Integrating Computer Vision Models into SfM Photogrammetry: An Automated Pipeline for Occlusion Mask Generation

Riccardo Tavolare(✉)

Politecnico di Bari, 70125 Bari, Italy
riccardo.tavolare@poliba.it

Abstract. This paper presents an automated and scalable pipeline for occlusion mask generation in Structure-from-Motion (SfM) photogrammetric workflows, addressing a well-known limitation related to dynamic elements, such as pedestrians, vehicles, and other transient objects, that introduce geometric artifacts and reduce the metric reliability of 3D reconstructions. To overcome the inefficiencies and subjectivity of manual masking, the proposed approach integrates pre-trained computer vision models into the preprocessing stage, enabling a reproducible solution that does not require domain-specific training data. The pipeline is structured into three modular components. The first module performs automated image quality assessment based on sharpness, exposure, and resolution, ensuring dataset consistency and reducing error propagation during feature extraction. The second module implements a detect-then-segment strategy, combining instance-level detection via YOLOv8-seg with optional pixel-wise refinement using the Segment Anything Model (SAM2), improving segmentation accuracy in high-resolution and close-range scenarios. The third module handles automatic mask generation and formatting for compatibility with major SfM platforms, including morphology-based post-processing to enhance robustness. A proof-of-concept implementation was developed in Python using exclusively pre-trained models, allowing immediate applicability while minimizing computational and operational overhead. Experimental results highlight the efficiency of the automated masking process, while maintaining improved or comparable reconstruction quality. The approach proves particularly effective in architectural and cultural heritage surveys conducted in complex and uncontrolled environments.

Keywords: Structure-from-Motion · Photogrammetric Masking · Computer Vision

1 Introduction

Short-range photogrammetry and the Structure-from-Motion (SfM) technique have established themselves as standard practices in architectural surveying, cultural heritage documentation, urban mapping and the inspection of civil infrastructure. Continuous improvements in photogrammetric software, combined with the growing availability of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16762, pp. 106–122, 2027.
https://doi.org/10.1007/978-3-032-30524-4_8

high-resolution consumer cameras and Unmanned Aerial Vehicles (UAVs), have drastically reduced the technical barriers to producing georeferenced point clouds and accurate textured mesh models [1].

Despite these advances, the quality of SfM-based reconstructions depends critically on the composition and consistency of the photographic dataset. One of the most persistent sources of error in photogrammetric pipelines is the presence of non-stationary objects (such as pedestrians, cyclists, parked or moving vehicles, and other transient elements) within the scene at the time of acquisition. Such elements introduce inconsistencies during the feature matching phase, as the same physical point is projected onto different positions in the overlapping images, violating the fundamental assumption of scene rigidity upon which block-based adjustment relies [2, 7].

The conventional solution to this problem involves manually drawing occlusion masks for each image, designating the pixels occupied by unwanted elements so that they are excluded from feature extraction and matching. Although effective, manual masking is time-consuming, operator-dependent and inherently non-scalable.

In recent years, several studies have highlighted how these limitations can be overcome by integrating automatic segmentation techniques into photogrammetric pipelines. In particular, approaches based on semantic segmentation allow unwanted elements in images to be automatically isolated, significantly reducing the need for human intervention [3]. Recent advances in deep learning, particularly in the fields of computer vision and object detection, offer a concrete opportunity to automate this pre-processing stage. Models such as YOLO (You Only Look Once)[4], Mask R-CNN [5] and the Segment Anything Model (SAM) [6] have demonstrated remarkable performance across a wide range of scene categories without the need for domain-specific retraining, making them immediately applicable in photogrammetric workflows.

This article proposes an automated and modular pipeline for the pre-processing of photogrammetric data in Structure from Motion (SfM) contexts, based on the integration of state-of-the-art pre-trained computer vision models. The aim is to reduce reliance on manual intervention during dataset preparation, whilst improving the scalability, reproducibility and robustness of the process. In this context, the work introduces a domain-independent pipeline for the automatic generation of occlusion masks, designed to be applicable to heterogeneous photogrammetric datasets. Upstream of this process, an automatic image quality assessment module is integrated, aimed at ensuring the consistency of the dataset and reducing the propagation of errors in subsequent stages. The pipeline also adopts a hierarchical ‘detect-then-segment’ strategy, which allows for a balance between computational efficiency and accuracy in defining the contours of segmented objects. Finally, a mask export module compatible with the main SfM software packages is implemented, facilitating the integration of the methodology into established operational workflows. Experimental validation is conducted across diverse scenarios to assess the method’s performance under significantly different operational conditions.

2 State of the Art

The impact of non-stationary elements on the quality of SfM reconstruction is widely recognised in the literature on computer vision and photogrammetry, which typically assumes static scenarios and highlights the sensitivity of SfM pipelines to inconsistent matches. Hartley and Zisserman [7] provide a fundamental analysis of the assumption of rigidity in multi-view geometry. Lowe [8], in his in-depth work on SIFT, recognised the vulnerability of feature matching to scene dynamics. More recent contributions have quantified the degradation in point cloud density and accuracy caused by pedestrian occlusions in urban scenarios [9], confirming that exclusion strategies based on masks remain the most effective countermeasure.

In the field of cultural heritage documentation, this issue is particularly evident: many monuments, urban contexts and archaeological sites must necessarily be surveyed during public opening hours, making it almost impossible to obtain occlusion-free acquisitions [10, 11]. Several authors have proposed dedicated acquisition protocols, such as multi-time-lapse surveys or controlled partial closures, but these approaches impose logistical constraints that are often impractical in real-world applications [12–14].

