3D Modeling and Rendering with a Tesla Model 3 Highland

Luigi Barazzetti[1], Mattia Previtali[1], Fabio Roncoroni[2]

[1] Dept. of Architecture, Built environment and Construction engineering (ABC),
Politecnico di Milano, Via Ponzio 31, Milan, Italy
(luigi.barazzetti, mattia.previtali)@polimi.it
[2] Polo territoriale di Lecco, via Previati 1/c, Lecco, 23900, Italy - fabio.roncoroni@polimi.it

Commission II

Keywords: 3D, Accuracy, Calibration, Modeling, Multi-camera rig, Gaussian splatting, Photogrammetry, Tesla Vision.

Abstract

This paper investigates the feasibility of using video sequences recorded by a Tesla Model 3 Highland for photogrammetric 3D reconstruction and neural rendering. The onboard cameras, originally designed for autonomous navigation, were calibrated as a multi-camera rig using bundle adjustment. The resulting intrinsic and extrinsic parameters were validated across several test projects and subsequently applied to real-world driving sequences to generate oriented image datasets, 3D mesh reconstructions, and gaussian splatting renderings. The experiments demonstrate that complex scenes can be reconstructed, although artefacts persist due to limited acquisition geometry, temporal desynchronization, compression, and dynamic scene elements. The study highlights the photogrammetric potential of consumer vehicles and provides a quantitative evaluation of Tesla Vision data for 3D applications, addressing limitations, achievable accuracy, and prospects for automated artefact correction and large-scale reconstruction from vehicular fleets. A video with selected examples is available at https://youtu.be/pOFpUp-vlHU.

1. INTRODUCTION

Since 2022, Tesla has removed ultrasonic sensors from its vehicles and now relies exclusively on Tesla Vision, a multi-camera-based autopilot system. This system processes images using AI, and an occupancy network (Mescheder et al., 2019) is currently employed in supervised Full Self-Driving (FSD) to perform various automated tasks such as route navigation, steering, lane changes, and parking (under the driver's active supervision). From a photogrammetric perspective, the system consists of synchronized cameras mounted on the front, rear, left, and right sides of the vehicle. These can be modeled as a multi-camera rig, characterized by individual camera calibration parameters and relative orientations among the cameras (Chiodini et al., 2018; Maset et al., 2024; Perfetti et al., 2024a – 2024b). However, this system was designed for autonomous driving rather than for 3D reconstruction tasks typically associated with photogrammetric applications.

The aim of this work is not to consider a Tesla vehicle as a mobile mapping system. Such systems are already commercially available and well-documented in the literature, with numerous established applications (Elhashash et al., 2022; Takahashi and Masuda, 2019; Tardy et al., 2023). This work was inspired by presentations given by Ashok Elluswamy (Tesla) during the Autonomous Driving workshops at CVPR 2022 and CVPR 2023, where he discussed the development of foundation models and the potential to generate 3D models from the customer fleet using Neural Radiance Fields (NeRFs, Mildenhakk et al., 2020).

Motivated by this vision, we set out to investigate this possibility, adopting a different methodology that combines digital photogrammetric techniques (such as multi-camera system calibration, image orientation, dense point cloud generation, and textured mesh creation, Luhmann et al. (2023)) with AI-powered 3D reconstruction methods, specifically gaussian splatting (Kerbl et al., 2023, Chen and Wang, 2024) (Figure 1). The results were validated using reference datasets obtained from total station measurements and laser scanning, evaluating accuracy, level of detail, and completeness, while also summarizing the main advantages, limitations, and potential improvements.

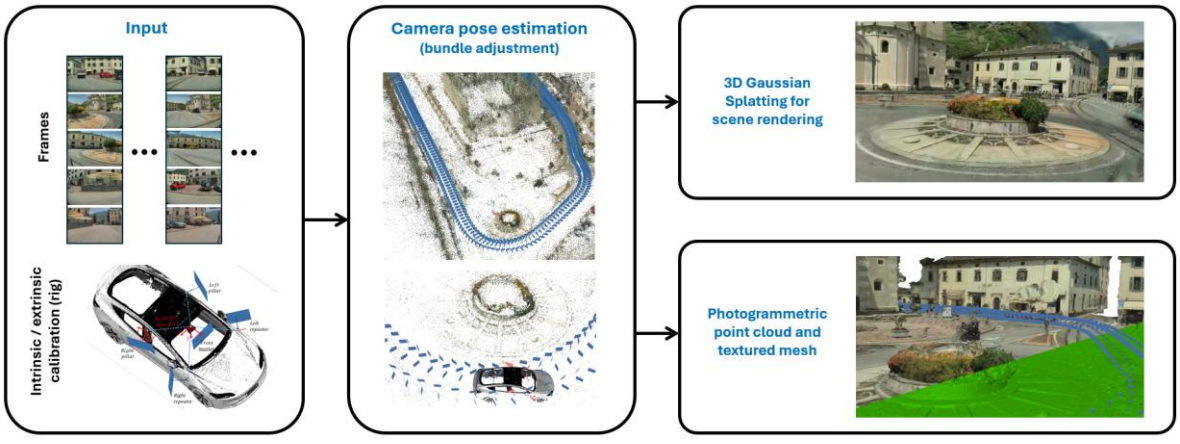

Figure 1. Overall workflow for 3D rendering and modeling using video frames captured while driving a Tesla Model 3 Highland. A video with examples is available at https://youtu.be/pOFpUp-vlHU.

