Point Cloud Digital Terrain Modeling from Video Data in Rail environment

ARTUR DYLEWSKI, University of Twente, The Netherlands

Digital Elevation Models (DEM), represent the topography of the bare Earth and can be processed using computer software. These models are often utilized in industries such as transportation and engineering for creating digital twins and ensuring the efficient operation of machinery and infrastructure. This paper discusses the methodology for generating Digital Terrain Models (DTMs) within rail infrastructure using monocular video footage. DTMs being DEMs that exclude man-made infrastructure and vegetation. The techniques covered include GPS interpolation, feature matching and extraction, triangulation, and computer vision models for converting DEMs into DTMs. The generation and comparison of point clouds are demonstrated, showing acceptable accuracy for given context. The results indicate that this technology has potential for low to medium accuracy use cases, though it has certain limitations that are discussed. The paper covers the technical challenges and solutions associated with this approach and compares its characteristics to the current standard, LIDAR. The main takeaway for future research is that this method is a viable alternative to LIDAR, but achieving high accuracy requires significant effort in setting up high-quality systems and using more reliable object detection models with more training data.

Additional Key Words and Phrases: Digital terrain model, Point Cloud, Video topography extraction

1 INTRODUCTION

Rail transportation systems play an important role in modern society, facilitating efficient movement of goods and passengers across vast distances. The reliability and safety of these systems are paramount, necessitating robust modeling techniques to optimize maintenance. Traditional methods of assessing rail infrastructure often rely on labor-intensive surveys or costly LIDAR (Light Detection and Ranging) technology. However, recent advancements in computer vision and photogrammetry have opened up new avenues for capturing detailed spatial information through video data.

This paper displays an approach to DTM (Digital Terrain model) modeling in rail infrastructure by using video-derived point clouds. By leveraging video footage captured from trains, the aim is to reconstruct detailed three-dimensional representations of rail assets and their surrounding environments, creating a DEM (Digital Elevation model). Following successful generation of a DEM, this paper presents solutions for converting a DEM into a DTM (see Figure 1), which entails removing the infrastructure and vegetation from the feature extraction and matching. This approach offers several advantages over conventional methods, including cost-effectiveness.

In this paper, we will first review existing literature on infrastructure modeling techniques, mentioning the limitations of current approaches and the potential benefits of utilizing video-derived point clouds. We will then present a pipeline for generating point clouds from video data, highlighting key steps such as camera calibration, feature tracking, and point cloud reconstruction. Additionally, research in the direction of converting Digital Elevation models to Digital Terrain Models will be explored.

Overall, this paper aims to demonstrate the potential of video-derived point clouds as a valuable tool for enhancing service modeling in rail infrastructure by providing detailed and accurate spatial information.

2 RESEARCH QUESTIONS

The main research question that will be tackled in this paper is the following: ***How can video-generated point clouds be utilized as a viable alternative to LIDAR scans for generating Digital Elevation models in the context of rail infrastructure?*** In addressing the main question, the related sub-questions and adjacent topics will be explored.

2.1 Sub-research Question 1:

What are the limitations and constraints of using video-based methods for generation in comparison to LIDAR?

2.2 Sub-research Question 2:

How do the precision characteristics of video-generated point clouds compare to those of LIDAR, and are they sufficient for modeling purposes within this specific context?

2.3 Sub-research Question 3:

How can machine learning and computer vision algorithms be applied to automate the process of transforming the Digital Elevation models to Digital Terrain models?

3 RELATED WORKS

In this section, the relevant literature pertaining to this topic of interest will be reviewed. Photogrammetry is a well-researched topic, particularly in fields like social engineering. While stereo cameras are commonly used, research on depth estimation and point cloud creation using monocular cameras is also substantial. Below, the specific topics relevant to this project will be explored.