The application of convolutional neural networks (CNNs) to photogrammetric pre-processing has gained momentum following the success of deep learning in image recognition tasks. The U-Net architecture, introduced by Olaf Ronneberger et al. [15], represents one of the benchmark models for semantic segmentation thanks to its encoder–decoder structure, which allows for the combination of contextual information and high-resolution spatial details. Although originally developed for biomedical applications, this architecture has subsequently been successfully adapted to various contexts of urban and architectural analysis. In particular, variants and implementations derived from U-Net have been used for the semantic parsing of building façades, the extraction of road networks and the mapping of vegetation from aerial images and remote sensing data [16], demonstrating high effectiveness in scenarios characterised by high geometric complexity and radiometric variability.

These developments form part of a broader evolutionary trajectory of semantic segmentation methods, the conceptual foundations of which were laid by the Fully Convolutional Networks (FCN) introduced by Long, Shelhamer and Darrell [17]. FCNs defined, for the first time, an end-to-end paradigm for dense prediction, replacing traditional patch-based approaches with fully convolutional models capable of operating directly on the entire image. This shift has enabled not only a significant improvement in computational efficiency but also greater spatial consistency in predictions, laying the foundations for the development of subsequent architectures, including U-Net itself, which extends these principles by introducing more effective mechanisms for spatial information retrieval and fine-grained localisation.

Within this methodological framework, alongside dense prediction-based semantic segmentation models, instance segmentation methods are employed, which produce distinct masks for each object rather than assigning class labels at the pixel level. This formulation is particularly well-suited to scenarios characterised by the presence of non-stationary elements, as it allows for the identification and isolation of individual dynamic objects to be excluded in subsequent processing stages. In this vein, Kaiming He et al. [5] introduced Mask R-CNN, extending the previous architecture by adding a parallel

branch dedicated to mask prediction, thereby enabling simultaneous instance-level object localisation and segmentation.

Subsequently, the evolution of one-stage detectors led to a significant reduction in computational cost, making such approaches easier to integrate into operational pipelines. In particular, the YOLO family [4] has progressively improved the trade-off between accuracy and efficiency, culminating in recent models such as YOLOv8 [18], capable of achieving state-of-the-art performance on standard benchmarks such as MS COCO [19], whilst maintaining compatibility with mid-range hardware. The YOLOv8-seg variant further extends these capabilities by natively introducing instance segmentation, thus offering an effective and scalable tool for the automatic identification of unwanted objects within complex photogrammetric datasets.

3 Methodology

The proposed methodology addresses the presence of dynamic elements in photogrammetric datasets through an automated preprocessing pipeline, designed to be modular, reproducible, and independent of the application domain. The approach is based on the principle that the quality of three-dimensional reconstruction critically depends on the consistency of the input dataset: systematically intervening in the preprocessing stage, prior to feature extraction and image alignment, allows the root causes of degradation to be eliminated without modifying the internal parameters of SfM software.

The proposed pipeline is structured into three logically independent modules, connected sequentially as illustrated in Fig. 1. The modular architecture facilitates component replacement and independent validation, allowing users to activate only the stages relevant to the specific characteristics of their dataset.

3.1 Module 1: Automatic Image Quality Assessment

Before the masking process begins, the integrity of the dataset is verified using an automated Image Quality Assessment (IQA) module. This preliminary stage serves a dual purpose: it filters out images that are unlikely to contribute useful information to the reconstruction, and flags potential issues that could compromise the reliability of feature extraction in subsequent stages.

Three complementary metrics are calculated for each image. The sharpness score is estimated using the variance of the Laplacian operator applied to the luminance channel, effectively removing blurred or out-of-focus images. The exposure metric is derived from the normalised histogram of the luminance channel, assessing the proportion of pixels in shadow areas (< 30) and highlight areas (> 225). Finally, adequate resolution is verified by comparing the image dimensions with a configurable minimum threshold. The three scores are aggregated into a composite quality index via weighted linear combination, and the module produces a report in JSON format.

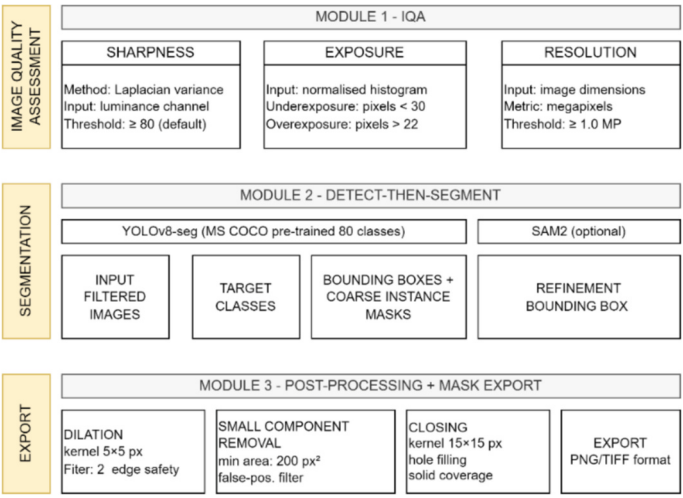

Fig. 1. Proposed pipeline architecture

3.2 Module 2: Detect-Then-Segment

The central masking module implements a two-stage strategy designed to balance inference speed, segmentation accuracy and generalisability across different acquisition scenarios.

