Towards autonomous shotcrete construction: semantic 3D reconstruction for concrete deposition using stereo vision and deep learning

Patrick Schmidt[1], Dimitrios Katsatos[2], Dimitrios Alexiou[2], Ioannis Kostavelis[2], Dimitrios Giakoumis[2], Dimitrios Tzovaras[2], Lazaros Nalpantidis[1]

[1]Technical University of Denmark, Kongens Lyngby, Denmark
[2]Centre for Research and Technology Hellas, Thermi, Greece

pasch@dtu.dk, {dkatsatos, dalexiou, gkostave, dgiakoum, dimitrios.tzovaras}@iti.gr, lanalpa@dtu.dk

Abstract -

The adoption of autonomous systems is a foreseeable necessity in the construction sector due to work hazards and labor shortages. This paper presents a semantic 3D understanding module that creates 3D models of construction sites with highlighted regions of interest for shotcrete application. The approach uses YOLOv8m-seg and SiamMask for robust semantic segmentation together with RTAB-Map and InfiniTAM for visual odometry and 3D reconstruction. Our method is the first step towards a novel, autonomous robot for shotcrete spraying and finishing. The effectiveness of our approach is shown on a mock-up construction site and provides evidence for the applicability of robotic construction.

Keywords -

Construction Robotics, 3D Reconstruction, Semantic Segmentation, Shotcrete Automation

1 Introduction

The construction sector is among the industries that have not undergone a major digital transformation. Within the European Union (EU), it stands out as one of the least digitalized industries [1]. Furthermore, construction work is skilled labor, posing demands on the labor market that are not met today: A European Labor Authority report shows that 13 out of 30 regions are reporting shortages of Concrete Placers and Finishers [2]. Construction work is also known to be highly hazardous with Eurostat reporting the highest incidence of non-fatal work-related accidents in the EU, reaching almost 3000 accidents per 100,000 employed persons [3].

Aiming to alleviate these challenges and enhance overall efficiency, the construction industry is gradually embracing digital advancements [4]. In particular, the application of semantic 3D reconstruction, through computer vision systems and deep learning algorithms, plays a critical role in enabling precise 3D digital models of the scene. This advancement allows for the creation of rich Building and Construction Information Models (BIM/CIM) and Mechanical, Electrical, and Plumbing (MEP) systems [5], targeting the development of digital models of the construction site to ensure time-sensitive decision-making and streamline project progress monitoring.

Additionally, real-time 3D reconstruction featuring semantic annotations can be used to measure the thickness level of ground support walls. This is especially valuable in shotcrete tasks, where dry concrete is sprayed onto the wall surface. Shotcrete processes subject workers to health hazards due to concrete rebound and rely heavily on the expertise of nozzlepeople to determine the amount of concrete necessary for effective filling [6]. Consequently, this method often leads to a substantial waste of concrete. Enabling advanced imaging systems to monitor and control the procedure can significantly improve accuracy and worker safety and reduce excess use of material [7].

This paper presents a robust real-time method for semantic 3D reconstruction tailored for robotic construction applications. It introduces a novel computer vision system for the detection and reconstruction of shotcrete construction sites, using YOLOv8m-seg for semantic segmentation and InfiniTAM for 3D reconstruction.

The main contributions of this paper are summarized as follows:

- Compilation of a new semantic segmentation dataset and training of a robust segmentation model.
- Introducing an integration strategy between 3D reconstruction and semantic segmentation, adapted for robotic applications within construction sites.

The paper is organized as follows: In Section 2, the latest advancements in semantic segmentation, 3D reconstruction, and computer vision datasets tailored to construction sites are outlined. Section 3 provides a comprehensive analysis of each proposed pipeline component and the interaction between them. Section 4 demonstrates the experimental assessment, and Section 5 concludes this paper.

