Autonomous UAV 3D Reconstruction using Prediction-Based Next Best View

Ziwen Wang, Bashar Alsadik, Francesco Nex

Department of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Hallenweg 6, 7522 NH Enschede ziwen.wang@utwente.nl, b.s.a.alsadik@utwente.nl, f.nex@utwente.nl

Keywords: NBV, Autonomous, MACARONS, 3D Reconstruction, Self-supervision, Blender.

Abstract

High-quality 3D reconstruction of infrastructure using UAVs is essential for inspection, monitoring, and digital twin applications. Traditional flight planning methods rely on predefined paths and often struggle with complex geometries, leading to incomplete models and inefficiencies. This paper evaluates a state-of-the-art autonomous Next Best View (NBV) of MACARONS model (Mapping And Coverage Anticipation with RGB Online Self-Supervision), which enables online, self-supervised 3D reconstruction of large-scale scenes using only a monocular RGB sensor. The MACARONS NBV model autonomously adjusts UAV trajectories in real time based on predictions of unseen scene structure to improve reconstruction accuracy and surface detail recovery. Despite its advantages, a key limitation is its lack of consideration for camera coverage percentage from a photogrammetric perspective, which makes it challenging to consistently obtain an informative point cloud. The simulation results demonstrate that the autonomous NBV strategy significantly enhances both reconstruction quality and operational efficiency. To evaluate its effectiveness, we applied the MACARONS NBV model to two open-access 3D bridge models. The generated camera trajectories were imported into Blender, where we rendered high-resolution images using realistic camera intrinsics to overcome the limitations of the low-resolution depth predictions. From these images, we reconstructed point clouds and compared them to those produced by a traditional flight planning approach, as well as to the ground truth models. The comparison highlights the added value of autonomous view planning for accurate and efficient UAV-based 3D reconstruction. The two experiments showed a high coverage percentage of 88 % compared to the ground truth and 90% compared to traditional flight planning based on a 37.5% efficiency raise. This work highlights the potential and current limitations of prediction-based NBV in UAV photogrammetry and motivates further research into integrating coverage-aware planning.

1. Introduction

The application of Structure-from-Motion algorithms to images captured from Unmanned Aerial Vehicles (UAVs) has made it feasible to reconstruct 3D models of expansive outdoor settings, for instance, in order to create a Digital Twin of the scene. Autonomous path planning methods like the modified A* algorithm (Duchoň et al., 2014), and RRT algorithm (Kuwata et al., 2009) are aimed at navigating robots in a three-dimensional domain with obstacles, which are formulated as an optimization problem for the shortest path. However, those flight path algorithms are not designed for the 3D reconstruction task and can lead to incomplete models, especially in complex environments. To provide high-quality 3D reconstruction data of infrastructure, precise UAV image data acquisition is important. Nowadays, automatic navigation in 3D reconstruction can be effectively achieved through various methods such as Object-Aware Guidance and Scene Reconstruction (Liu et al., 2018), and sensor fusion and accuracy (Li et al., 2020). These methods focus on maximizing information gain and minimizing path inefficiencies, such as excessive turns, to improve reconstruction quality and reduce operational costs. Despite advancements, challenges remain in achieving fully automatic 3D reconstruction and navigation. Issues such as scale accuracy, especially in GPS-denied environments, require systematic analysis and error evaluation. Therefore, informative UAV motion planning should be considered to capture the images and cover the whole scene from different perspectives. The Next Best View (NBV) approach is increasingly used as an outstanding methodology for viewpoint selection optimization in autonomous UAV, demonstrating substantial potential to enhance the operational efficacy of informative path planning.

In this paper, we will implement a self-supervised online NBV method to reconstruct the scene using an RGB sensor. Among the growing number of NBV strategies developed in recent years, we customize the MACARONS method (Guédon et al., 2023) due to its unique strengths in autonomous online learning and scalability. Compared to traditional NBV methods, it relies on supervised training with limited datasets or handcrafted heuristics. MACARONS prediction-based policy allows to dynamically adapt trajectories during flight without predefined maps or human intervention which aligns with our goal of UAV autonomy. MACARONS NBV method demonstrates strong performance in self-supervised online 3D reconstruction. However, it often struggles in complex or large-scale indoor environments. One key limitation is that it focuses only on the next immediate view and then the UAV tends to get trapped in locally reconstructed areas and missing under-explored regions of the scene. This behaviour reduces overall coverage and reconstruction completeness (Li et al., 2025). To our knowledge, no prior study has evaluated prediction-based NBV methods like MACARONS from a photogrammetric perspective, particularly in terms of image coverage and reconstruction quality. Accordingly, this paper seeks to answer the following research question: Can the MACARONS NBV method achieve high-quality 3D reconstruction with sufficient scene coverage during online exploration of complex infrastructure objects? To answer this, we design a pipeline that evaluates the coverage quality of the reconstructed 3D models by comparing them to ground truth point clouds, offering insight into the performance and limitations of this prediction-based NBV of MACARONS in autonomous UAV mapping tasks.