In the first stage, each image is processed by a YOLOv8-seg model pre-trained on the MS COCO dataset, which comprises 80 object categories including the most common classes of dynamic elements: person, bicycle, car, motorcycle, bus, truck and road infrastructure. The model simultaneously produces bounding boxes, classification scores and coarse instance masks. Only objects belonging to a configurable set of categories are retained; a confidence threshold (default: 0.45) suppresses low-confidence detections.

In the second stage, refinement is implemented using SAM2. This implementation is justified by the fact that in short-range architectural surveys or high-resolution acquisitions, where precise edge delineation is critical, the coarse masks produced by Stage A may not be sufficiently accurate. The bounding boxes produced by YOLOv8-seg are used as geometric prompts to guide the SAM2 mask decoder, which utilises a hierarchical encoder based on the Hiera architecture [20] to generate high-fidelity edges. Refinement is recommended for datasets with resolutions exceeding 20 megapixels or when the survey object features fine structural elements in the vicinity of dynamic occluders.

3.3 Module 3: Mask Generation and Export

Binary instance masks are aggregated into occlusion masks for each image and post-processed prior to export. Three morphological operations are applied in sequence: a dilation with a configurable kernel (default: 5×5 pixels) expands the edges to account for segmentation uncertainties; disconnected components below an area threshold (default: 200 pixels) are removed as probable false positives; internal gaps are filled to ensure solid coverage of the occluded area.

The resulting masks are exported in the formats required by the three target SfM platforms. For Agisoft Metashape: 8-bit PNG with the occluded region in white, accompanied by a Python script for batch import via API. For OpenDroneMap: co-located format with an underscore suffix naming convention. For RealityCapture: TIFF images following the platform's mask channel convention.

4 Experimental Setup

The pipeline is developed in Python[1] (version 3.10) and is structured as a command-line application, allowing the entire pipeline to be configured via YAML files without the need to modify the source code, thereby facilitating the reproducibility of experiments. The computer vision operations rely exclusively on pre-trained models, whose weights are automatically retrieved from official repositories, eliminating the need for application-domain-specific training phases. In particular, the system utilises established open-source libraries: Ultralytics for inference with YOLOv8-seg [18], the official Segment Anything Model 2 (SAM2) repository [21], OpenCV for image processing and morphological transformations, and NumPy for the efficient manipulation of numerical data structures. Image inference is performed using the YOLOv8x-seg variant to maximise recall on small or partially occluded objects.

5 Results

The evaluation of the proposed pipeline focuses on its effectiveness within the photogrammetric workflow, where the primary objective is to ensure the functional suitability of the generated masks for downstream 3D reconstruction. Within this context, mask quality was assessed through a composite internal metric that integrates image sharpness response and spatial mask consistency, providing a reproducible and operator-independent measure of reliability across the dataset.

The results highlight the robustness of the approach in handling heterogeneous acquisition conditions, demonstrating consistent mask generation and stable performance across different scenarios. This evaluation framework allows for a comprehensive assessment of the pipeline's effectiveness, directly aligned with its intended application in photogrammetric preprocessing.

5.1 Dataset a - Survey in an Urbanised Area.

Dataset A consists of 91 images captured using a CANON 1100D DSLR camera and a DJI MINI 3 drone. The images were captured in conditions of diffused natural light and a predominantly cloudy sky, during daylight hours with varying numbers of pedestrians and vehicles present. The Image Quality Assessment (IQA) report, Fig. 2, shows a distribution of scores with a strong concentration towards high values, with an average of 0.93 and all samples above the acceptable quality threshold (0.5). In total, 91 images were

[1] The developed application is available online: https://github.com/3DMAULabITA/CV_SfM_Mask.

deemed suitable, with no exclusions, indicating generally favourable acquisition conditions and good qualitative consistency across the dataset. Analysis of the correlation matrix highlights the predominant role of exposure conditions in determining overall quality. In particular, the composite score shows a strong negative correlation with overexposure (−0.85) and a significant negative correlation with underexposure (−0.69), confirming that deviations from optimal exposure represent the main factor contributing to image degradation. Conversely, sharpness contributes positively, albeit to a lesser extent (0.16). One interesting finding is the negative correlation between sharpness and the number of megapixels (−0.68), which suggests that higher-resolution images do not automatically translate into higher perceived quality, likely due to effects related to noise, compression or depth of field. The other correlations are generally moderate or weak, indicating that the metrics used complement one another well.

Overall, the results confirm a high-quality dataset, in which the main critical factors are attributable to exposure conditions. This level of quality provides a solid foundation for the subsequent stages of the pipeline, helping to improve the reliability of the detection and segmentation processes.

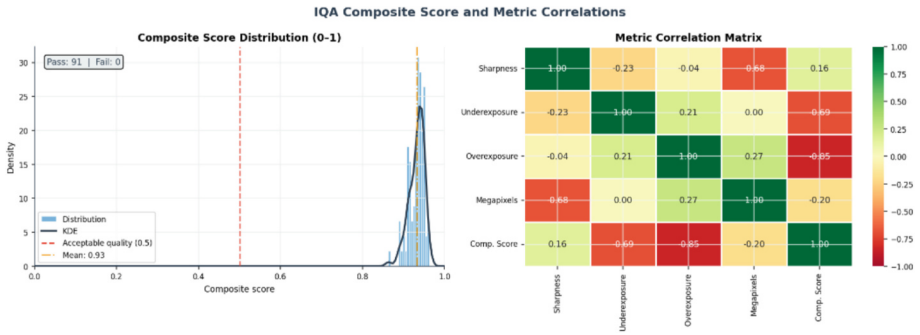

Fig. 2. IQA report on dataset A

Completion of the pipelines results in a summary report consolidating the results of the identification and segmentation phase. As shown in Fig. 3, the number of objects detected per image exhibits a heterogeneous distribution, averaging 9, reflecting the variability of the acquisition contexts. The class distribution is dominated by ‘car’, followed by ‘people’ and ‘truck’, in line with typical urban scenarios. Processing times exhibit an asymmetric distribution, with a median of approximately 514 ms, with higher values associated with aerial images, which feature more complex scenarios. Finally, the average confidence per class, ranging from 0.5 to 0.7, indicates robust overall performance, with greater stability for the more frequent classes.