2 Related work

2.1 Semantic segmentation

Semantic segmentation is one of the key components necessary for a robotic system to analyze its surroundings. With the advent of deep learning, a boost in performance gave the computer vision community momentum to research new, deep-learning-based, methods for semantic segmentation. These methods allow for parameter-less inference methods and do not rely on extensive domain knowledge. U-Net [8] is a popular, one of the first deep-learning-based semantic segmentation models [9] with an application in biomedicine. DeepLabV3 [10] is another popular model for semantic segmentation. It made its debut in 2017 and is the result of incremental developments based on the first [11] and second [12] versions of DeepLab. The use of atrous convolutions and atrous spatial pyramid pooling makes DeepLab have a larger receptive field and thus higher-resolution feature maps, retaining more information in deeper layers. This enables the integration of both local and global contexts when extracting features. HRNet [13] was released in 2020 and also addresses the issue of retrieving and maintaining high-resolution representations from the input image. The main contributions of HRNet are the so-called parallel multi-resolution convolutions and repeated multi-resolution fusions.

YOLO (You Only Look Once) [14] is a popular series of Convlutional Neural Network (CNN)-based object detection models, first released in 2016 and consequently developed up to this date in various versions [15, 16, 17, 18]. Its latest version, YOLOv8, provides a unified framework for image classification, object detection, instance segmentation, object tracking, and pose estimation. Architectural details concerning this model can be found in [19]. It provides an easy-to-use interface for training the various tasks and integrates well with experiment tracking tools, ensuring quality control.

Other recent advancements like Segment Anything [20] or BEiT-3 [21] are departing from CNN-based architectures and use Transformer-based models trained on large-scale datasets, providing foundation models capable of tackling multiple downstream tasks.

In our work, we combine YOLOv8m-seg with a mask tracking model to stabilize predictions obtained from a low-data model.

2.2 3D reconstruction

Recently there has been a notable effort to integrate deep learning techniques into 3D reconstruction approaches. Specifically, this effort focuses on Neural Radiance Fields (NeRF) [22], and its variants such as Instant-NGP NeRF [23]. These methods are pivotal elements that boosted 3D reconstruction. Typically, NeRF-based methods use a regression technique for opacity and color together with numerical integration, predicting the real step of the volumetric rendering function based on images with known camera poses. Robotic applications of 3D reconstruction in the construction industry include Inspection-NeRF [24] as a method for capturing surface defects in the form of RGB-D images, collected by a wall-climbing robot to create a 3D model and its bounding box, aligning it with the NeRF implicit boundary. In the work of [25], semantic segmentation is applied to a 3D model generated by Instant-NGP to construct BIM models from a sequence of construction site images. Additionally, conventional methods such as Clustering Multi-View Stereo (CMVS) and Patch-based Multi-View Stereo (PMVS) [26], are deployed for dense 3D reconstruction in construction sites.

Notwithstanding the advancements in NeRFs, these techniques require absolute scale recovery and entail significant computational time requirements, especially when handling large-scale scenes such as construction sites [27]. More specifically, 3D reconstruction pipelines that integrate camera pose estimation methods and dense meshing frameworks have demonstrated their superior feasibility for 3D reconstruction in time-sensitive construction projects [28]. KinectFusion [29] paved the way for approaches using real-time volumetric Truncated Signed Distance Fields (TSDF), resulting in InfiniTAM [30]. This method uses RGB-D input to perform real-time reconstruction. To accomplish that, it enables a module for estimating camera poses with a keyframe-based relocalization system and provides globally consistent reconstruction, using either TSDF or surfel methods. Additionally, most recent RGB-D or stereo approaches are using real-time TSDFs from Euclidean Signed Distance Fields (ESDFs) to formulate implicit surfaces [31].

In this study, InfiniTAM is investigated as a real-time modular method and its integration with the well-established visual SLAM algorithm, namely RTAB-MAP [32], is explored to achieve more accurate results.

2.3 Computer Vision datasets in construction

Examples of datasets focusing on computer vision in construction include the Alberta Construction Image Dataset [33], datasets for safety helmet detection [34, 35], and the SODA dataset [36], designed for general-purpose object detection in construction environments. For specific construction activities, datasets such as CODEBRIM [37] address concrete inspection, while others like [38] focus on window installation. Regardless, publicly available datasets captured in construction environments, specifically tailored for robotic application in construction areas, remain a scarce resource. Consequently, we compiled custom datasets for our application.