Building upon the NBV strategy in Macarons, optimized global exploration trajectories are generated by iteratively selecting camera poses that maximize surface coverage gain. The derived poses are validated in Blender for geometric consistency, where synthetically projected multi-view RGB images are analysed for visibility and occlusion check. These validated images and poses are then processed using the Metashape photogrammetric pipeline, applying automated feature matching, dense point cloud

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-W2-2025-207-2025

reconstruction, and global bundle adjustment to generate a high-fidelity 3D model. At the end, the reconstructed 3D model is compared to the ground truth point cloud (section 3.4). Furthermore, a reconstructed 3D model using a traditional flight plan is also demonstrated.

In the following section 2, we will first introduce the methodology, including the MACARONS module description (section 2.1) and our evaluation pipeline (section 2.2). Then, the experiments will be described (section 3) showing the results of two bridge models. Next, we will discuss the result (section 4) and followed by a conclusion and future work (section 5).

2. Methodology

This section is to introduce the methodology of MACARONS and the evaluation pipeline for the autonomous reconstruction performance of the MACARONS NBV approach. First, the core principles and functionality of the MACARONS model are described. Then, our custom evaluation pipeline is designed.

2.1 Prediction-Based NBV with MACARONS

As mentioned, MACARONS represents a significant advance in 3D scene reconstruction and exploration. With RGB online self-supervision, it offers an efficient and scalable solution to the NBV problem, and as a result, it is a promising tool for large-scale environmental mapping applications where depth sensors are not applicable.

This method works without perfectly known 3D objects and explicit 3D supervision and can explore the scene and reconstruct it only by using RGB images. The MACARONS takes coverage maximization as its core optimization criterion to solve the NBV problem. The method is assumed to provide a 3D bounding box to accurately reconstruct the area.

As we know, the NBV problem is the identification process of the next most informative sensor position for reconstructing a 3D object or scene efficiently and accurately (Hepp et al., 2018; Mendoza et al., 2020; Zeng et al., 2020). Here, we look for the view that increases the most of the total coverage of the scene surface: the number of new visible surface points.

To describe the problem in a mathematical way, first the scene is represented as occupied points $x \in R^3$ and boundary ∂_x. Then, the observation of the scene are presented as images captured by cameras, therefore images $(I_0 \dots I_t)$ and cameras $(c_0 \dots c_t)$ are used for NBV strategy. The total coverage of the scene surface is described as surface coverage gain $G_t(c)$. See Figure 1.

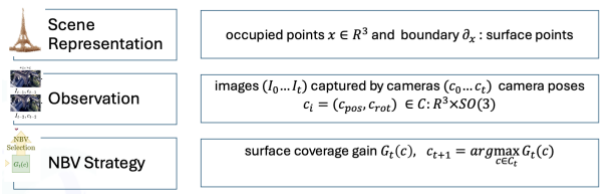

Figure 1. Mathematical description in Macarons

For the scene representation, a bounding box is asked to as an input to the algorithm, which allows the user to define the target reconstruction area within the scene and newly reconstructed surface points will be registered inside the coarse 3D grid to compute accurate surface coverage gains optimizing viewpoint parameters and coverage boundaries for the 3D reconstruction. Based on the given scene bounding box, an adaptive camera bounding box is calculated by the scene bounding box to ensure the candidate camera poses exploration. Each bounding box is discretized with grids. The amount of the grid in height, width and length can be tuned based on the scene. See Figure 2.

For the cameras $(c_0 \dots c_t)$, camera poses can be written as $c_i = (c_{pos}, c_{rot}) \in C: R^3 \times SO(3)$ All the camera poses are discretized on a 5D grid. The 3D camera position is represented as $c_{pos} = (x_c, y_c, z_c)$ and 2D camera rotation c_{ros} is encoded as azimuth and elevation. See Figure 3.

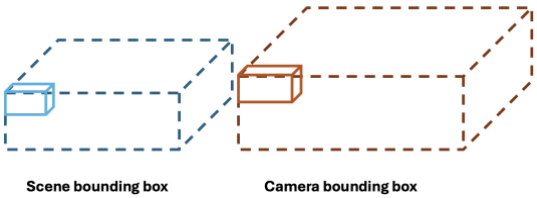

Figure 2. Bounding boxes and girds

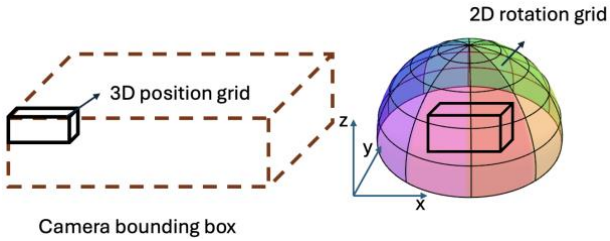

Figure 3. Camera 5D grid

Online exploration is operated by a self-supervised online learning procedure, where parameter optimization in each module is achieved by an iterative training executing three steps at each time step t: (i) *Decision Making*, (ii) *Data Collection & Memory Building*, and (iii) *Memory Replay*. The steps are described in the following Figure 4 while the three modules: depth module, volume occupancy module and surface coverage gain module, which are executed during the first step are shown in Figure 5 and described in the following section. In our evaluation process, the pre-trained module is used, which was already trained by large infrastructure provided.

These three steps will work as follow:

(i) In *the Decision-making* step, the next best camera pose c_{t+1} is selected by calculating the highest coverage gain of all sampled camera poses.