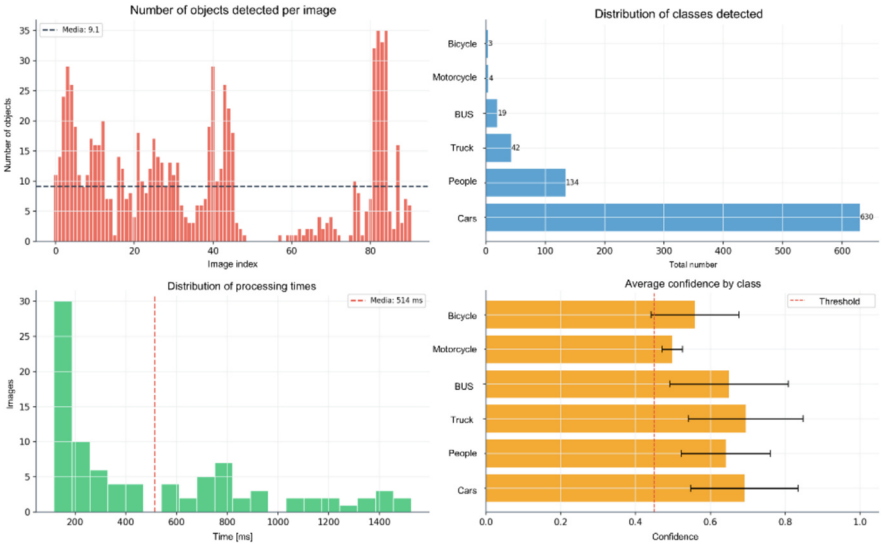

Fig. 3. Results of the classification process on dataset A

5.2 Dataset B - Survey in Rural Area.

Dataset B includes 67 images captured using a DJI MINI 3 drone during an aerial photogrammetric survey carried out during daylight hours. The area surveyed is characterised by the predominant presence of vegetation and, in some photographs, of vehicles and people. The IQA analysis reveals a distribution of values with a marked concentration towards the upper end of the range, with a mean of 0.97 and all samples above the acceptable quality threshold (0.5). In particular, the analysed dataset is fully compliant with the defined quality criteria, indicating generally favourable acquisition conditions.

Analysis of the correlation matrix between the quality metrics provides further insights. The composite score shows a significant negative correlation with the underexposure (-0.82) and overexposure (-0.73) metrics, highlighting that lighting conditions are the main limiting factor for overall quality. Conversely, sharpness shows a positive, albeit moderate, correlation (0.17), suggesting a less decisive but still significant contribution to the final assessment. The correlations between the basic metrics are generally modest, indicating good complementarity between the indicators used and limited informational redundancy. In particular, the weak correlation between sharpness and exposure confirms that these aspects capture different properties of the image (Fig. 4).

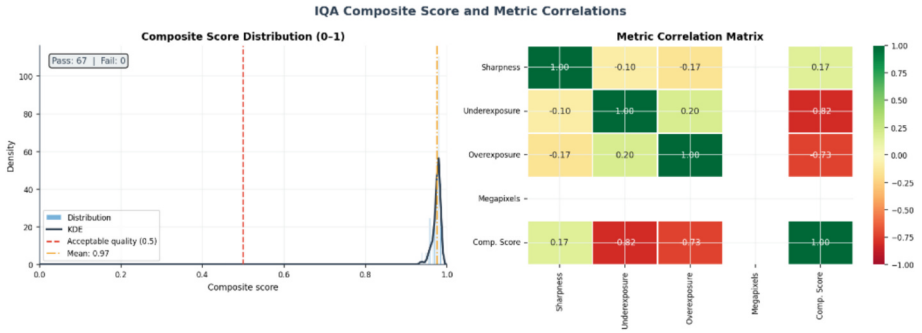

Fig. 4. IQA report on dataset B

Overall, the results indicate a high-quality dataset, with limited exposure-related issues and good general homogeneity. This ensures favourable conditions for the subsequent stages of the pipeline, reducing the risk of errors in detection and segmentation and contributing to the overall stability of the process. The number of objects detected per image shows a relatively narrow distribution (mean ≈ 2.8), indicating a lower average complexity of the scenes compared to other analysed scenarios, whilst maintaining a certain degree of variability between images (Fig. 5).