A natural question arises: why has this opportunity only been explored now? Beyond the "practical factor" of gaining access to a Tesla vehicle, a key enabler was a software update introduced in 2025 that extended access to the two B-pillar cameras. Previously, only four video recordings were available: front, rear, and two side-repeater cameras. The addition of the B-pillar views now enables a complete 360° coverage around the vehicle, accessible through Sentry Mode as video files. It is important to note that users do not have access to raw image data but only to compressed mp4 files, which inevitably impacts the quality of the resulting models. The synchronization of such recordings is crucial, particularly when the camera rig is moving at significant speed. Nevertheless, the combination of these video streams made it possible to generate 3D models and rendered visualizations.

The paper is structured as follows: Section 2 outlines the calibration of the vehicle's camera rig and the experiments conducted to validate both the intrinsic parameters and relative orientation of the cameras. Section 3 presents several real-world case studies in which gaussian splatting was employed to render 3D environments. Section 4 discusses the results obtained using a traditional photogrammetric workflow, comparing them with those from other digital 3D reconstruction techniques. Finally, Section 5 discusses potential improvements and broader considerations for the entire processing pipeline, including georeferencing, error propagation, and camera model selection.

2. TESLA VISION: PHOTOGRAMMETRIC CALIBRATION AND VALIDATION

2.1 Description of the Tesla Model 3 Highland Camera Rig

The vehicle used in our experiments is a Tesla Model 3 Highland (software version 2025.20.3), from which six video recordings were obtained, as outlined in the introduction. The camera system consists of six fixed units positioned at strategic locations: one above the rear license plate, one on each B-pillar, two near the windshield above the rearview mirror (only one of which was used in this study), and one on each front fender.

Video recordings were accessed via Sentry Mode, which provides six compressed mp4 files at an approximate frame rate of 36 frames per second. The cameras differ in resolution: the front camera records at 2896×1876 pixels, whereas the remaining five operate at 1448×938 pixels. Each video lasts approximately one minute, with recordings spanning multiple minutes stored in separate folders. Although the nominal duration is consistent, minor variations were observed (e.g., 59.890, 59.893, and 59.889 seconds). While these discrepancies may appear negligible, they are critical in multi-camera setups—especially when the vehicle is in motion—as even slight temporal misalignments can affect downstream processing. In some instances, videos differed by one or two frames in total length. Compression, required to limit file size, also degrades visual quality. These recordings are not raw sensor outputs, which remain inaccessible to end users. Consequently, the quality of any derived processing is inherently constrained by the limitations of the available data.

2.2 Calibration of the Camera Rig

When the cameras are treated as a unified rig, the first step is to estimate both intrinsic and extrinsic parameters. The intrinsic parameters —including focal length (f), principal point offsets (c_x, c_y), and distortion coefficients (radial: k_1, k_2, k_3; tangential: p_1, p_2) —are specific to each camera. In contrast, the extrinsic parameters define the relative position and orientation of each camera within the rig's coordinate system. In this setup, the front camera is designated as the master, and the extrinsic parameters of the other cameras are expressed as baselines (Δx, Δy, Δz) and rotation angles (ω, φ, κ), as shown in Figure 2.

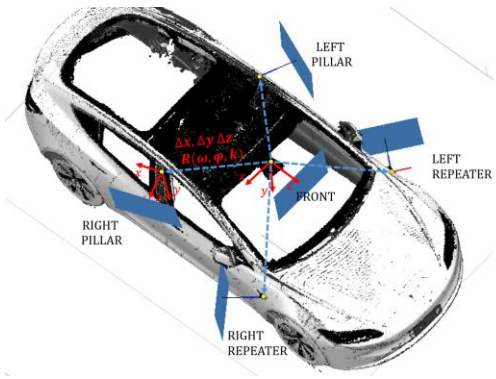

Figure 2. Camera layout on the Tesla Model 3 Highland, considered as a multi-camera rig.

The estimation of these parameters was performed through dedicated photogrammetric projects, in which the vehicle was driven in a parking lot. A bundle adjustment was used to jointly estimate both intrinsic and extrinsic parameters, along with the exterior orientation of each camera and the 3D coordinates of tie points. As is well known, photogrammetric calibration requires image blocks with suitable geometry, including roll variations and convergent views (Remondino and Fraser, 2006; Luhmann et al., 2016). In this case, the challenge is even greater, as multiple cameras are involved and both intrinsic and extrinsic parameters must be simultaneously estimated.

A key aspect of the calibration process was selecting the appropriate camera model during bundle adjustment. Due to the lack of official specifications, assumptions were made particularly for the front camera, which has a narrower field of view than the other fisheye-type cameras. It was tested both as a pinhole camera and as an equidistant fisheye (Schneider et al., 2009), with both models yielding acceptable results. However, the equidistant model was chosen for consistency, as the pinhole model required unusually large distortion coefficients, suggesting a less natural fit. The fisheye model produced smoother results with smaller parameters.

To strengthen the network geometry, the vehicle was driven along closed loops with varying trajectories, in both directions, including several transverse paths. A set of 23 ground control points (targets) was installed on the walls and pillars of the parking lot and measured using a Leica TS30, yielding 3D coordinates with a precision of approximately ±0.002 m. Control points were not directly included in the bundle adjustment but were instead used to estimate a 7-parameter similarity transformation. This allowed for scaling, rotation, and translation of the photogrammetric reconstruction to align it with the total station reference system through a rigid transformation.

Figure 3 shows four photogrammetric projects conducted in the parking lot, each featuring different network geometries and frame counts. The vehicle was driven at a speed of approximately 5–7 km/h, and frames were extracted from the videos at a

frequency of 1 frame/s. Various combinations of control points and check points were used for validation (Table 1), and the results showed comparable accuracy across all four projects.