(ii) In the *Data Collection & Memory Building* step, the camera is moving from c_t to c_{t+1}, which is done by n linear interpolation steps. During this camera moving procedure, n images will be captured. The images with the depth data and camera poses will be taken as three different supervision signals to three different modules. These signals will be used to instruct the training process of the three modules by iterations, to calculate the Loss and to update the modules parameters. They will be stored in the Memory.

(iii) In the *Memory Replay* step, newly acquired data is added to the sample data, which is already stored in Memory in the second step to update the loss of each module by comparing the current scene state to the state of the same scene at the previous camera pose.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-W2-2025-207-2025 | © Author(s) 2025. CC BY 4.0 License.

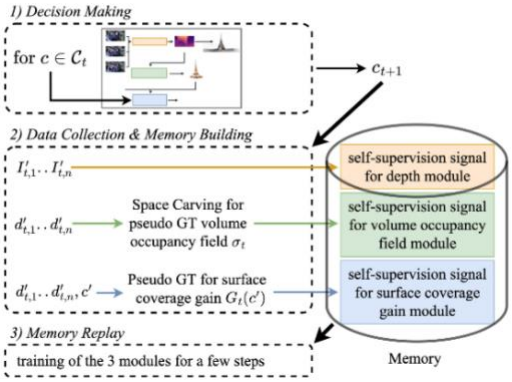

Figure 4. Three iteration steps (Guédon et al., 2023)

2.1.1 Depth module: This module (Watson et al., 2021) is aimed at online surface point reconstruction by depth prediction. During the exploration camera will capture a sequence of images I_t, I_{t-1}, I_{t-2},… I_{t-m}. These images and corresponding camera poses c_t, c_{t-1}, c_{t-2},… c_{t-m} will be used as inputs for the module to predict the depth map d_t based on the last observation I_t ($0 \ll m \ll t$). In the end, S_t as reconstructed surface point clouds (Godard et al., 2017) (Heise et al., 2013) which will be kept updated during observation using predicted depth maps.

2.1.2 Volume occupancy module: This module can derive a "volume occupancy field" σ_t from the predicted depth maps. It provides a basis for sample points for visibility integration of all the sampled points in the next step. $\sigma_t(p) = 0$ demonstrates that point p is empty in space; $\sigma_t(p) = 1$ demonstrates that point p is occupied in space. The inputs of this module (Vaswani et al., 2023) are the point p surface point cloud S_t and previous camera poses c_i, with the pseudo ground truth occupancy this module will predict a partial volumetric representation regarding a scalar value [0,1].

2.1.3 Surface coverage gain module: The number of new visible surface points is defined as the surface coverage gain. Given any camera pose c based on the predicted occupancy field V_c, the results of visibility integration $G_t(c)$ of the sample points can be derived. The criterion to select the next camera pose c_{t+1} is to choose the maximum surface coverage gain (Guédon et al., 2022).

MACARONS is a method that enables efficient exploration and reconstruction of large-scale scenes from a single monocular RGB input. However, the method assumes a static scene environment, though this limitation can be mitigated by existing self-supervised depth prediction models that demonstrate robustness to dynamic objects. Furthermore, a rather simple path planning is conducted by sampling the camera pose in the neighbourhoods of the current camera to estimate c_{t+1}. Therefore, developing trajectory planning algorithms could improve exploration efficiency.

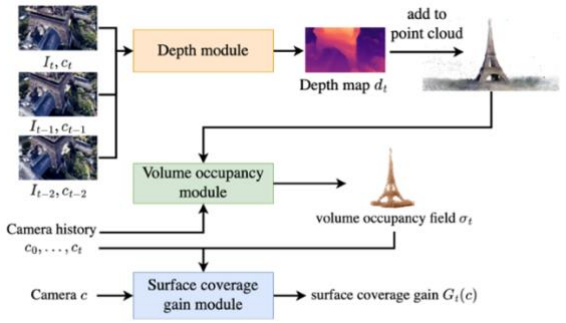

Figure 5. Three modules to generate the next camera pose (Guédon et al., 2023)

2.2 Evaluation pipeline

Our evaluation pipeline Figure 6 provides an assessment method of the 3D reconstruction quality using camera trajectories generated by MACARONS.

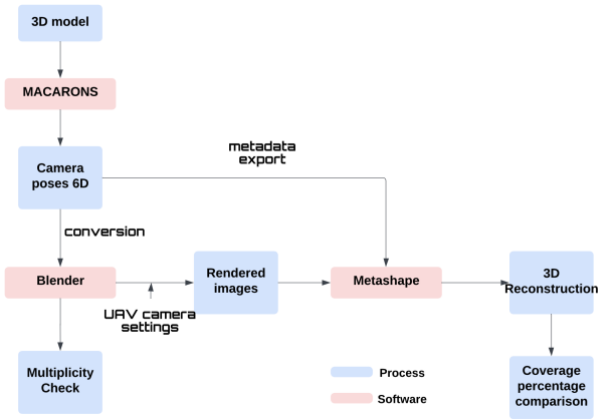

Figure 6. Evaluation pipeline

Here, we follow the pipeline to discover and improve the 3D reconstruction result using MACARONS. It will provide a basis for the validation procedure in this paper. The input in the pipeline are existing 3D models, which are obtained from an open source website (https://sketchfab.com/feed), Sketchfab.