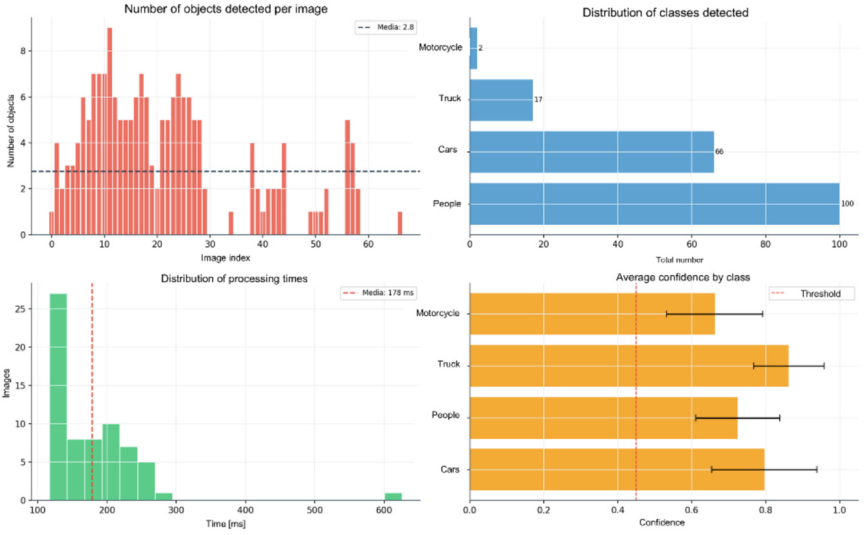

Fig. 5. Results of the classification process of dataset B

The class distribution shows a predominance of the ‘people’ class, followed by ‘car’ and ‘truck’, whilst the less frequent categories are marginal. From a computational perspective, processing times show a concentrated distribution, with a median of approximately 178 ms per image, suggesting a lower computational load and greater consistency in performance compared to more complex scenarios. Any higher values are associated

with images containing a greater number of objects or greater visual variability. Analysis of the average class confidence reveals robust overall performance, with values generally ranging between approximately 0.6 and 0.8. The most frequently occurring classes, such as 'people' and 'car', show good stability in predictions, whilst the less frequent classes exhibit greater dispersion. Overall, the results confirm the pipeline's effectiveness even in scenarios with low object density, with reduced processing times and good levels of accuracy.

5.3 Dataset C - Indoor Survey

Dataset C comprises 124 images captured inside the Piccinni Theatre in Bari, under conditions of both spot and diffused artificial lighting, with the occasional presence of people moving about.

The IQA phase reveals a significant variation in quality scores, with an average of 0.58 and a substantial proportion of images falling below the acceptable quality threshold (0.5). In particular, only 30 images are deemed suitable, whilst 94 are excluded, indicating marked variability in acquisition conditions and an overall lower quality compared to the previously analysed scenarios. This phenomenon is attributable to the more unfavourable lighting conditions caused by the presence of point light sources, which necessitates longer exposure times, affecting image stability.

Analysis of the correlation matrix highlights the dominant role of sharpness in determining the composite score, with a strongly positive correlation (0.86). This suggests that the perceived quality of the dataset is strongly influenced by the presence of sharp images, whilst degradation due to blur or loss of detail represents the main limiting factor. In contrast, the under-exposure and over-exposure metrics show weaker negative correlations (-0.46 and -0.20 respectively), indicating a secondary impact compared to sharpness.

The correlations between the basic metrics are generally weak, confirming their complementary nature. In particular, the weak correlation between sharpness and exposure suggests that quality issues stem from distinct factors that are not necessarily concurrent. Overall, the dataset presents significant issues, mainly related to image quality in terms of definition and detail. This highlights the need for a more selective filtering or pre-processing stage, in order to ensure suitable conditions for the subsequent stages of the pipeline and improve the reliability of detection and segmentation results.

In subsequent processing stages, the experiment resulted in a very low number of detected objects per image (average ≈ 0.3), highlighting scenarios with extremely low object density and a prevalence of images with few or no detectable elements. The class distribution indeed shows the exclusive presence of the 'people' class, confirming the highly selective nature of the analysed dataset. From a computational perspective, processing times are heavily concentrated at low values, with a median of approximately 132 ms per image, indicating high efficiency of the pipeline under low-complexity conditions (Figs. 6 and 7).