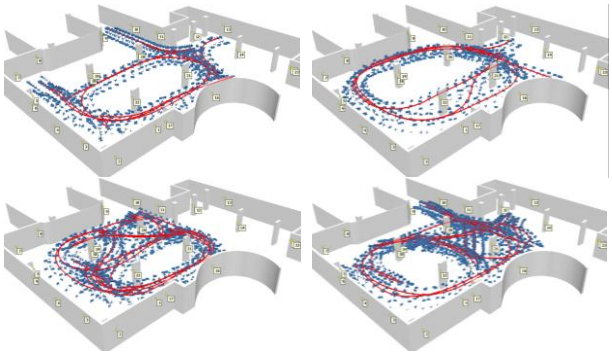

Figure 3. Overview of the four calibration projects in the parking lot, showing trajectories and corresponding camera poses.

	P1	P2	P3	P4
N. frames	252×5	252×5	315×5	441×5
RMS (pixel)	0.51	0.48	0.62	0.73
3D RMSE (m)				
23 Control	0.017	0.015	0.014	0.020
0 Check	-	-	-	-
12 Control	0.015	0.018	0.016	0.023
11 Check	0.024	0.016	0.015	0.018
6 Control	0.016	0.011	0.007	0.016
17 Check	0.033	0.028	0.029	0.026

Table 1. RMSE values obtained for the four calibration projects.

The computed extrinsic parameters of the camera rig across the different projects exhibited baseline differences of approximately 0.01–0.02 m, which align with the expected precision indicated by the variance–covariance matrix (approximately ±0.015 m). Similarly, the variations in rotation angles between projects were around 0.1°–0.2°, consistent with the estimated precision obtained after bundle adjustment.

An important consideration concerns the use of the rear camera. When its images were included in the photogrammetric projects, the results proved inconsistent, with unreliable parameter estimation. As a result, we decided to exclude the rear camera and proceed with the remaining five, which yielded stable and accurate outcomes. The cause of the poor performance associated with the rear camera remains unclear. It may stem from difficulties during image matching, though other factors cannot be ruled out. Investigating this issue would require further dedicated analysis. For the current work, we opted to rely on the five cameras that appear more critical for obtaining robust 3D reconstruction results.

2.3 Validation of the computed camera rig parameters

Two different validation procedures were carried out using the computed calibration parameters. First, projects P1, P2, and P4 were reprocessed using the intrinsic and extrinsic parameters estimated from project P3. In this approach, the intrinsic parameters for each camera were treated as fixed values, while the extrinsic parameters were incorporated with weighted constraints, assuming a precision of ±0.015 m for translations and ±0.1° for rotations. Image matching and bundle adjustment were performed again, but this time only the exterior orientation parameters (camera poses) and 3D point coordinates were estimated. Validation was conducted using the ground control targets, and the resulting RMSE values (Table 2) were consistent with those obtained in the previous section.

	P1*	P2*	-	P4*
N. frames	252×5	252×5	-	441×5
RMS (pixel)	0.58	0.56	-	0.71
3D RMSE (m)				
23 Control	0.012	0.013	-	0.027
0 Check	-	-	-	-
12 Control	0.010	0.014	-	0.031
11 Check	0.021	0.013	-	0.022
6 Control	0.008	0.009	-	0.022
17 Check	0.027	0.026	-	0.034

Table 2. RMSE values for the calibration projects reprocessed with the intrinsic and extrinsic parameters derived from P3.

An additional experiment was conducted on a completely new validation project (Figure 4), set in a different parking lot. This dataset included fewer images and lacked sequences where the vehicle was driven in both directions, thus resembling more closely a typical acquisition scenario. The project consists of 141×5 frames. The bundle adjustment yielded an overall RMS of 0.42 pixels. Ground control targets were again placed throughout the area and measured using a total station, providing a validation dataset with millimeter-level accuracy. Bundle adjustment was performed using the intrinsic and extrinsic parameters previously estimated in project P3. The resulting reconstruction was registered to the total station reference system using a rigid seven-parameter transformation. The statistics reported in Table 3 indicate that the precision again reached the centimeter level.

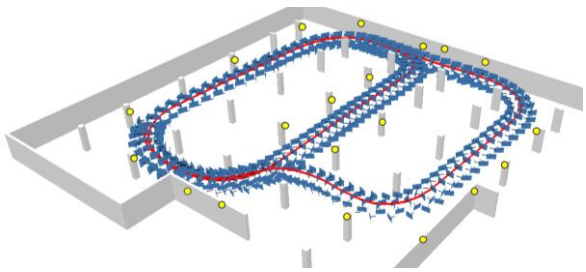

Figure 4. New validation polygon in another parking lot.

New parking lot (PL_NEW)				
	RMSE X (m)	**RMSE Y (m)**	**RMSE Z (m)**	**3D RMSE (m)**
20 Control	0.017	0.017	0.003	0.024
0 Check	-	-	-	-
10 Control	0.013	0.014	0.003	0.019
10 Check	0.022	0.034	0.004	0.041
5 Control	0.011	0.010	0.002	0.015
15 Check	0.020	0.030	0.005	0.037

Table 3. Results from a new parking lot project at a different site, using intrinsic and extrinsic rig parameters derived from P3.

3. 3D RENDERING WITH TESLA CAMERAS: RESULTS FROM ON-ROAD TESTING

We captured a variety of scenarios by driving the vehicle through different environments, including urban settings and more remote rural areas. After video acquisition, the data were processed photogrammetrically using frames extracted at a rate of 4 frames per second. Vehicle motion is fundamental for establishing a baseline between frames, particularly given that the original Tesla Vision configuration does not provide sufficient image overlap for reliable photogrammetric 3D reconstruction. However, using a constant sampling rate is not an optimal solution. Ideally, the frame extraction rate should adapt based on vehicle speed and trajectory to ensure optimal overlap. The absence of synchronized GNSS or odometry data—potentially available if raw vehicle data were accessible—further limits the ability to refine this process.