(1) In their paper, Wang et al.[6] provide a solution for making DTMs with monocular cameras from aerial photos that provide very satisfactory results. They investigate into more specifics and technologies that they applied. Some of their techniques could transfer in the use of videos made from ground level.
(2) Ekström [2] shows that the validity of creating a system that develops a depth estimation of road infrastructure is possible with the use of monocular cameras and Structure From Motion (SFM) approach on the condition that it is not something to be used in real time.

TScIT 41, July 5, 2024, Enschede, The Netherlands

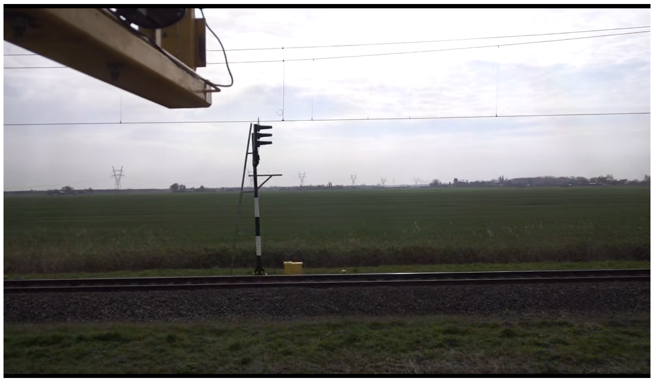

Fig. 1. Frame from the video

(3) The creation of point clouds from video has been a longstanding topic of discussion, with existing datasets showcasing well-researched methods for generating point clouds and depth estimation. For instance, the KITTI validation dataset [1] serves as a comprehensive resource in this regard.
(4) In 2019, Weng, X., and Kitani, K [5]. introduced a solution that combines image recognition from 2D images with point cloud object recognition from monocular images. This approach offers a reliable and rapid object detection method, which could prove to be an intriguing solution when contemplating the transformation of DEM into DTM.

4 METHODOLOGIES

This part of the study details the foundational structure that this research paper will follow. It begins by describing the setup and datasets. Next, it explains the procedure for carrying out the experiments and specifies the evaluation metrics used.

4.1 Datasets

4.1.1 Strukton Leonardo Video. Strukton possesses an old stomping train equipped with multiple cameras that capture data from various viewpoints. This setup forms the primary dataset for forthcoming operations and testing. Each video is accompanied by GPS data for the train and cameras, along with calibration details for all cameras. In theory, this dataset contains all the necessary information to perform photogrammetry. For this paper a sideways view will be the main focus as seen in Figure 1.

4.1.2 LIDAR Point Clouds from SpoorInBeeld. As a benchmark, LIDAR point clouds of regions will be utilized, sourced from the ProRail Spoorbeeld open dataset. These LIDAR point clouds provide high-resolution, precise spatial data, serving as a standard for evaluating the accuracy and performance of other datasets.

4.2 Procedure

This section introduces a pipeline for creating DSM point clouds and transforming them into DTMs. Each step of the process is discussed in detail. An overview of the pipeline is provided in Figure 2.

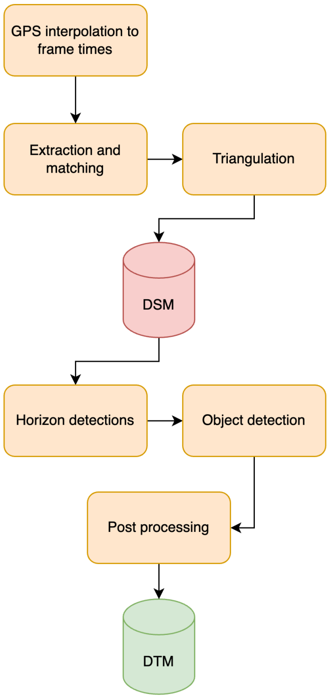

Fig. 2. Pipeline for DTM generation

4.2.1 GPS Data Interpolation. GPS coordinate interpolation and matching with video involve the alignment and synchronization of video frames with geographical locations. Interpolation, which entails estimating values between recorded data points, facilitates the precise determination of each video frame's location. This technique is particularly beneficial in situations where GPS data points are sparse or insufficiently frequent to accurately correspond with the video frames. For instance, during the acceleration of a train, minor speed variations can result in significant scale distortions. Therefore, this method ensures higher-precision measurements and accurate spatial representation.