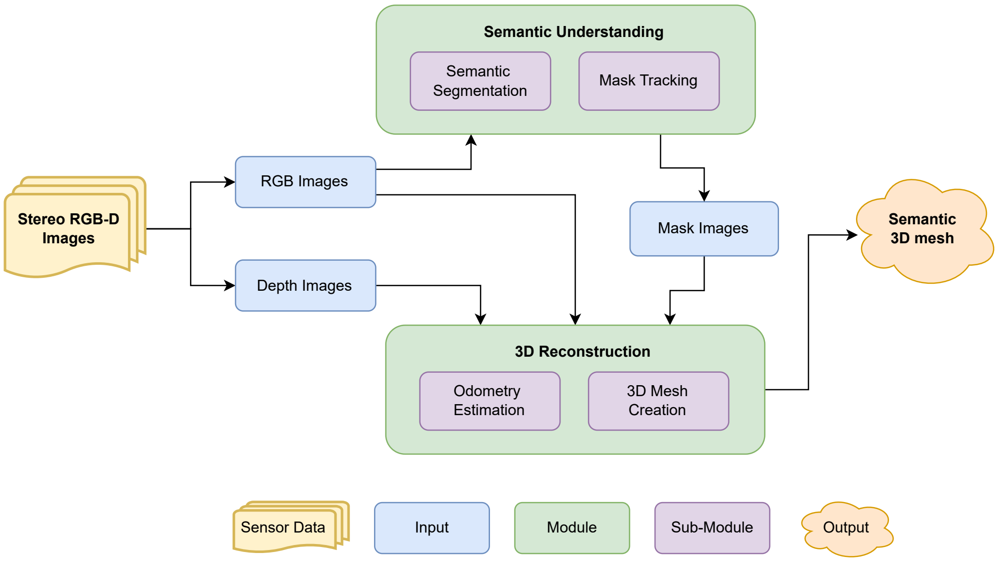

Figure 1. System overview of the proposed and integrated methods

3 Methodology

The structure of our method is illustrated in Figure 1. It is composed of several modules with their respective submodules which are covered in detail in the following.

3.1 Semantic understanding

The proposed end-to-end deep-learning-based pipeline consisting of YOLOv8 and SiamMask needs to be trained to yield sensible results. YOLOv8 and SiamMask were selected due to their state-of-the-art performance. Furthermore, YOLOv8 has a well-known architecture, ensuring reliable and interpretable results. We use an unmodified, pre-trained SiamMask model. Thus, the training efforts concern YOLOv8m-seg. The following section describes experiments carried out to obtain a well-performing semantic segmentation module.

3.1.1 Training setup

The training goal is to teach the model how to segment areas where shotcrete needs to be applied. We stipulate that such areas are easily identified by exposed reinforcement bars. The model is therefore trained to detect exposed reinforcement bars. We compiled a small dataset for this task: detecting exposed reinforcement bars in wooden structures, without a significant presence of other objects. The dataset restricts the appearance of areas in which shotcrete is applied, to rectangular areas with a rather uniform, wooden background and good lighting conditions. In real-life conditions, these areas are less regular in shape and have a bigger variety of backgrounds and lighting conditions. The dataset consists of three splits:

- Training: 515 frames, 580 instances
- Validation: 191 frames, 167 instances
- Testing: 210 frames, 210 instances

We train the YOLOv8m-seg model for 100 epochs with default hyperparameters recommended by [19].

3.1.2 Pre-training procedure

We perform a custom pre-training schedule as follows:

1. Train the YOLOv8m bounding box detection model from COCO pre-trained weights provided by [19] on the COCO dataset with augmented, synthetic reinforcement bars [39], for 100 epochs. All COCO classes are used, plus an “ExposedBars” category.
2. Considering the epochs from step 1., we use the weights that achieved the best bounding box mean Average Precision (50-95) (mAP50-95) as a starting point and train on the CODEBRIM dataset [37] with CODEBRIM classes, as well as on an augmented version of the dataset with synthetic reinforcement bars.
3. Use weights from the last epoch of step 2. as the custom pre-trained weights to start training the segmentation model on the dataset described in 3.1.1.