The use of pre-calibrated intrinsic and extrinsic parameters for the five-camera rig is essential, especially for frames lacking a sufficient number of keypoints for successful bundle adjustment. In such cases, intrinsic parameters were applied as fixed values, while extrinsic (relative orientation) parameters were incorporated with weighted constraints: ±0.015 m for baseline distances and ±0.1° for rotation angles—consistent with the assumptions used in the earlier validation projects. The calibration parameters were taken from project P3.

Field experiments were conducted in various locations across the Lombardy Region of Italy, including the cities of Sondrio, Lecco, and Tirano, as well as nearby villages. Figure 5 shows the locations of the experiments, the computed camera poses and the resulting 3D renderings generated through gaussian splatting.

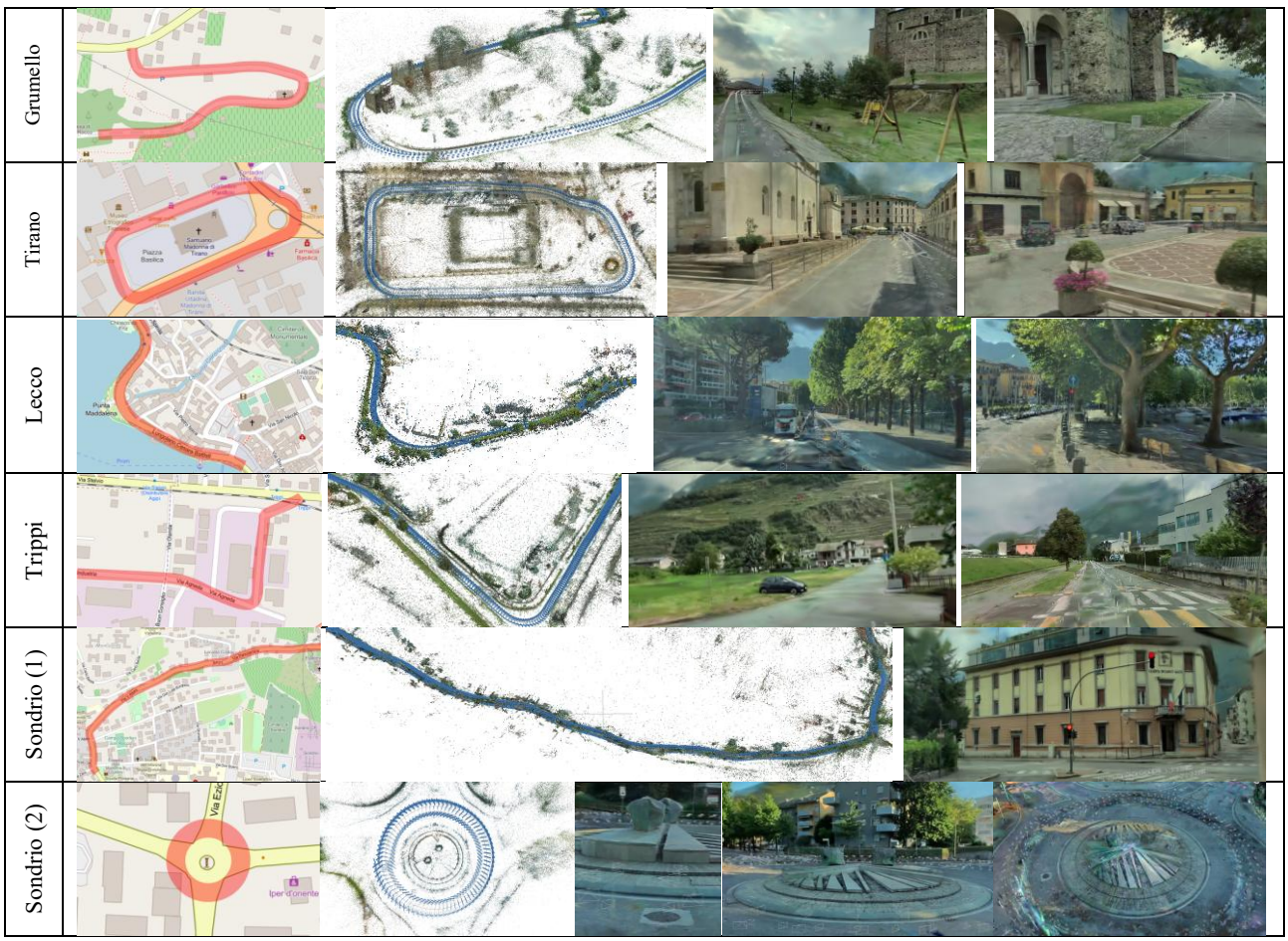

Figure 5. Examples of photogrammetrically oriented sequences using rig parameters and 3D rendering with Gaussian splatting. Results can be viewed at https://youtu.be/pOFpUp-vlHU.

The acquired sequences contain a variable number of frames and were initially oriented using a matching strategy that considers not only temporal ordering but also potential closed loops and matches between non-consecutive frames. Although more computationally demanding, this approach can lead to improved orientation results. In some cases, initial orientation attempts failed, requiring adjustments such as exhaustive image matching across all image pairs—significantly increasing processing time. Additional challenges arose from dynamic elements in the scenes, such as moving vehicles, which further complicated the estimation of camera poses. The absence of GPS data, which could provide an initial approximation of image acquisition geometry, represents a significant limitation—not only in terms of speeding up the workflow but also for improving pose robustness, especially in road-based sequences where image overlap and keypoint density may be insufficient for reliable matching. Once exterior orientation parameters were estimated via bundle adjustment, Gaussian splatting was applied using the computed camera poses, camera calibration parameters (interior orientation), and the sparse point cloud. Unlike traditional rendering methods based on meshes or dense point clouds, Gaussian splatting represents 3D points as Gaussian

distributions—visualized as translucent blobs. This approach offers significant advantages in complex scenarios, particularly for handling sparse or irregularly sampled data. It is especially effective in challenging environments with reflective or transparent surfaces, non-uniform point densities, variable object-camera distances, or inconsistent image resolutions. In some cases, gaussian splatting can yield superior visual results compared to textured meshes, and recent photogrammetric platforms have started integrating this rendering technique.