4.2.2 Feature Extraction and Matching. Feature extraction plays a pivotal role in image analysis by identifying distinctive points or regions within an image. Subsequently, feature matching compares these extracted features across multiple images to identify commonalities between them. Within this pipeline, feature matching is a critical step as it facilitates the identification of shared features across images. Initially, conventional SIFT algorithms were employed for feature detection and matching, but their efficacy proved inadequate for the complexity of this use case. This primarily due to a similar repeating environment. Consequently, a transition to the DNN-based LightGlue[8] detection algorithm was undertaken to ensure both accuracy and speed. In this use case, generating dense point clouds is prioritized over the speed of feature matching-extraction. Therefore, the parameters for limiting the number of matches are turned off or set to the maximum possible value.

4.2.3 Terrain Detection Image Recognition Model. Within the extensive array of computer vision tools, certain methods may facilitate

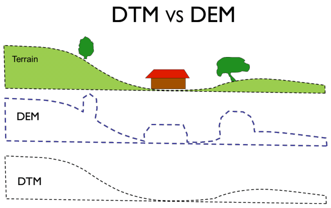

Fig. 3. Difference between DEM and DTM. (Wiki du Master Géographies Numériques, 2018).

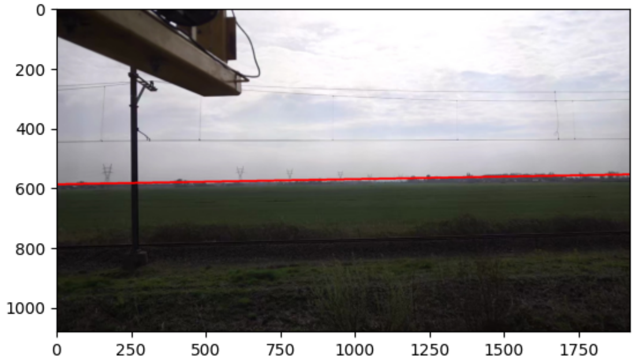

Fig. 4. Horizon detection results

the conversion of a DEM to a DTM (Figure 3) without necessitating the standard compute intensive point cloud segmentation[12]. This discussion explores techniques that may prove effective in accomplishing this task.

(1) Horizon Detection Algorithm - The most straightforward yet susceptible to inaccuracies method involves employing a horizon detection algorithm. This technique utilizes a basic segmentation algorithm, leveraging an existing library[9]. While it yields helpful results, it is not entirely reliable as it frequently includes infrastructure elements such as traffic lights and trees as part of the horizon. Despite these limitations, it serves as a rapid and lightweight preliminary step towards generating a DTM. Additionally a correction was made by pulling down the horizon line 5% of the image height, due to the fact that the interest is of the area close to the rail tracks not the actual horizon. The results of this be viewed in Figure 4.

(2) YOLOv8n, pretrained on the OpenImagesV7[10] dataset - includes several labels relevant to our task. It will be used to identify and eliminate vertical obstacles during the DTM generation process. Specifically, we will focus on labels such as "Tree", "Street Light", "Street Sign", and "Traffic Light", as they are the most likely to correspond to the trees and poles obstructing the view. The model that will be used is YOLOv8n(nano) because of its efficiency in terms of speed and good accuracy.[13]

4.2.4 Triangulation. Triangulation is a technique utilized in computer vision and photogrammetry to calculate the 3D coordinates of a point in space by using two or more images taken from different viewpoints. By extracting features from these images and triangulating their coordinates, a dense collection of points known as a point cloud is generated.