2.2.1 Camera poses Conversion: The goal of the conversion is to get the same camera pose in Blender as in MACARONS to render the correct images. As one of the most important outputs from MACARONS, camera poses in $c_i = (c_i{}^{pos}, c_i{}^{rot})$ provide both the location and rotation of the cameras in sequence which is in world (local) coordinates. Additionally, the rotation is given as azimuth and elevation, where the conversion is needed later. Then, we can import these camera sequences into the blender tool using an embedded Python interpreter in the scripting function.

From the given azimuth and elevation, we convert them into $zyx - euler$ as the representation used in Blender. Moreover, the camera settings also need to be added to Blender.

The method mentioned in (Alsadik et al., 2023) is introduced to calculate the transformation despite the initial orientation difference between camera coordinates and world coordinates as shown in Figure 7.

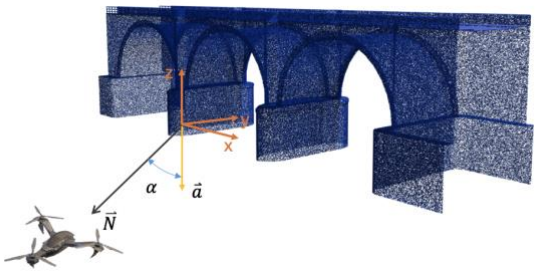

Figure 7. camera normal vector and world coordinates

Where:
$\vec{N}$: normal vector of the camera
$\vec{a}$: camera original view direction without rotation in Blender
α : angle between the normal vector and the original view direction

The viewing direction $\vec{N}$ of the camera should be defined by the given azimuth and elevation angles.
In Blender, the default downward camera orientation $\vec{a}$ (0, 0, -1) is -Z axis. A Rodrigues formula is used to calculate the rotation matrix based on the given conditions shown in Figure 6.

$$R = I + \sin\alpha K + (1 - \cos\alpha)K^2 \quad (1)$$

$$K = \begin{bmatrix} 0 & -k_z & k_y \\ k_z & 0 & -k_x \\ -k_y & k_x & 0 \end{bmatrix} \quad (2)$$

Where:
R is 3×3 *rotation matrix*
I is 3×3 *identity matrix*
K is the skew-symmetric matrix regarding to unit axis vector, here is represented as $\vec{a}$

2.2.2 3D reconstruction: The 3D reconstruction alignment of the rendered frames is applied in Agisoft Metashape with the initial camera poses exported from MACARONS.
The 3D reconstruction can be generated following the procedures outlined in Metashape. This reconstructed 3D model can be compared against the ground-truth dataset (SketchFab as mentioned) to quantitatively assess the coverage accuracy through percentage-based metrics, thereby evaluating the geometric accuracy of the reconstruction.
2.2.3 Multiplicity check: After transformation, rendering, image alignment, and 3D reconstruction, the coverage percentage of the images generated by MACARONS is also computed (Mousavi et al., 2021). An illustration of the principle of coverage computation and visualization is shown in Figure 8.

Figure 8. Multiplicity check

The visibility analysis for 3D point cloud validation is implemented through a multi-stage computational pipeline. In a visibility check, a camera-centered reference system is used, where 3D points are transformed into the image reference system using the camera's projection matrix. After transformation, the resulting homogeneous coordinates are analyzed. If the normalized x, y, and z of the transformed point fall in the range [0, 1], the point is considered to be in the camera's view frustum (Ilie, 2003), which means the point is visible to the camera. Subsequently, an occlusion check via ray-casting is performed on these visible points. A ray is cast from each camera's optical center to the target point, using collision detection constrained within ray casting distance (from camera to the target point).
Furthermore, points visible to fewer than two cameras were annotated with a red marker to indicate insufficient multi-view coverage. For points observable by two or more cameras, a baseline-to-depth ratio analysis was conducted on all observing camera combinations pairwisely based on the photogrammetric principles to ensure the 3D reconstruction quality. This metric quantified the space distribution between camera baselines (B) and point-to-camera distances (d_1, d_2) through the ratios $\frac{B}{d_1}$ and $\frac{B}{d_2}$. Points satisfying the empirical range of $0.1 \leq \frac{B}{d} \leq 0.4$
In the multiplicity check, it is verified if the point can be seen for more than 2 cameras. The color of the point is classified based on the result of visible camera count with the color transitioning from cool to warm to represent increasing visible camera counts as shown in Figure 8.

3. Experiment

In this section, our results will be introduced. The experiments aimed to assess the 3D reconstruction of bridge structure, which compares conventional flight planning (López et al., 2013)(Santamaria et al., 2012) with the selected autonomous NBV planning of MACARONS.

3.1 Experiment setup

Both experiments used 3D bridge models are sourced from the open-access platform Sketchfab , which are selected for different distinct structural types and varying geometric complexity.
Bridge A: A brick-arched stone bridge, characterized by thick masonry walls, arched spans, and relatively enclosed geometry. The surfaces exhibit high textural richness and some underdeck occlusion zones.

Figure 9. stonebridge (height: 14m length: 55.5m width: 11.9m)

Bridge B: A steel truss bridge with open lattice-like geometry formed by repeated triangular steel elements and elevated side railings. This model presents more internal visibility but includes complex self-occlusion due to its framework.