The gaussian splatting workflow was carried out using Jawset PostShot, a software specifically designed for advanced 3D rendering with gaussian primitives. To maintain consistency with the photogrammetric processing pipeline, image orientation parameters were imported from Metashape, rather than relying on PostShot's internal orientation system. This ensured alignment with the spatial relationships established during bundle adjustment. Gaussian splats were generated directly from the sparse point cloud, with each 3D point represented as a gaussian blob to produce a smooth and continuous visualization. All available frames were used in the training process, and the number of iterations was increased beyond the default value to improve rendering quality. A higher number of iterations generally allows the model to converge toward a more detailed and accurate representation of the scene. We adopted the Splat MCMC Profile, which applies a Markov Chain Monte Carlo (MCMC) method to control the number of generated splat primitives. This reduces memory and storage requirements—particularly useful when working with large datasets or limited computational resources—while maintaining high visual fidelity.

As shown in Figure 5 and the accompanying video, the final results contain visible artifacts. These are particularly evident along the road, where geometry is reconstructed primarily from the front-facing cameras. In such cases, the perspective centers follow a linear trajectory aligned with the road, which results in a suboptimal acquisition geometry. Additional artifacts appear throughout the scenes, depending on the characteristics of each environment. Notably, no manual editing or post-processing was applied: the results are presented exactly as produced by the fully automated workflow. Further discussion on artifact types and possible correction strategies is provided in the conclusions.

4. FROM THE ROAD TO 3D: SCENE RECONSTRUCTION WITH A TESLA

The accurate 3D reconstruction of real-world objects from digital images relies on a processing workflow that typically includes the generation of point clouds or mesh models, followed by texture mapping. Starting from a set of photogrammetrically oriented images, dense image matching algorithms can be used to reconstruct the external surfaces of objects. However, achieving good metric accuracy requires images with sufficient resolution and a well-designed acquisition geometry to effectively model fine details.

Video sequences recorded while driving a Tesla along a road can potentially be exploited for metric reconstruction. Nonetheless, several limitations—such as constrained camera trajectories, variable lighting conditions, and relatively low image resolution—significantly affect the final model quality. For these reasons, we chose not to apply the full photogrammetric workflow for textured mesh generation on some of the sequences introduced in Section 3, as the results would likely be of limited quality. In contrast, gaussian splatting proves more suitable for rendering complex scenes under such conditions. Photogrammetric methods, however, can still be effectively applied to selected case studies where the acquisition geometry around a specific object more closely resembles traditional survey setups used in digital documentation.

To illustrate this, we present the results obtained from the Tirano and Sondrio (2) datasets, which include imagery captured around a basilica and within a roundabout, respectively. Examples of the resulting meshes are shown in Figure 6. It is important to note that applying a texture to a surface may create the visual impression of a geometrically accurate 3D model; however, this is not always the case, as textures can visually conceal underlying geometric artifacts.

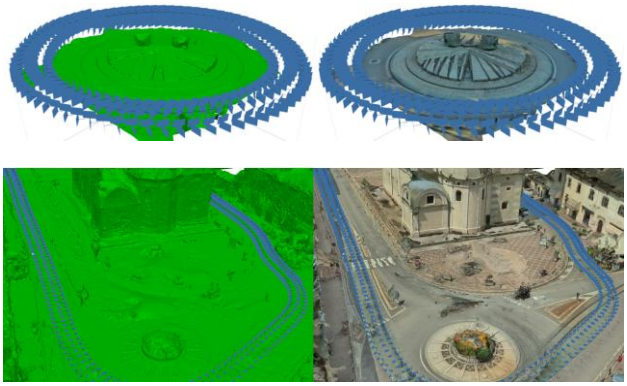

Figure 6. Textured 3D models generated from selected case studies: Tirano and Sondrio (2).

A quantitative evaluation of the achievable metric accuracy was carried out by comparing the 3D model generated from Tesla video frames with a reference laser scanning point cloud acquired using a Leica RTC360. The image sequences were recorded while driving around the Church of S. Peter in both directions, simulating a double-pass acquisition to improve the network geometry. This aspect is further discussed in the conclusion, which also explores the potential of incorporating data acquired from multiple vehicles and addresses limitations such as variable lighting conditions—an effort that clearly goes beyond the scope of a single-user project.

In this experiment, the two opposing sequences were acquired simultaneously; however, only part of the church was captured from both directions. After image orientation, a dense point cloud was generated from depth maps and then meshed to produce a surface, which was subsequently textured. The dense point cloud derived from the Tesla footage was registered to the laser scanning point cloud using the Iterative Closest Point (ICP) algorithm. A variant of ICP that estimates scale was also tested, but the resulting scale factor was about 0.99, confirming the validity of the scale derived from the camera rig's extrinsic parameters. Therefore, co-registration was finalized with a fixed (unitary) scale factor, and discrepancy maps between the two point clouds were computed, as shown in Figure 7.