4.2.5 Camera Projection Matrix Calculation. To proceed with triangulation, it is necessary to create projection matrices that incorporate both the intrinsic and extrinsic parameters of the camera. The intrinsic parameters refer to the internal characteristics of the camera, such as focal length and sensor size. The extrinsic parameters include spatial data, such as the translation and rotation of the camera. The intrinsic will be provided with the dataset while the extrinsic are going to be calculated from the GPS data.

4.2.6 Visualisation, post processing. Given the likelihood of noise in feature matching, post-processing, such as statistical outlier removal, proves advantageous. Visualization serves as a valuable tool for evaluating point cloud performance through visual examination. As the focus is on dense point clouds that are situated in a close proximity to each other the method of statistical post processing here is radius outlier removal.

4.3 Evaluation

The video sequences will be split into manageable segments. Leveraging the publicly available LIDAR point clouds from SpoorInBeeld for evaluation, both sets of point clouds, being in homogeneous coordinates, necessitate manual alignment as a prerequisite for the evaluation process. Consequently, preliminary manual alignment of the samples will be conducted, followed by an automated matching procedure. The evaluation criteria guiding the assessment comprise:

(1) Proportion of points within a specified threshold
(2) Minimum distance recorded
(3) Maximum distance recorded
(4) Point with the highest data density
(5) Mean absolute distance
(6) Standard deviation of absolute distance

4.4 Environment

In this paper, Python, leveraging multiple scientific libraries, serves as the basis for developing the pipeline. Utilizing Jupyter Notebooks for interactive development. CloudCompare is used for analysis of generated point clouds and visualization of data.

5 COMPARASION TO LIDAR AND LIMITATIONS

5.1 Comparasion to LIDAR

This part will discuss the different aspects of characteristics and ways of data collection both from video photogrammtery as well as LIDAR scans. The comparasion can be seen in Table 1

5.1.1 Relevant takeaways. Photogrammetry is considerably cheaper and, when LIDAR scanners are out of range price-wise, it becomes a viable alternative that can produce appropriate results. However, LIDAR is superior for precise use cases, making photogrammetry not suitable for all applications. In present context, a potential drawback of photogrammetry is its high sensitivity to light conditions, unlike LIDAR, which remains unaffected. This makes LIDAR particularly advantageous in environments such as tunnels, or footage taken in the evenings.

One advantage of photogrammetry over LIDAR scanners is that it immediately captures color values. This is beneficial for applications such as service modeling, where the models are viewed by humans. Providing RGB values allows for clearer identification of the environment.

Aspect	**Monocular Photogrammetry**	**LIDAR Scanning**
Accuracy	High-resolution, well calibrated images provide detailed surface information	Highly accurate point clouds
Precision	Sensitive to image quality and calibration	Direct distance measurement
Range	Limited by camera lens and distance from the object	Long-range scanning capability
Data Density	Highly dependent on image resolution	Dense point cloud
Lighting Conditions	Performance may degrade in poor lighting conditions	Unaffected by lighting conditions
Costs	Lower cost due to use of consumer-grade cameras	Higher initial costs due to specialized equipment
Color	Includes RGB values	Does not include RGB values

Table 1. Comparison of Monocular Photogrammetry and LIDAR Scanning Point Cloud Results

5.2 Limitations of current setup

There are several limitation characteristics of this approach in its current state, they are listed below and it is discussed how it can change the accuracy:

(1) Inaccuracies in calibration - The cameras are used for numerous different purposes, leading to frequent recalibration with changes in focal lengths, recording modes, and auto focus settings. As a result, achieving a highly accurate intrinsic calibration of the cameras is nearly impossible. In this case, a default calibration from a separate camera is used, introducing a certain margin of error.