Figure 2 shows the values of the segmentation loss (both evaluated on the training and validation dataset split) and the development of mask precision and recall (evaluated on the validation dataset split) over training epochs. The losses and metrics show no sign of overfitting, i.e., the validation loss increasing while the training loss is decreasing.

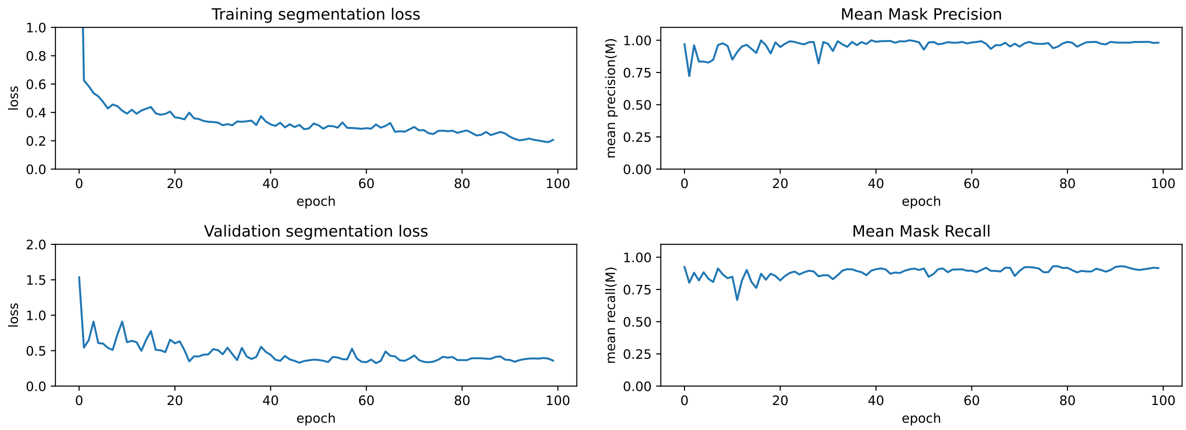

Figure 2. Training and validation losses per epoch (left) and Mean Mask Precision and Mean Mask Recall per epoch (right)

3.1.3 Robustness enhancement

Perturbations of the input image sequences, e.g., rotations and distortions, which are often encountered from cameras mounted on mobile robots can lead to cases where the segmentation model cannot re-detect previously seen areas with exposed reinforcement bars. To counteract this, an object-tracking model is deployed. The semantic segmentation model in combination with the tracking model constitutes the complete semantic understanding module. The tracking model used for this task is SiamMask [40], using the default pre-trained model on the DAVIS dataset [41] as well as the default configuration provided by the developers.

When the segmentation model detects an object, the tracker is initialized with the respective image and bounding boxes. They are stored in a first in, first out (FIFO) queue of images and bounding boxes. When the segmentation model fails, the tracker pops a bounding box and the corresponding image from the queue to find the content of the bounding box in the current image.

We evaluate the effect of the tracking model on the semantic understanding module. The module is tested on the validation and testing sequence of the dataset, both with disabled and enabled tracking. We use a FIFO queue length of 1 and default confidence thresholds for both the segmentation model and the tracker model. Figure 3 shows plots of the mean Intersection over Union (mIOU) calculated on all available IOUs up to that frame. Both plots indicate the superior performance of the semantic understanding module when the tracking model is used as an additional layer to recover false negatives. The benefit is mainly pronounced on the test dataset split, as the segmentation model fails and thus activates the tracking submodule more often. A video showcasing the benefit of the module can be found at [42]. In this video, frames with red masks show the output of a tracked mask when the segmentation module failed to produce an output.

In conclusion, the SiamMask tracking model can increase robustness in the case of the segmentation model failing on objects/areas it has detected once before.