The resulting error was 0.049 m ± 0.051 m, indicating an overall accuracy within a few centimeters, though with a bias. Most of the discrepancies were observed on the main façade and the rear of the church, where the acquisition geometry was less favorable—only one sequence covered the façade, and the road gradually diverged from the apse. In contrast, the long lateral wall exhibited a lower error of approximately 0.03 m, where the average distance between the car and the wall was around 20 meters. An orthophoto was also generated, with an estimated

Ground Sampling Distance (GSD) of 0.02 m, which is comparable to the observed discrepancy on the façade in the laser scanning comparison.

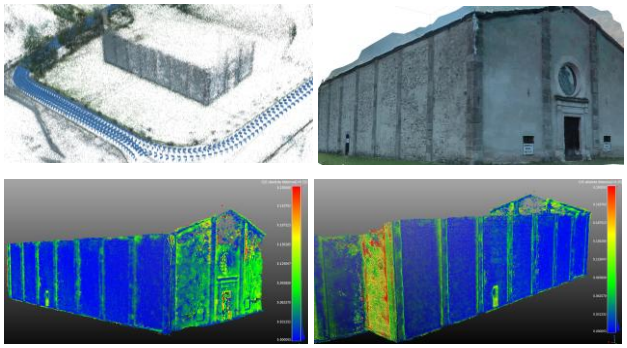

Figure 7. S. Pietro dataset and error analysis from comparison between tesla and laser scanner point clouds.

Finally, we present an illustrative example captured inside an underground parking lot (Figure 8). The dataset consists of a linear sequence of frames, from which a textured mesh was generated (top). The bottom image shows mesh confidence, evaluated as the average number of depth maps contributing to the reconstruction.

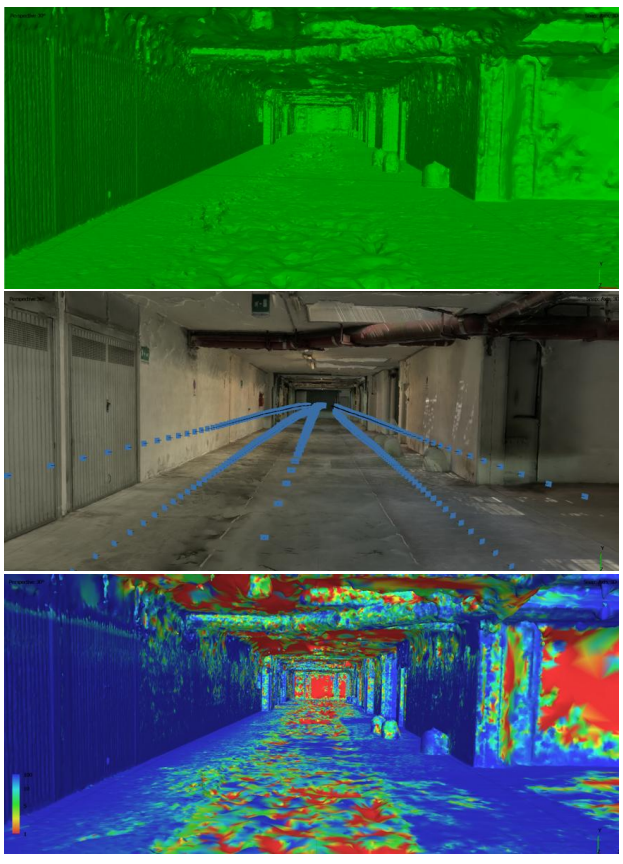

Figure 8. 3D reconstruction along a straight sequence, with confidence maps projected onto the final mesh.

As expected, confidence is higher along the sides of the vehicle—where multiple cameras provide overlapping views—and lower in areas primarily captured by the front camera alone. Notably, the road surface is among the most affected areas, highlighting the importance of acquiring data from multiple directions. This is technically feasible, as users typically drive in and out of a parking lot, creating natural bidirectional coverage.

However, a greater challenge lies in merging images acquired under varying lighting conditions—especially in external environments, where natural light changes rapidly and can significantly affect image quality and consistency. These variations can compromise feature matching and surface reconstruction, making lighting one of the most critical factors for achieving reliable results in outdoor scenarios.

5. ADDITIONAL CONSIDERATIONS

5.1 Project georeferencing and error propagation

The use of image frames and constraints derived from the rig parameters enables metric reconstructions; however, such reconstructions are generated in a generic, local reference system. Due to the absence of GNSS data, the resulting models are not georeferenced. Although one could employ a GNSS antenna synchronized with the video recordings to integrate absolute positioning, we deliberately chose to rely exclusively on image-based information, without incorporating external sensors.

A potential strategy for approximate georeferencing is the integration of Google Street View imagery, which is available along many roads worldwide. These images can be retrieved as equirectangular projections, each associated with known geographic coordinates (latitude and longitude) corresponding to their projection centers (Agarwal et al, 2015; Bruno and Roncella, 2019; Tsai and Chang, 2012). We adopted this approach by first processing the Tesla image sequences and integrating corresponding GSV images—configured with a spherical camera model—into a joint bundle adjustment. Following orientation, the project was aligned to an orthographic coordinate system (UTM Zone 32N WGS84) using a seven-parameter similarity transformation.

Figure 8 presents the results for the Grumello dataset, where an RMSE of approximately 0.42 m was achieved for Easting and Northing, and 0.10 m for elevation. A second test, conducted on the Trippi dataset, resulted in larger errors: approximately 2.81 m in horizontal position and 0.44 m in elevation. These results are consistent with the geolocation accuracy reported for Google Street View imagery in the literature, which typically ranges from several decimeters to a few meters. In summary, this georeferencing method can serve as a practical solution for approximate project localization, though its accuracy is considerably lower than that achievable with integrated GNSS systems.