(2) Vibrations on the train - The movement and vibrations of the train introduce additional errors

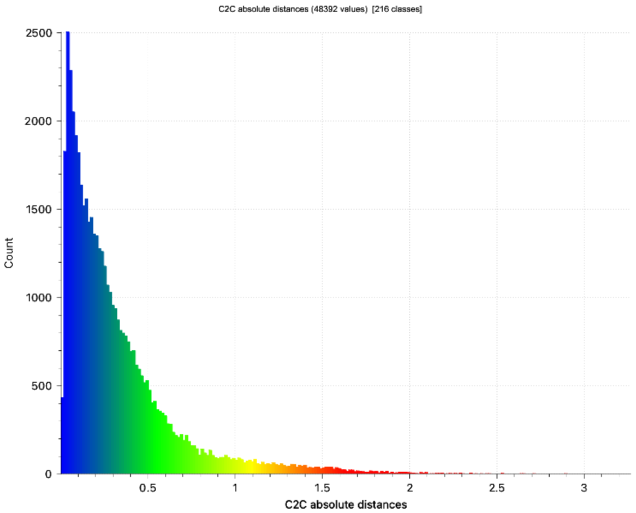

Fig. 5. Histogram showing results

(3) Internal software stabilization - As previously mentioned, the camera footage is used for various services by Strukton, making it impossible to turn off the stabilization. In the context of photogrammetry, this stabilization reduces precision and generally degrades performance[11].

(4) The interpolation of GPS data - poses challenges in contexts where the train accelerates or decelerates (Generally more sensitive in at low speeds). This difficulty arises because scale of the point clouds are dependent on the precise distance traveled between frames. Consequently, any inaccuracies in the interpolation process can result in operational malfunctions.

(5) Vegetation - around the points of interest poses a significant challenge for accuracy, as the feature matching algorithms can be disrupted by it. Additionally, reference point clouds used to measure accuracy are created at different times of the year or day, leading to discrepancies in grass levels growth between the footage, which can introduce further errors.

6 RESULTS

6.1 Leonardo Videos

A point cloud compiled from 71 frames of a video was initially manually aligned with the corresponding spot from a SpoorInBeeld point cloud in CloudCompare. Following this, a tool for fine cloud-to-cloud matching was used to accurately align the point clouds. Subsequently, a cloud-to-cloud distance computation (C2C) was performed. The histogram of point distances is shown in Figure 5, and the statistics of this test are presented in Table 2. This figure 6 shows the relative point-to-point distances between the generated point cloud and the reference point clouds. The colors used to indicate these distances correspond to the colors of the histogram from Figure 2. Further practical implications are detailed in the discussion section.

Table 2. Distance Statistics

Metric	Value
80% of points within	0.5 m
Min distance	0.0 m
Max distance	3.3 m
Distance with most counts	0.04 m
Mean distance	0.32943 m
Standard deviation	0.33728 m

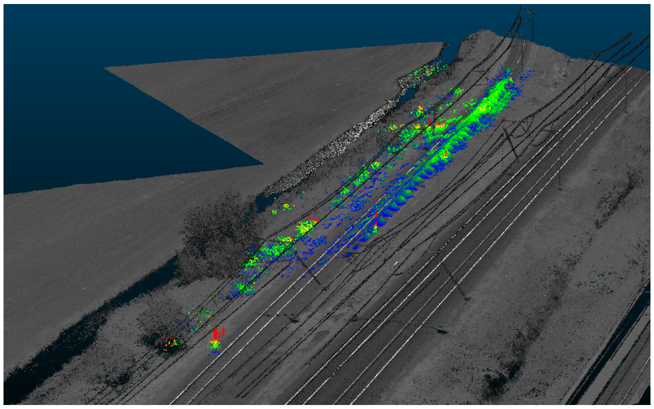

Fig. 6. Comparasion between generated point clouds and SpoorInBeeld LIDAR Point Clouds

7 DISCUSSION

7.1 General Observations

In general, upon examining the generated point clouds and the results of comparison, it can be concluded that the reconstruction of the terrain around the tracks was successful. The primary discrepancies observed in the cloud-to-cloud comparison were mainly due to changes in vegetation growth between the time of video capture and the reference point clouds. Some mislabeling from the object detection algorithm were also noted, contributing to a slight loss of points. With discrepancies in mean distance and standard deviation around 0.3 meters.