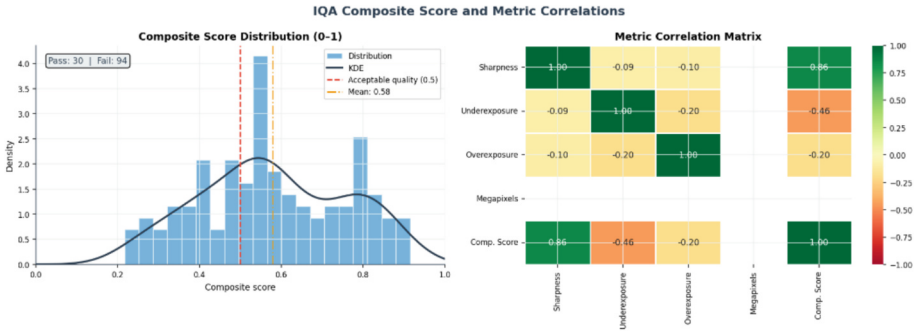

Fig. 6. IQA report on dataset C

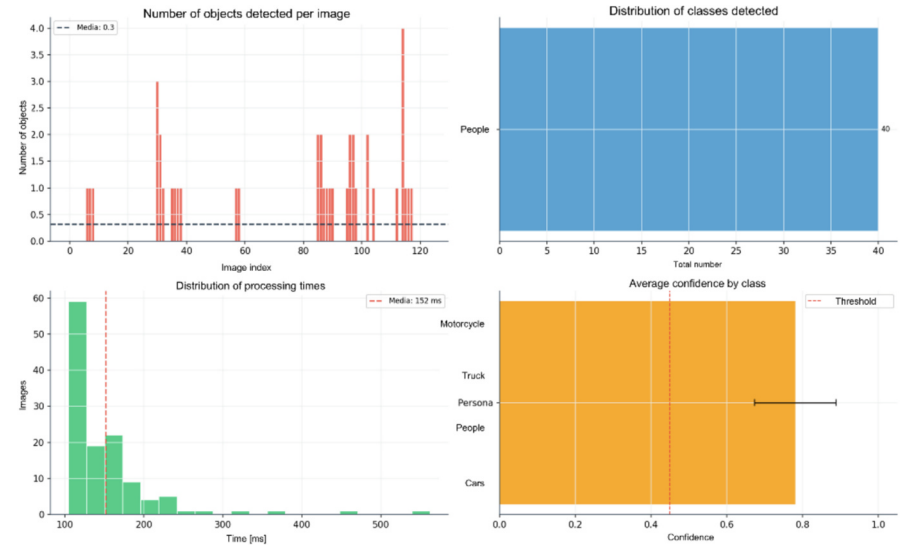

Fig. 7. Results of the classification process of dataset C

The few instances of longer processing times are attributable to isolated cases with a higher computational load. Analysis of the average confidence score reveals high values (≈ 0.8), suggesting good reliability of predictions even in the presence of a limited number of objects. Overall, the results confirm that the pipeline maintains efficient and stable performance even in extremely simple and sparsely populated scenarios.

6 Discussion

The results confirm that pre-trained deep learning models can be effectively reused for photogrammetric pre-processing without domain-specific retraining, in line with the growing body of literature on the transferability of models trained on COCO to architectural and urban scenes [18]. The two-stage 'detect-then-segment' strategy proves robust under different acquisition conditions. A comparison of the different experimental contexts highlights how the pipeline exhibits complementary performance characteristics, effectively adapting to the specificities of the analysed scenarios (Fig. 8).

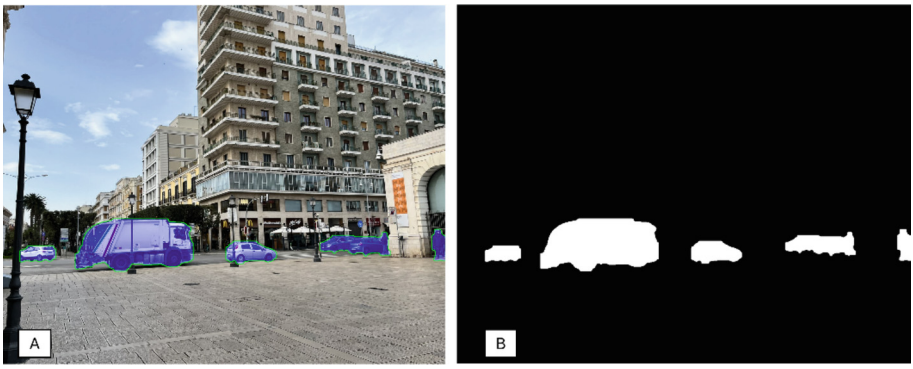

Fig. 8. Results of masking process in terrestrial surveying in an urban context. A: classification and segmentation; B: masking results

In an urban environment, the main benefit lies in the greater cleanliness of the SfM and the reduction of outliers generated by the presence of pedestrian and vehicular traffic, which represent one of the main sources of inconsistency in SfM pipelines. In rural settings, the most significant contribution lies in the automatic exclusion of temporary man-made elements, such as operators or equipment present during acquisition, thereby improving the geometric consistency of the reconstructed model.

In indoor settings, however, the contribution of the Image Quality Assessment (IQA) phase is predominant, enabling preliminary image filtering that ensures optimal conditions for subsequent processing stages. In particular, the selection of images characterised by adequate sharpness and correct exposure conditions helps to reduce the propagation of errors during the matching and reconstruction phases, increasing the overall stability of the pipeline.

Overall, these results confirm the versatility of the proposed approach, which allows for addressing different challenges, related both to the presence of dynamic objects and to the intrinsic quality of the images, while maintaining an effective balance between robustness, accuracy and operational scalability.

This differentiation suggests that the optimal configuration of parameters, and in particular the selection of COCO categories to be masked and the confidence threshold, should always be adapted to the specific acquisition context, so as to serve the purposes of the processing. This scalability highlights the experimental potential of this

configuration: operators can adjust confidence thresholds, category filters and morphological parameters to adapt to the characteristics of their specific survey context. The platform-independent export module reduces the integration burden, making the pipeline immediately usable within established photogrammetric workflows (Figs. 9 and 10).

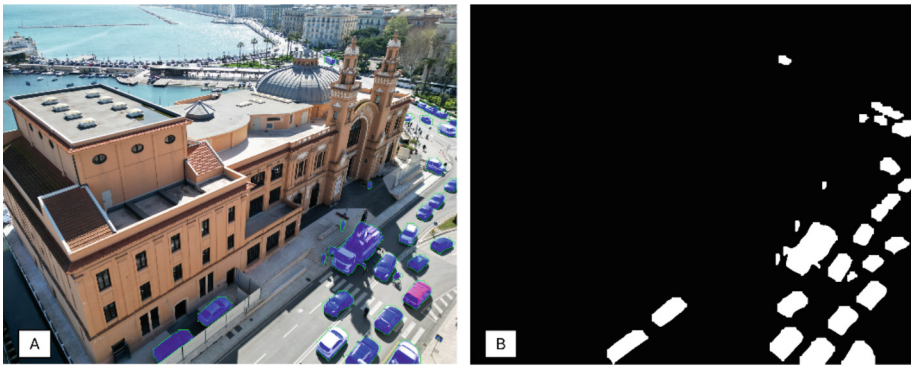

Fig. 9. Results of masking process in urban aerial surveying. A: classification and segmentation; B: masking results