Figure 10. Truss bridge (height: 30.4m length: 100m width: 22.1m)

Experiment 1 is conducted on Bridge A, comparing both conventional flight planning and NBV planning.
For NBV planning, the MACARONS framework is used to select the next best view online, based on prediction-based coverage estimation. The plot visualizes the surface coverage curves for two bridges, showing how the coverage improves over multiple rounds. The red line represents the mean surface coverage across multiple runs, indicating the average performance of MACARONS. The shaded red band around the line represents the variability (standard deviation) in surface coverage across runs, providing insight into the consistency of the method. A narrower band means more consistent behaviour, while a wider band suggests higher variability. The x-axis represents the number of rounds, which corresponds to iterative steps (e.g., adding new cameras or actions), and the y-axis represents the surface coverage, which measures the percentage of the covered scene surface. This visualization shows the effectiveness and reliability of the MACARONS method in achieving high surface coverage over time. The system continues to collect views until no significant additional coverage gain is detected, which is shown in Figure 11.

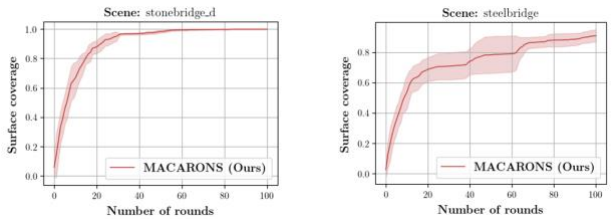

Figure 11. BridgeA (left) & Bridge B (right) surface coverage curve

Experiment 2 is conducted on Bridge B using both conventional flight planning and NBV-driven planning.
Important parameters are evaluated, such as reprojection errors, tie point multiplicity, and computational efficiency.
All the experiments use the same camera settings: the camera lens is an 18 mm fixed-focal-length, lens distortion-free, and the image resolution is 1920×1080. The GSD (ground sampling distance) (Felipe-García et al., 2012) in both planning methods is around 2cm and 1.5cm for Bridge A and Bridge B, respectively.

3.2 Experiment 1: Bridge A

3.2.1 Conventional flight planning: This planning is designed as a grid-based flight path. The GSD is calculated as 2.0cm. 485 images were captured during the flight. Coverage is 97.5%. The average tie point multiplicity is 5.7, and the RMS (root mean square) reprojection error (Yuan et al., n.d.) is 0.69 pixels. The UAV waypoints are shown in Figure 12 from different perspectives.

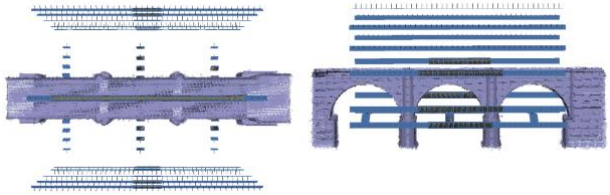

Figure 12. stonebridge in conventional flight planning

3.2.2 NBV driven flight planning: The flight path is dynamically online adjusted in MACARONS using NBV planning. The GSD is calculated around 1.8cm. This approach completes the reconstruction with only 303 images, achieving 91.5% coverage. The tie point multiplicity is 4.2, RMS error is 0.84 pixels. The waypoints of UAV can be seen in Figure 14.

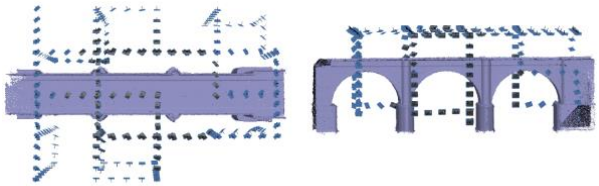

Figure 13. stonebridge in NBV flight planning

3.3 Experiment 2 – Bridge B

3.3.1 Conventional planning: Camera path settings are the same as the Bridge A. The GSD is calculated as 1.5cm. 1106 images are captured, covering 99% of the model, with an average tie point multiplicity of 5.2. The RMS reprojection error is 0.89 pixels. The UAV waypoints are shown in Figure 14 from different perspectives.

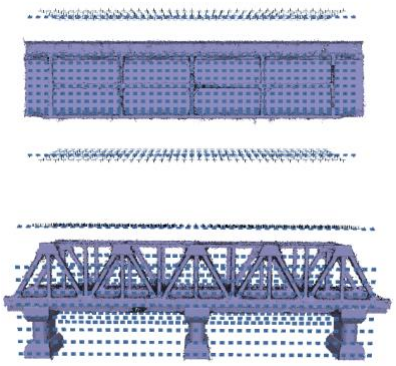

Figure 14. the truss bridge in conventional flight planning

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-W2-2025-207-2025 | © Author(s) 2025. CC BY 4.0 License.

3.3.2 NBV driven planning: The autonomous MACARONS approach completed the model using only 392 images, resulting in 88% total coverage. The GSD is calculated as 1.5cm. The tie point multiplicity is 4.7, and the RMS is 0.85 pixels. The waypoints of UAV are shown in Figure 15.