7.2 Practical Implications

To answer the main research question of this paper: *How can video-generated point clouds be utilized as a viable alternative to LIDAR scans for generating Digital Terrain models in the context of rail infrastructure?* a pipeline and replicable methodology for generating Digital Elevation Point Cloud models without relying on costly LIDAR scanners is proposed. This approach being particularly useful for businesses and researchers who lack the resources to invest in expensive LIDAR technology.

The paper additionally discusses the differences and characteristics of both technologies. Section 5 provides an analysis to determine the environments and use cases where photogrammetry can serve as an appropriate alternative to LIDAR, thereby addressing SQ1.

To answer SQ3, the paper introduces a unique method for converting DEM to DTM using 2D computer vision tools, such as horizon detection algorithms and the YOLO object detection algorithm. The YOLOv8n (nano) model [13], which is a model focused on efficiency and speed, is employed to achieve a more computationally efficient solution compared to the traditional point cloud segmentation, which remains a computationally expensive process [12].

7.3 Usability for Service Modeling

The purpose of this study is to align collected data with a digital twin of the rail infrastructure in the Netherlands. This involves accurately placing various infrastructure elements such as lights, poles, and service equipment in their correct coordinates, particularly in challenging environments like stations, urban areas, and tunnels. The primary objective is to provide a comprehensive overview of the surroundings along the railway tracks, enabling efficient object placement. While achieving very high precision is not critical in this context, the inherent limitations of the chosen method are deemed acceptable for the scope of this project which gives and answer to SQ2.

7.4 Dataset Quality

As previously discussed, the data collected from the train serves multiple purposes but requires frequent adjustments in settings, posing a significant challenge for photogrammetry. Precise calibration is instrumental to ensure accurate results, and while efforts were made to approximate optimal settings, precise calibration of each camera would be preferred in an ideal scenario.

Overall, the dataset exhibits significant variability across various dimensions such as time of day, lighting conditions, weather, varying vegetation, and changes in infrastructure between different regions. These variations present a considerable challenge for the object detection algorithm to perform effectively and consistently.

7.5 Future Works

In future research, exploring multiple image triangulation could be a promising approach. Triangulation, as a method capable of integrating more than two images, offers the potential to achieve a more precise and densely represented geometry. This avenue holds promise for enhancing accuracy in the point clouds.

An alternative approach of using stereo vision can enhance the accuracy of this study, leveraging multiple viewpoints from the train that overlap and provide varying perspectives. This method could offer a richer geometric dataset compared to monocular videos alone, enabling more detailed and precise reconstruction of the environment and infrastructure along the railway tracks.

The object detection algorithm for identifying traffic poles and trees could be enhanced by transitioning to a segmentation algorithm. This approach would minimize the loss of data points around the infrastructure, thereby improving the accuracy and completeness of the detection process.

Moreover, addressing the issue of inaccuracies in data collection setups is a big part of the process. Future experiments could benefit significantly from a more precise setup incorporating error-reduction methods. These methods might include the use of gimbals for stability, precise GPS data acquisition, and accelerometers to minimize errors caused by setup instability. Implementing such improvements could lead to more reliable and robust experimental results.

8 CONCLUSION

To address the research question of how video-generated point clouds can serve as an alternative to LIDAR scans for generating DTMs in the context of rail infrastructure, this paper has explored the viability of monocular photogrammetry. While monocular photogrammetry shows to be a feasible method for generating semi-accurate point clouds and Digital Elevation models, it does come with limitations from the technology used in this application. Particularly, its sensitivity to calibration and susceptibility to errors caused by train vibrations, along with reduced precision due to software stabilization.

It is essential to acknowledge that LIDAR technology, despite its higher cost, remains superior for achieving accurate point cloud surface reconstructions. However, monocular photogrammetry has relevance in scenarios where high precision is not crucial. Nonetheless, for applications demanding high accuracy, LIDAR should remain the primary technology.