3.1.4 Hyperparameter tuning

This section evaluates the influence of tracker and segmentation model hyperparameters on the overall performance measured in mean precision, mean recall, and mIOU. We perform a grid search on these parameters, to maximize mIOU on the test set:

1. Tracker FIFO queue length (QL), search range: $\{1, 2, ..., 11\}$
2. Tracker Confidence Threshold (TCT), search range: $\{0.1k | k \in \{1, 2, ..., 9\}\}$
3. Segmentation Confidence Threshold (SCT), search range: $\{0.1k | k \in \{1, 2, ..., 9\}\}$

Contrary to common practice, the sweep is conducted on the test split. This is due to the minimal effect of the tracker on the validation split.

We found that the 10 best-performing runs exclusively use a QL of 2 together with medium to high TCTs (0.6 – 0.8) and medium to low SCTs (0.1 – 0.4). The 10 worst-performing runs tend to use medium to high QLs (6-10), medium to high TCTs (0.6-0.9), and a high SCT (0.9). This suggests the usage of a QL of 2 together with a low SCT and a high TCT. We have chosen QL=2, SCT=0.4, TCT=0.9, achieving a mIOU of 0.85 together with mean Precision (mP)=0.96 and mean Recall (mR)=0.88 on the test split.

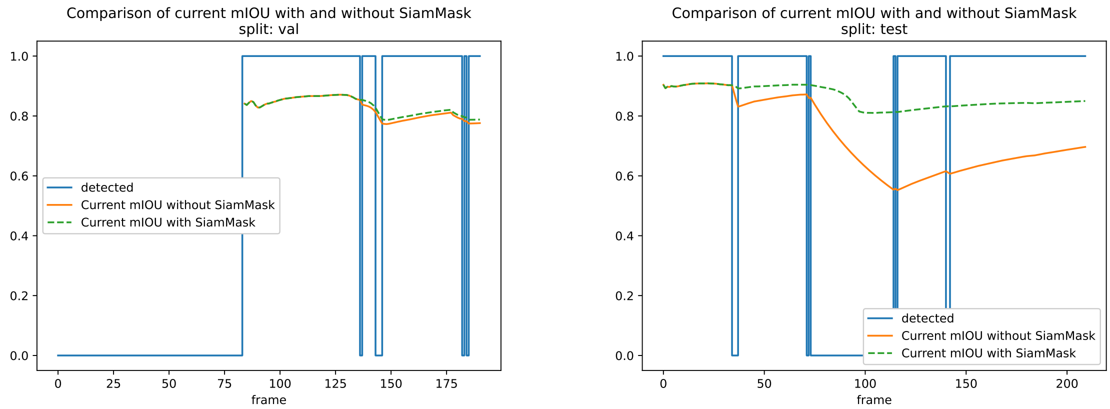

Figure 3. Temporal development of detections and mIOU, on the validation (left) and test (right) splits. Green dashed line shows the current mIOU with tracking enabled, whereas the orange line indicates the scenario with tracking disabled.

3.2 3D Reconstruction

The 3D reconstruction process uses RGB-D input, obtained through a stereo-vision camera to infer a real-time 3D mesh. The 3D reconstruction module comprises the subsequent submodules:

3.2.1 3D mesh creation

Aiming to develop an end-to-end modular pipeline, we strategically applied the InfiniTAM algorithm. Its key advantage over recent approaches is the capability to synthesize accurate 3D surfaces in real-time, addressing absolute scale recovery and long computational time issues common in other methods, as reported in [27], which is crucial for dynamic and expansive scenes such as construction sites. Furthermore, it seamlessly integrates with stereo sensors, typically used for robotic vision applications.

In particular, InfiniTAM infers TSDF-based volumetric reconstruction, using hash tables. It relies on RGB-D input to segment the scene into rigid subscenes and refines their relative poses to build a coherent overall map. During the camera pose prediction, it adjusts the current camera position to track the sensor frame, aligning surface measurements with the model estimation [30].

It is worth noting that our approach strongly depends on the performance of the stereo camera, as conventional RGB-D cameras may face limitations in composing depth images in semi-indoor scenes with changing illumination conditions. To fulfill these objectives, the Roboception RC-Visard 160 stereo camera was deployed.