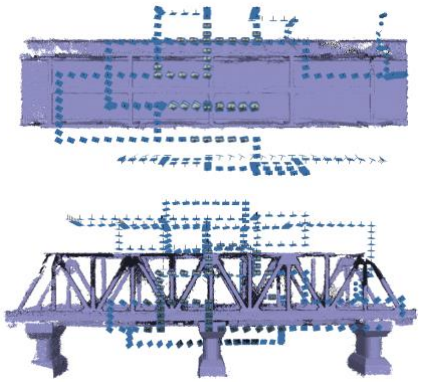

Figure 15. truss bridge in NBV flight planning

3.4 3D reconstruction quality evaluation parameters

Three different parameters to evaluate the comparison between conventional and NBV planning using two bridges model are shown below in Table 1 and Table 2. Reconstruction is quantified using reprojection errors and average tie point multiplicity. A quantitative comparison performance reveals distinct differences between conventional and autonomous bridge reconstruction methods.

BridgeA: Referring to reprojection errors, the autonomous method has a 10.3% increase in RMS error compared to the traditional method. It indicates that NBV planning has lower matching consistency in feature-sparse regions. When it comes to tie point Multiplicity, a 26.3% decrease in the autonomous method shows a sparser cross-view feature tracking in NBV planning, see Table 1.

These error is mainly caused by the image decrease, which is strongly influenced by UAV efficiency. Despite all, the rise of the error in NBV planning captured images are decreased by 37.5% with comparable reconstruction quality, see Figure 16.

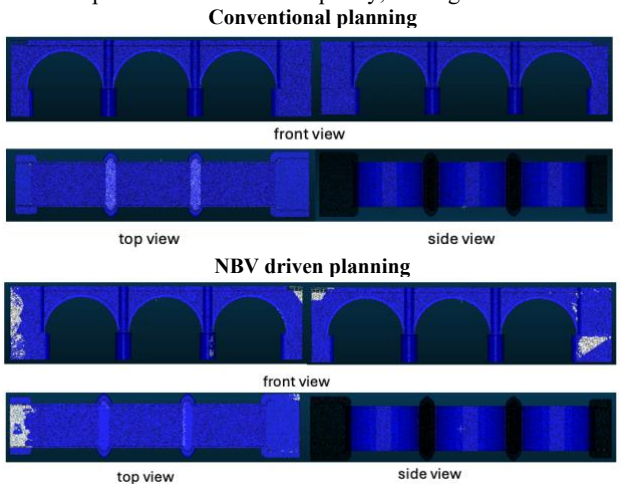

Figure 16. Coverage comparison with ground truth (white spots uncovered)

Bridge A	Bridge coverage	Images amount	RMS reprojection error	Average tie point multiplicity
Conventional	97.5%	485	0.69 pix	5.7
Autonomous	91.5%	303	0.84 pix	4.2

Table 1. Bridge A evaluation parameters

Bridge B: The conventional approach achieves 6.0% greater coverage requiring 60% more images to accomplish. Precision metrics show the conventional method reduced RMS reprojection error by 17.9%. Network robustness differed substantially, with the conventional technique exhibiting 36% higher average tie point multiplicity, see Table 2.

The lower coverage achieved in this experiment suggests that MACARONS had more difficulty handling self-occlusion within the lattice structure, potentially missing areas that are only visible through multiple indirect lines of sight. Nevertheless, the model quality remained visually acceptable, and the image count was reduced by nearly 65%, highlighting the method's efficiency and potential scalability. See Figure 17.

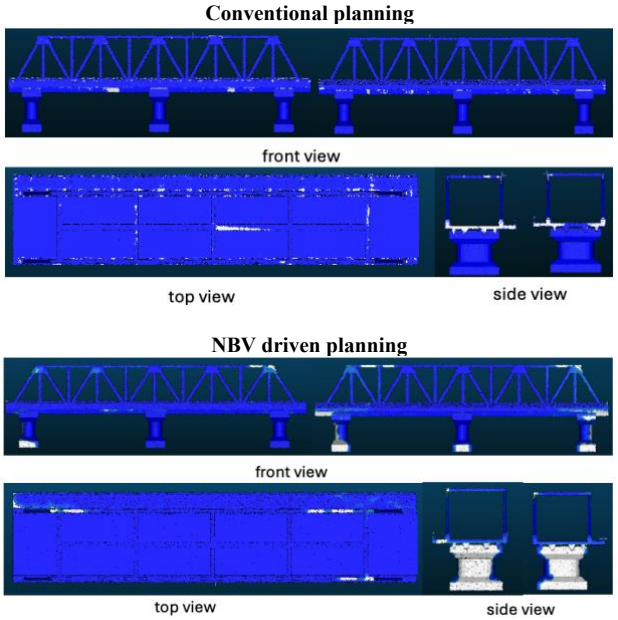

Figure 17. Coverage comparison with ground truth (white spots uncovered)

Bridge B	Bridge coverage	Images amount	RMS reprojection error	Average tie point multiplicity
Conventional	99%	1106	0.89 pix	5.2
Autonomous	88%	392	0.85 pix	4.7

Table 2. Bridge B evaluation parameters

4. Discussion

Examining the differences between conventional flight planning and autonomous NBV planning using the MACARONS framework helped to have a thorough understanding in terms of

model accuracy, efficiency, and completeness of coverage in 3D bridge reconstruction scenarios.

