-processing method for object detection,” in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2022, pp. 1560–1562.
- [14] M. Dronova, V. Cheremnykh, A. Kotcov, A. Fedoseev, and D. Tsetserukou, “Flynerf: Nerf-based aerial mapping for high-quality 3d scene reconstruction,” *arXiv preprint arXiv:2404.12970*, 2024.
- [15] J. Zheng, J. Tan, and G. Chen, “Research on uav coverage path planning in building visual inspection,” *Journal of Building Engineering*, vol. 111, p. 113556, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352710225017930>
- [16] W. Meng, X. Zhang, L. Zhou, H. Guo, and X. Hu, “Advances in uav path planning: A comprehensive review of methods, challenges, and future directions,” *Drones*, vol. 9, no. 5, 2025. [Online]. Available: <https://www.mdpi.com/2504-446X/9/5/376>
- [17] S. Papaioannou, P. Kolios, T. Theocharides, C. G. Panayiotou, and M. M. Polycarpou, “Integrated guidance and gimbal control for coverage planning with visibility constraints,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 2, pp. 1276–1291, 2022.
- [18] Z. Hu, M. Z. Butt, N. Nasir, and M. B. Salleh, “Vision-based autonomous navigation for quadrotor uavs in unknown and constrained environments,” *International Journal of Intelligent Robotics and Applications*, vol. 9, no. 4, pp. 1743–1754, 2025.
- [19] B. Kang, C. Wang, Y. Su, and J. Zeng, “Multi-uav forest area inspection path planning based on concave polygon region decomposition,” *Scientific Reports*, vol. 15, no. 1, p. 41990, 2025.
- [20] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “Carla: An open urban driving simulator,” in *Conference on robot learning*. PMLR, 2017, pp. 1–16.
- [21] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, “Nerfstudio: A modular framework for neural radiance field development,” in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH ’23, 2023.
- [22] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [23] V. Furnon and L. Perron, “Or-tools routing library,” <https://developers.google.com/optimization/routing/>, Google. [Online]. Available: <https://developers.google.com/optimization/routing/>