3.2.2 Odometry estimation

During our experiments, we observed poor relocalization performance of vanilla InfiniTAM, when dealing with partial and noisy surface measurements, captured under conditions of reduced overlap. To address this challenge, we conducted a thorough investigation of the camera pose estimation. Leveraging InfiniTAM's modularity, we looked into the integration of a more robust camera pose estimation module. For this, we deployed the well-established visual odometry method RTAB-Map [32], which is a flexible Graph-Based SLAM technique, to create dense 3D reconstructions. Consequently, the integration of an advanced external camera pose estimation played a significant role in developing a robust real-time 3D reconstruction method, demonstrating adaptability in handling complex environments such as construction sites.

3.3 Integrated system

The proposed system consists of two interconnected major submodules: semantic understanding and 3D reconstruction. The semantic understanding module deploys a YOLOv8m-seg [19] segmentation model, stabilized with the SiamMask [40] segmentation mask tracker, adept at robustly identifying exposed reinforcement bars. The 3D reconstruction module incorporates an external odometry estimation component and employs TSDF volumetric reconstruction, leveraging RTAB-Map [32] and InfiniTAM [30], respectively.

As illustrated in Figure 1, the integrated method follows these steps: Given the sensor data (RGB image, depth image) as an input, the semantic understanding module infers a mask, marking areas with exposed reinforcement bars in each image. Simultaneously, the sensor data is used to obtain a visual odometry estimation. Combined with the segmentation mask and sensor depth measurements, the 3D reconstruction module constructs a semantic 3D model of the environment. This model can be used to acquire a digital representation of the construction site in real-time, streamlining the planning and execution of robotic shotcrete operations.

4 Experimental evaluation

4.1 Experimental process

We tested our pipeline at a semi-indoor construction site featuring wooden panels, both with and without exposed reinforcement bars. We tested before shotcrete application, as depicted in the upper part of Figure 4. The 3D reconstruction module processed data captured by a Roboception RC-Visard 160 stereo camera. The semantic understanding module inferred masks for areas with exposed rebars.

Figure 4. Upper part: samples of two images of the testing area within the construction site. Lower part: semantically annotated 3D mesh model, where regions of interest are highlighted in cyan.

4.2 Results

Qualitative results are presented in the lower part of Figure 4. The semantic understanding module precisely identified regions of interest, denoted by cyan masks.

We measured the effectiveness of the integrated framework in terms of its geometric precision by applying a manual procedure to measure the point-to-point distances of the wooden frames within the 3D model, corresponding to the segmented area. The vertical and horizontal dimensions of the real wooden frame were measured to be 2 meters long and 1 meter wide. We selected several 3D point pairs in the point cloud (X, Y, Z) to measure their distances and assess the accuracy of the scene reconstruction. Specifically, the experiment involved 12 point-to-point measurements of the frame's dimensions, as illustrated in Figure 5. Subsequently, we compute the root mean square error (RMSE) by comparing the distances of the real wooden frame with the distances of the reconstructed 3D model. The RMSE amounts to 0.564 centimeters.

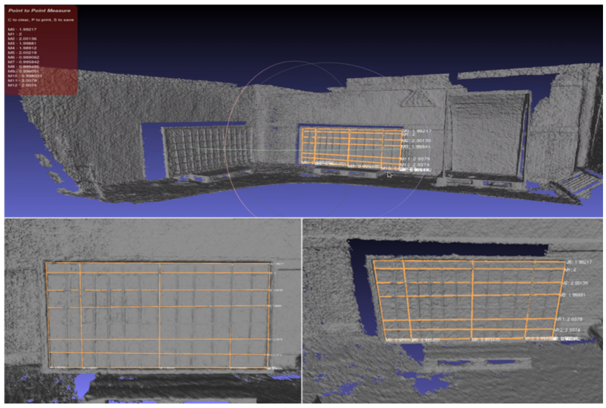

Figure 5. Samples of point-to-point measurements of the 3D reconstruction accuracy