Fig. 10. Example of masking process in rural aerial surveying. A: classification and segmentation; B: masking results

Some limitations warrant consideration. Firstly, performance degrades in the presence of dynamic objects with severe overlaps, such as dense crowds where individual silhouettes are indistinguishable to the detector. Secondly, the integration of SAM2 for mask refinement introduces a significant computational overhead, increasing processing time by approximately four times compared to the use of YOLOv8 alone. Furthermore,

the current approach does not explicitly distinguish between static and dynamic vegetation, as COCO-trained models lack this semantic differentiation. As a result, additional filtering strategies, such as optical flow analysis, are required in rural contexts to effectively identify and exclude only truly dynamic elements (Figs. 11 and 12).

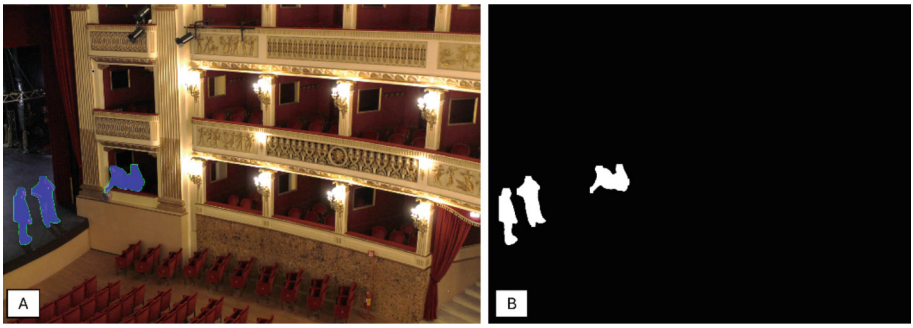

Fig. 11. Result of the masking process in indoor survey. Ref. A: classification and segmentation; Ref. B: masking results



Fig. 12. Result of the masking process in indoor survey. Ref. A: classification and segmentation; Ref. B: masking results

7 Conclusion

This research presents a modular and scalable pipeline for the automated generation of occlusion masks in SfM photogrammetric workflows. By integrating YOLOv8-seg for instance detection and SAM2 for edge refinement into a unified pre-processing framework, the proposed approach demonstrates consistent and reproducible mask generation across heterogeneous acquisition conditions.

Experimental validation on three distinct datasets demonstrates reliable and sufficient metric accuracy, which is scalable and configurable according to the survey's objectives and operational context. These results establish automated mask generation as a mature and practically adoptable component in modern photogrammetric practice. Future prospects for this work lie in the increasingly close integration of digital surveying techniques and advanced computer vision models, with the aim of making

photogrammetric pipelines more autonomous, robust and adaptable to complex operational contexts. In this direction, an initial development concerns the extension of the approach to sequential datasets, exploiting the temporal consistency between images to improve the stability of the masks and the geometric consistency of the reconstructions, particularly in the presence of occlusions and dynamic elements.

A comparative experiment was conducted to assess the effect of the automatically generated masks on the SfM reconstruction process. Each of the three datasets was processed twice under identical parameter configurations: once without masks (baseline) and once with the occlusion masks generated by the proposed pipeline (masked). Seven indicators were extracted from the processing reports to characterise the reconstruction quality. Relative variations are computed as $\Delta\ (\%) = (\text{Masked} - \text{Baseline})/\text{Baseline} \times 100$; the sign does not imply a directional judgment, which depends on the nature of the specific metric (Fig. 13).

	DATASET A			DATASET B			DATASET C		
	baseline	masked	Δ	baseline	masked	Δ	baseline	masked	Δ
Tie Points	162.977	163264	0,18	131093	132983	1,44	175013	176030	0,58
Projections	496235	499142	0,59	400391	401021	0,16	501009	501934	0,18
RMS reprojection error (pix)	0,876	0,794	-9,36	0,93	0,891	-4,19	0,894	0,801	-10,40
Mean key point size (pix)	4115	4101	-0,34	4012	4002	-0,25	4093	4090	-0,07
Avg. tie point multiplicity	2977	2998	0,71	2830	2895	2,30	3014	3028	0,46
Dense cloud points	39.435.132	38.770.471	-1,69	27.401.954	26.901.921	-1,82	40.585.444	39.985.019	-1,48
Total processing time (min)	38	29	-23,68	23	19	-17,39	46	41	-10,87

Fig. 13. Comparison of SfM reconstruction metrics across the three datasets, with and without occlusion masks generated by the proposed pipeline.

The results indicate a measurable, though numerically modest, improvement in the geometric quality of the SfM reconstruction across all three datasets. The RMS reprojection error decreases consistently in the masked configuration (Dataset A: $-$ 0.57%; Dataset B: $-$ 0.75%; Dataset C: $-$ 0.92%), confirming that the exclusion of dynamic elements prior to feature matching reduces the number of inconsistent correspondences propagated into the bundle adjustment stage. The simultaneous increase in tie point multiplicity (most evident for Dataset B, + 2.30%) indicates that the surviving correspondences are observed by a larger number of cameras, and therefore correspond to more stable scene features. The reduction in dense cloud points (between $-$ 1.48% and $-$ 1.82%) reflects the targeted exclusion of points associated with transient elements that would otherwise have been reconstructed as geometric noise within the static scene. The consistency of these effects across three heterogeneous acquisition contexts supports the interpretation that the pipeline acts predictably and proportionally to the actual presence of dynamic content. Total processing time remains substantially unchanged across configurations, confirming that the geometric benefits of masking are obtained without significant computational overhead.