One of the first differences we noted is the marked decrease in the number of images needed with the autonomous planning method (37–65% reduction). In experiment one, we needed 303 images with the autonomous method, compared to 485 for the conventional method. The difference was at its highest during experiment two, where a total of 392 images were required autonomously versus 1106 with the conventional method. This indicates that the autonomous flight planning significantly increases data acquisition efficiency by decreasing the total amount of images that were redundant or of little value. What we consider efficient data acquisition is wisely selecting camera locations rather than flying in a predefined, predictable required pattern, which the drone targets image areas that haven't already been taken. However, fewer images simply corresponded to fewer tie point multiplicity and fewer overlapping observations per feature. This change can impact the robustness of matching features in photogrammetric processing.

The two planning methods generated similar RMS reprojection errors within each experiment. For the conventional flights, the RMS errors were approximately. 0.69-0.89 pixels, while the NBV autonomous method yielded RMS errors of approximately 0.84 - 0.89 pixels. While RMS reprojection error is slightly higher with the autonomous approach, this is still at an acceptable variance considering there were more irregular and adaptive flight types which could contribute marginally to distortions in the image geometry.

The conventional flights consistently had the strongest advantage in overall coverage. In both tests, conventional planning achieved 97% and 99% coverage, while the autonomous system achieved 91.5% and 88% coverage. This highlights a possible challenge of NBV approaches in constrained settings like bridges. The fact that uncovered white spots remained in the autonomous coverage maps under structural elements was also evidence of this possibility. It seems to confirm that current NBV plans like MACARONS could benefit from hybrid integration with manual guidance or predefined safety margins to guarantee full scene completeness in mission-critical applications.

In summary, these findings indicate that although conventional planning still involves better coverage and slightly better accuracy, autonomous NBV-based planning also possesses considerable potential for data-efficient and valuable solutions for bridge inspection, particularly under time constraints or any storage limitations. As autonomous method systems continue to improve, particularly their ability to reason about occlusions and the use of scene priors, the current gap in coverage may not be an issue.

While MACARONS computes Next Best Views using depth maps generated from a learned depth estimation model, our evaluation pipeline relies on conventional photogrammetric reconstruction methods based on multi-view image matching with roughly same GSD metioned in Experiment section. Specifically, we use the orientations given by MACARONS to run the dense reconstruction in Metashape. This difference introduces a key limitation in our evaluation: the quality of the reconstructed 3D model is highly dependent on successful feature matching across views.

In practice, certain surfaces such as textureless regions, repeating patterns, or areas with intense illumination changes can cause classical feature matching to fail. As a result, even if a camera viewpoint selected by MACARONS is theoretically optimal based on its internal depth predictions, the corresponding real-world or simulated RGB images may not yield sufficient 3D points using conventional photogrammetric methods. Due to the lack of cameras marked in red as shown in Figure 18 , this mismatch can lead to underestimating the true potential of the NBV strategy in our evaluation and should be considered when interpreting the results

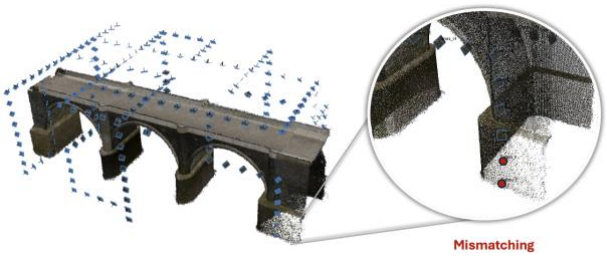

Figure 18. Mismatching using MACARONS

5. Conclusion

In this study, we set out to evaluate whether a prediction-based Next-Best-View (NBV) method specifically the MACARONS framework can perform effectively during online exploration for 3D reconstruction of complex infrastructure objects such as bridges. Our central research question asked whether this autonomous planning strategy could ensure sufficient coverage and high reconstruction quality comparable to conventional flight planning.

Based on the two controlled experiments, the autonomous method demonstrated strong potential. It was significantly more efficient in terms of image acquisition, reducing the number of images by more than half in both experiments while maintaining comparable RMS reprojection errors. The findings suggest that prediction-based NBV strategies such as MACARONS are able to assist accurate, efficient, and scalable 3D reconstruction for real-time UAV mapping. Our results did, however, highlight some important disadvantages: the autonomous method had consistently worse coverage percentages (91.5% and 88%) than conventional approaches (97.8% and 99%) and maximum reprojection errors were also found to be higher in specific areas. This may point to issues with occlusions and achieving consistent coverage, in particular in areas that are hard to see and lack distinctive texture. Thus, although consequently capable in-principle, MACARONS implementation in (or without) any additional strategies is unlikely to be acceptable in safety-critical or highly-complex inspection tasks. In order to further improve the performance and robustness of autonomous NBV methods in infrastructure mapping we suggest that further work should examine the incorporation of efficient path planning, real-time occlusion detection, and semantic scene understanding to assist more effectively with guiding the NBV selection process. Providing MACARONS with visibility prediction models or hybrid approaches that rely on guidance from the operator alongside, or instead of, autonomous decision making may also improve coverage in problematic areas.