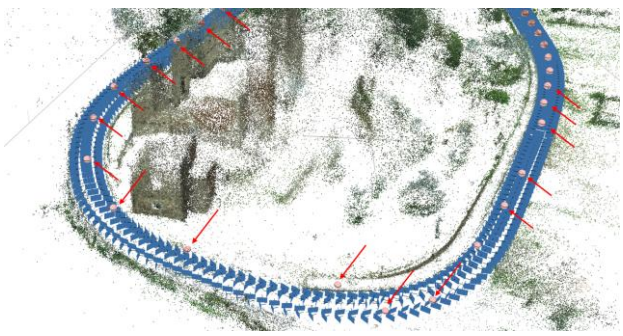

Figure 9. Combined image matching between Tesla video frames and Google Street View panoramas.

Another factor affecting metric accuracy is error propagation, which is particularly difficult to control in open sequences. Figure 10 (top) shows an example from the city of Sondrio, where a long image sequence—approximately 3,000 meters in length—was acquired along a path that includes both straight urban roads and narrow curves. The dataset consists of a total of 9,335 frames. The project was processed using a sequential approach and was manually georeferenced to align the beginning of the trajectory with the main access road to the city (on the left side of the image).

Figure 10 (bottom) presents the trajectory of the front camera overlaid on digital cartography. After a 3 km segment, the accumulated positional error reaches several meters. While improved accuracy could be achieved by forming closed loops, such configurations require a more generalized image matching strategy, rather than matching only consecutive frames.

The absence of GNSS data also impacts processing efficiency, as it prevents pre-selection of image pairs that likely share keypoints, thus increasing the computational load during initial image matching. Given that Tesla vehicles are equipped with onboard navigation systems, access to raw GNSS data and frame timestamps could significantly improve both error propagation control and overall project georeferencing. However, to the authors' knowledge, such data are currently not accessible to end users.

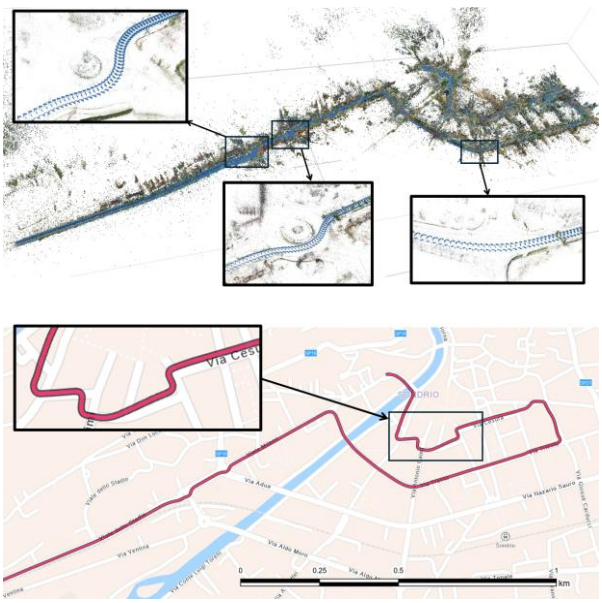

Figure 10. Error propagation along a 3 km image sequence.

5.2 Pinhole or Fisheye Camera Model?

As mentioned in Section 2, the processing was carried out using five cameras, all calibrated with the equidistant camera model. This choice was well justified for four out of the five cameras. However, the applicability of this model to the front camera raised some doubts. To support our decision, we present results from the parking lot project number 2 (P2), which was processed using two configurations: one with all five cameras treated as fisheye (equidistant model), and one where only the front camera used a frame (pinhole) model, while the remaining four were kept as fisheye.

The two configurations yielded very similar results in terms of final image reprojection error: an RMS of 0.48 pixels with five fisheye models, and 0.59 pixels when using one pinhole model (front) with four fisheye models.

The comparison with ground truth data obtained from total station measurements also showed close agreement, with overall 3D RMSE values of 0.015 m for the all-fisheye setup and 0.012 m for the mixed model configuration. Different distributions of control and check points led to consistent outcomes across both cases. The estimated relative orientation parameters (Δx, Δy, Δz and ω, φ, κ) also differed only marginally—well within the standard deviations provided by the variance-covariance matrix. The computed focal lengths for the front camera were nearly identical in both configurations: 3638.74 pixels (fisheye) and 3638.51 pixels (pinhole).

Overall, the analysis suggests that both the fisheye and the pinhole models applied to the front camera yield satisfactory results, and it is not straightforward to select the more appropriate model based solely on statistical indicators. However, a comparison of the distortion curves for the two reveals clear differences in their behavior. The pinhole model exhibits significantly stronger distortion, particularly in the radial component. The high magnitude of the distortion coefficients suggests that the model compensates for substantial lens distortion by relying on higher-order polynomial terms. In contrast, the fisheye model yields smaller and more balanced distortion coefficients, indicating that it represents the observed distortion more naturally. This behavior suggests that the fisheye model better matches the optical characteristics of the lens. Comparable results were also observed in other parking lot projects. Although these findings are not conclusive, they provide consistent evidence in favor of the fisheye model. For this reason, we continued to treat the camera rig as composed entirely of fisheye cameras in subsequent processing steps.

6. CONCLUSIONS

The experiments demonstrate that Tesla Vision imagery, although not designed for metric applications, can support reliable 3D reconstruction when rigorous calibration and orientation procedures are applied. The intrinsic and extrinsic parameters of the five-camera rig proved stable across independent tests, with baseline variations within 0.01–0.02 m and angular differences below 0.2°. When used for scene reconstruction, the calibrated rig achieved metric accuracies of a few centimeters compared with reference laser-scanning data, confirming the photogrammetric consistency of the approach. The experiments relied on compressed Sentry-Mode recordings, whereas access to original raw image data could substantially improve image quality and the accuracy of reconstruction results.