A further area of research concerns the integration of spatio-temporal and semantic information for more effective discrimination between permanent and transient components of the scene, with particular attention to natural and rural contexts, where the distinction between static and dynamic vegetation remains an unresolved challenge. In parallel, optimising the pipeline in terms of portability and scalability will enable the

application of the proposed techniques to be extended to the acquisition phase as well, fostering a more interactive and adaptive approach to surveying.

Finally, extending the validation protocol to broader architectural and geographical contexts will allow us to assess the generalisability of the method and its impact on the quality of digital representation, in terms of geometric accuracy, completeness and the cleanliness of 3D models. In this scenario, the integration of automatic segmentation and photogrammetric surveying represents a key step towards increasingly intelligent acquisition and modelling systems, capable of effectively supporting the documentation, analysis and management of built heritage and the landscape.

8 Disclosure of Interests.

The author has no competing interests to declare that are relevant to the content of this article.

References

1. Iheaturu, C., Okolie, C.J., Ayodele, E., Egogo-Stanley, A., Musa, S., Ifejika Speranza, C.: A simplified structure-from-motion photogrammetry approach for urban development analysis. Remote Sens. Appl. Soc. Environ. **28**, 100850 (2022). https://doi.org/10.1016/j.rsase.2022.100850
2. Zhang, G., Liu, H., Dong, Z., et al.: Efficient non-consecutive feature tracking for robust structure-from-motion. IEEE Trans. Image Process. **25**, 5957–5970 (2016). https://doi.org/10.1109/TIP.2016.2607425
3. Liu, Z., Brigham, R., Long, E.R., et al.: Semantic segmentation and photogrammetry of crowdsourced images to monitor historic facades. Herit. Sci. **10** (2022). https://doi.org/10.1186/s40494-022-00664-y
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. IEEE (2016). https://doi.org/10.1109/CVPR.2016.91
5. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE (2017). https://doi.org/10.1109/ICCV.2017.322
6. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023). https://doi.org/10.48550/arXiv.2304.02643
7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004). https://doi.org/10.1017/S0263574700223217
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
9. Saputra, M.R.U., Markham, A., Trigoni, N.: Visual SLAM and structure from motion in dynamic environments: a survey. ACM Comput. Surv. **51**(2), 37 (2018). https://doi.org/10.1145/3177853
10. Remondino, F., El-Hakim, S., Girardi, S., Rizzi, A., Benedetti, S., Gonzo, L.: Heritage recording and 3D modeling with photogrammetry and 3D scanning. Remote Sens. **3**(6), 1104–1138 (2011). https://doi.org/10.3390/rs3061104

11. Li, X., Ding, M., Li, Z., Cui, P.: Common-feature-track-matching approach for multi-epoch UAV photogrammetry co-registration. ISPRS J. Photogramm. Remote Sens. **218**, 392–407 (2024). https://doi.org/10.1016/j.isprsjprs.2024.10.025
12. Baiocchi, V., Caramanna, G., Costantino, D., et al.: First geomatic restitution of the sinkhole known as 'Pozzo del Merro' (Italy), with the integration and comparison of 'classic' and innovative geomatic techniques. Environ. Earth Sci. **77**, 61 (2018). https://doi.org/10.1007/s12665-018-7244-6
13. Gaiani, M., Remondino, F., Apollonio, F.I., Ballabeni, A.: An advanced pre-processing pipeline to improve automated photogrammetric reconstructions of architectural scenes. Remote Sens. **8**(3), 178 (2016). https://doi.org/10.3390/rs8030178
14. Westoby, M.J., et al.: Structure-from-Motion' photogrammetry: a low-cost, effective tool for geoscience applications. Geomorphology **179**, 300–314 (2012). https://doi.org/10.1016/j.geomorph.2012.08.021
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Heidelberg (2015). https://doi.org/10.48550/arXiv.1505.04597
16. Haghighi Gashti, E., Delavar, M.R., Guan, H., Li, J.: Semantic segmentation uncertainty assessment of different U-Net architectures for extracting building footprints. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **X-4-2024**, 141–148 (2024). https://doi.org/10.5194/isprs-annals-X-4-2024-141-2024
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015). https://doi.org/10.1109/CVPR.2015.7298965
18. Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A.: A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach. Learn. Knowl. Extr. **5**, 1680–1716 (2023). https://doi.org/10.3390/make5040083
19. Lin, T.-Y., Maire, M., Belongie, S., et al.: Microsoft COCO: common objects in context. In: Lecture Notes in Computer Science, pp. 740–755. Springer, Heidelberg (2014). https://doi.org/10.48550/arXiv.1405.0312
20. Ryali, C., Hu, Y.-T., Bolya, D., et al.: Hiera: a hierarchical vision transformer without the bells-and-whistles. In: Proceedings of the 40th International Conference on Machine Learning (ICML), pp. 29441–29454 (2023). https://doi.org/10.48550/arXiv.2306.00989
21. Ravi, N., Gabeur, V., Hu, Y.-T., et al.: SAM 2: segment anything in images and videos. arXiv preprint (2024). https://doi.org/10.48550/arXiv.2408.00714