Regarding the transformation from DEM to DTM, employing object recognition models such as YOLO presents a viable yet not foolproof solution. The variability in infrastructure and vegetation necessitates extensive training data and effort to develop a robust model for reliable detection.

In summary, monocular photogrammetry for DTM reconstruction using triangulation offers a viable solution. However, achieving high utility and precision requires meticulous hardware preparation and additional development efforts, as outlined in the future works section.

9 ACKNOWLEDGEMENTS

Sjoerd Tiecken from Team Leonardo for assistance with calibration, data access, and their overall help with the project.

Wouter Groen from Strukton for providing feedback, and support in getting the topic.

10 REFERENCES

(1) Liao, Y., Xie, J., & Geiger, A. (n.d.). *KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D.* arXiv.org. https://arxiv.org/abs/2109.13410

(2) Ekström, M. (2018). *Road surface preview estimation using a monocular camera.* DIVA. https://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A1253882&dswid=2924

(3) Weng, X., & Kitani, K. (2019). *Monocular 3D Object Detection with Pseudo-LIDAR Point Cloud.* https://openaccess.thecvf.com/content_ICCVW_2019/html/CVRSUAD/Weng_Monocular_3D_Object_Detection_with_Pseudo-LIDAR_Point_Cloud_ICCVW_2019_paper.html

(4) Rashidi, A., Brilakis, I., & Vela, P. A. (2015). *Generating Absolute-Scale point cloud data of built infrastructure scenes using a monocular camera setting. Journal of Computing in Civil Engineering, 29(6).* https://doi.org/10.1061/(asce)cp.1943-5487.0000414

(5) Zeng, W., Karaoglu, S., & Gevers, T. (2018, December 4). *Inferring Point Clouds from Single Monocular Images by Depth Intermediation.* arXiv.org. https://arxiv.org/abs/1812.01402

(6) TerrainFusion: Real-time Digital Elevation model Reconstruction based on Monocular SLAM. (2019, November 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8967663

(7) Ku, J., Pon, A. D., & Waslander, S. L. (2019). *Monocular 3D object detection leveraging accurate proposals and shape reconstruction.* https://openaccess.thecvf.com/content_CVPR_2019/html/Ku_Monocular_3D_Object_Detection_Leveraging_Accurate_Proposals_and_Shape_Reconstruction_CVPR_2019_paper.html

(8) Lindenberger, P., Sarlin, P., & Pollefeys, M. (2023, June 23). *LightGlue: Local feature matching at light speed.* https://arxiv.org/abs/2306.13643

(9) Sall, S. (n.d.). *Sallamander/Horizon-Detection: A lightweight utility for detecting the horizon line in natural images.* https://github.com/sallamander/horizon-detection

(10) Kuznetsova, A., Rom, H., et al. (2020). *The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision (IJCV).*

(11) Nocerino, Erica Menna, Fabio Verhoeven, Geert. (2022). *Good vibrations? How image stabilisation influences photogrammetry. ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences.* XLVI-2/W1-2022. 395-400. 10.5194/isprs-archives-XLVI-2-W1-2022-395-2022.

(12) Grandio, J., Riveiro, B., Lamas, D., Arias, P. (2023). *Multimodal deep learning for point cloud panoptic segmentation of railway environments.* Automation in Construction, 150, 104854. https://doi.org/10.1016/j.autcon.2023.104854

(13) Jocher, G., Chaurasia, A., Qiu, J. (2023). *Ultralytics YOLO (Version 8.0.0) [Computer software].* https://github.com/ultralytics/ultralytics

A APPENDIX

A.1 Ai Tool use

During the preparation of this work the author(s) used ChatGPT to ensure coherence and correct scientific language. All content provided to ChatGPT was original and created by the author(s). ChatGPT was utilized purely for linguistic and coherence purposes. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.