Several artefacts remain visible, primarily along linear trajectories with limited parallax and in dynamic environments containing moving objects or vegetation. Additional degradation arises from compression and slight temporal desynchronization between cameras. Future improvements may involve (i) manual or semi-automatic post-editing to remove erroneous geometry, and (ii) automated artefact detection using machine-learning or semantic-segmentation methods to classify and suppress inconsistent regions.

The orientation strategies presented in this paper rely on photogrammetric and computer vision workflows that combine image matching and bundle adjustment, ensuring both accuracy

and reliability. However, recent advances in deep learning provide alternative or complementary approaches that could be exploited or integrated into these pipelines. For example, the Vision Geometry Graph Transformer (VGGT, Wang et al. 2025) is a neural network capable of jointly inferring all the key 3D attributes of a scene—including camera parameters, point maps, depth maps, and 3D point tracks—directly from image data. Variants such as VGGT-Long (Deng et al., 2025) extend this capability to long RGB sequences, where traditional foundation models typically fail. Another development, Rig3R (Li et al. 2025), builds upon the same single-pass design but can incorporate rig metadata, enabling structure discovery even in the absence of explicit priors.

Overall, this study provides a first quantitative assessment of Tesla Vision data for photogrammetric and AI-based 3D modeling. Despite inherent limitations, the results confirm the feasibility of producing coherent and visually compelling models from compressed onboard video. Integrating odometry, GNSS, or other positional information—currently not accessible to end users—would also be beneficial for refining specific processing tasks and further enhancing overall model reliability. The experiments were conducted using a single vehicle; scaling the approach to an entire fleet would introduce additional challenges but also unlock significant potential for large-scale digitalization.

References

Agarwal, P., Burgard, W., & Spinello, L. (2015). Metric Localization using Google Street View. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3111–3118).

Bruno, N., & Roncella, R. (2019). Accuracy Assessment of 3D Models Generated from Google Street View Imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W9, 181–188.

Chen, G., & Wang, W. (2024, January 8). A Survey on 3D Gaussian Splatting. arXiv preprint arXiv:2401.03890, 22 pages.

Chiodini, S., Pertile, M., Giubilato, R., Salvioli, F., Barrera, M., Franceschetti, P., & Debei, S. (2018). Camera Rig Extrinsic Calibration Using a Motion Capture System. In 2018 5th IEEE International Workshop on Metrology for AeroSpace (MetroAeroSpace).

Deng, K.; Ti, Z.; Xu, J.; Yang, J.; Xie, J. (2025). VGGT-Long: Chunk it, Loop it, Align it – Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences. arXiv preprint arXiv:2507.16443.

Elhashash, M., Albanwan, H., & Qin, R. (2022). A Review of Mobile Mapping Systems: From Sensors to Applications. Sensors, 22(11), 4262. https://doi.org/10.3390/s22114262

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG), 42(4), 1–14.

Li, S., Kachana, P., Chidananda, P., Nair, S., Furukawa, Y. & Brown, M. (2025). Rig3R: Rig-Aware Conditioning for Learned 3D Reconstruction. arXiv preprint arXiv:2506.02265.

Luhmann, T., Fraser, C., Maas, H. G. (2016). Sensor modelling and camera calibration for close-range photogrammetry. ISPRS Journal of Photogrammetry and Remote Sensing 115, 37–46.

Luhmann, T., Robson, S., Kyle, S., & Boehm, J. (2023). Close-Range Photogrammetry and 3D Imaging (4rd ed.). Berlin, Boston: De Gruyter. 820 pages.

Maset, E., Magri, L., Toschi, I., and Fusiello, A. (2020). Bundle block adjustment with constrained relative orientations. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2020, 49-56. https://doi.org/10.5194/isprs-annals-V-2-2020-49-2020

Maset, E., Magri, L., Fusiello, A. (2024). Principled bundle block adjustment with multi-head cameras. ISPRS Open Journal of Photogrammetry and Remote Sensing, 11, 100051, 12 pages.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A. (2019). Occupancy Networks: Learning 3D Reconstruction in Function Space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4460–4470).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 1–17).

Perfetti, L., Bruno, N., & Roncella, R. (2024a). Multi-Camera Rig and Spherical Camera Assessment for Indoor Surveys in Complex Spaces. Remote Sensing, 16(23), 4505.

Perfetti, L., Fassi, F., & Vassena, G. (2024b). Ant3D—A Fisheye Multi-Camera System to Survey Narrow Spaces. Sensors, 24(13), 4177.

Remondino, F. & Fraser, C. (2006). Digital camera calibration methods: considerations and comparisons. In Image Engineering and Vision Metrology, ISPRS Commission V Symposium, Dresden, 25–27 September 2006 (Vol. XXXVI, Part 5, pp. 266–272).

Schneider, D., Schwalbe, E., Maas, H.-G. (2009). Validation of geometric models for fisheye lenses. ISPRS Journal of Photogrammetry and Remote Sensing, 64(3), 259–266.

Takahashi, G., & Masuda, H. (2019, June). Trajectory-Based Visualization of MMS Point Clouds. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W13, 1127–1133.

Tardy, H., Soilán, M., Martín-Jiménez, J. A., & González-Aguilera, D. (2023). Automatic road inventory using a low-cost mobile mapping system and based on a semantic segmentation deep learning model. Remote Sensing, 15(5), 1351

Tsai, V. J. D., & Chang, C.-T. (2012). Three-dimensional positioning from Google Street View panoramas. IET Image Processing, 6(6), 675–684.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. (2025). VGGT: Visual Geometry Grounded Transformer. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–12.