Furthermore, it would be beneficial to investigate how MACARONS compares to other learning-based NBV techniques, whether they employ reinforcement learning, or uncertainty modelling. Such comparisons could help illuminate the trade-offs between planning speed, reconstruction quality, and computational complexity, potentially furthering the development of autonomously deployable photogrammetry systems capable of operating in a diverse set of real-world infrastructures. Path planning and further investigation is also critical aspects in drone navigation in MACARONS to increase the efficiency.

References

Alsadik, B., Spreeuwers, L., Dadrass Javan, F., Manterola, N., 2023. Mathematical Camera Array Optimization for Face 3D Modeling Application. Sensors 23, 9776. https://doi.org/10.3390/s23249776

Duchoň, F., Babinec, A., Kajan, M., Beňo, P., Florek, M., Fico, T., Jurišica, L., 2014. Path Planning with Modified a Star Algorithm for a Mobile Robot. Procedia Eng., Modelling of Mechanical and Mechatronic Systems 96, 59–69. https://doi.org/10.1016/j.proeng.2014.12.098

Felipe-García, B., Hernández-López, D., Lerma, J.L., 2012. Analysis of the ground sample distance on large photogrammetric surveys. Appl. Geomat. 4, 231–244. https://doi.org/10.1007/s12518-012-0084-2

Godard, C., Aodha, O.M., Brostow, G.J., 2017. Unsupervised Monocular Depth Estimation with Left-Right Consistency. https://doi.org/10.48550/arXiv.1609.03677

Guédon, A., Monasse, P., Lepetit, V., 2022. SCONE: Surface Coverage Optimization in Unknown Environments by Volumetric Integration.

Guédon, A., Monnier, T., Monasse, P., Lepetit, V., 2023. MACARONS: Mapping And Coverage Anticipation with RGB Online Self-Supervision.

Heise, P., Klose, S., Jensen, B., Knoll, A., 2013. PM-Huber: PatchMatch with Huber Regularization for Stereo Matching, in: 2013 IEEE International Conference on Computer Vision. Presented at the 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, Sydney, Australia, pp. 2360–2367. https://doi.org/10.1109/ICCV.2013.293

Hepp, B., Dey, D., Sinha, S.N., Kapoor, A., Joshi, N., Hilliges, O., 2018. Learn-to-Score: Efficient 3D Scene Exploration by Predicting View Utility.

Ilie, A., 2003. Computing a View Frustum to Maximize an Object's Image Area.

Kuwata, Y., Teo, J., Fiore, G., Karaman, S., Frazzoli, E., How, J.P., 2009. Real-Time Motion Planning With Applications to Autonomous Urban Driving. IEEE Trans. Control Syst. Technol. 17, 1105–1118. https://doi.org/10.1109/TCST.2008.2012116

Li, C., Yu, L., Fei, S., 2020. Large-Scale, Real-Time 3D Scene Reconstruction Using Visual and IMU Sensors. IEEE Sens. J. 20, 5597–5605. https://doi.org/10.1109/JSEN.2020.2971521

Li, S., Guédon, A., Boittiaux, C., Chen, S., Lepetit, V., 2025. NextBestPath: Efficient 3D Mapping of Unseen Environments. https://doi.org/10.48550/arXiv.2502.05378

Liu, L., Xia, X., Sun, H., Shen, Q., Xu, J., Chen, B., Huang, H., Xu, K., 2018. Object-aware guidance for autonomous scene reconstruction. ACM Trans. Graph. 37, 1–12. https://doi.org/10.1145/3197517.3201295

López, D., Felipe, B., González-Aguilera, D., Arias-Pérez, B., 2013. An Automatic Approach to UAV Flight Planning and Control for Photogrammetric Applications: A Test Case in the Asturias Region (Spain). Photogramm. Eng. Remote Sens. Volume 79, Pages 87-98. https://doi.org/10.14358/PERS.79.1.87

Mendoza, M., Vasquez-Gomez, J.I., Taud, H., Sucar, L.E., Reta, C., 2020. Supervised Learning of the Next-Best-View for 3D Object Reconstruction. Pattern Recognit. Lett. 133, 224–231. https://doi.org/10.1016/j.patrec.2020.02.024

Mousavi, V., Varshosaz, M., Remondino, F., 2021. EVALUATING TIE POINTS DISTRIBUTION, MULTIPLICITY AND NUMBER ON THE ACCURACY OF UAV PHOTOGRAMMETRY BLOCKS. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLIII-B2-2021, 39–46. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-39-2021

Santamaria, E., Pastor, E., Barrado, C., Prats, X., Royo, P., Perez, M., 2012. Flight Plan Specification and Management for Unmanned Aircraft Systems. J. Intell. Robot. Syst. 67, 155–181. https://doi.org/10.1007/s10846-011-9648-3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Watson, J., Aodha, O.M., Prisacariu, V., Brostow, G., Firman, M., 2021. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. https://doi.org/10.48550/arXiv.2104.14540

Yuan, S., Yang, B., Fang, H., n.d. Direct Root-Mean-Square Error for Surface Accuracy Evaluation of Large Deployable Mesh Reflectors, in: AIAA Scitech 2020 Forum. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2020-0935

Zeng, R., Zhao, W., Liu, Y.-J., 2020. PC-NBV: A Point Cloud Based Deep Network for Efficient Next Best View Planning, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Presented at the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Las Vegas, NV, USA, pp. 7050–7057. https://doi.org/10.1109/IROS45743.2020.9340916