5 Conclusion

We proposed a real-time method for semantic 3D reconstruction for robotics-based construction applications. Our method uses a robust semantic understanding module using a custom YOLOv8m-seg segmentation model and the SiamMask mask tracker, together with RTAB-Map for odometry estimation and InfiniTAM for 3D reconstruction. The resulting semantic 3D mesh model is an important step towards introducing robotic systems in shotcrete construction to improve worker safety and alleviate labor shortages. Our experimental evaluation concluded that the developed method is applicable for semantic understanding and reconstruction of semi-indoor construction scenes, highlighting regions of interest before performing shotcreting. The novel approach was tested under realistic construction site conditions, showcasing good performance in terms of RMSE, mIOU, mean precision and mean recall. We believe that the method can be extended to accommodate other common shotcrete application scenarios, as well as to spark general interest in introducing advanced computer vision approaches to challenging construction tasks.

6 Acknowledgements

This work has been funded and supported by the EU Horizon Europe project "RobetArme" under the Grant Agreement 101058731. We extend our gratitude to Christiansen & Essenbaek A/S (CEAS) for providing access to their premises and for the mock-up construction site.

References

[1] European Construction Sector Observatory. *Digitalisation in the construction sector*. Apr. 2021. URL: https://ec.europa.eu/docsroom/documents/45547/attachments/1/translations/en/renditions/pdf.

[2] European Labour Authority. *Report on labour shortages and surpluses: November 2021.* en. LU: Publications Office, 2021. URL: https://data.europa.eu/doi/10.2883/746322.

[3] URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_-_statistics_by_economic_activity.

[4] Zhiliang Ma and Shilong Liu. "A review of 3D reconstruction techniques in civil engineering and their applications". In: *Advanced Engineering Informatics* 37 (2018), pp. 163–174.

[5] Pingbo Tang et al. "Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques". In: *Automation in construction* 19.7 (2010), pp. 829–843.

[6] Ioannis Kostavelis et al. "RoBétArmé Project: Human-robot Collaborative Construction System for Shotcrete Digitization and Automation through Advanced Perception, Cognition, Mobility and Additive Manufacturing Skills". In: *Open Research Europe* 4 (Jan. 3, 2024), p. 4. ISSN: 2732-5121. DOI: 10.12688/openreseurope.16601.1.

[7] A. Sawhney, M. Riley, and J. Irizarry. *Construction 4.0: An Innovation Platform for the Built Environment (1st ed.)* Routledge, 2020.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: arXiv:1505.04597 (May 2015). arXiv:1505.04597 [cs]. URL: http://arxiv.org/abs/1505.04597.

[9] Shervin Minaee et al. "Image Segmentation Using Deep Learning: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2021.3059968.

[10] Liang-Chieh Chen et al. "Rethinking Atrous Convolution for Semantic Image Segmentation". In: arXiv:1706.05587 (Dec. 2017). arXiv:1706.05587 [cs]. URL: http://arxiv.org/abs/1706.05587.

[11] Liang-Chieh Chen et al. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". In: arXiv:1412.7062 (June 2016). arXiv:1412.7062 [cs]. URL: http://arxiv.org/abs/1412.7062.

[12] Liang-Chieh Chen et al. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: arXiv:1606.00915 (May 2017). arXiv:1606.00915 [cs]. URL: http://arxiv.org/abs/1606.00915.

[13] Jingdong Wang et al. "Deep High-Resolution Representation Learning for Visual Recognition". In: (2019). DOI: 10.48550/ARXIV.1908.07919.

[14] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: arXiv:1506.02640 (May 2016). arXiv:1506.02640 [cs]. URL: http://arxiv.org/abs/1506.02640.

[15] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection". In: *arXiv:2004.10934 [cs, eess]* (Apr. 2020). arXiv: 2004.10934. URL: http://arxiv.org/abs/2004.10934.

[16] Zheng Ge et al. "YOLOX: Exceeding YOLO Series in 2021". In: *arXiv:2107.08430 [cs]* (Aug. 2021). arXiv: 2107.08430. URL: http://arxiv.org/abs/2107.08430.

[17] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *arXiv:1804.02767 [cs]* (Apr. 2018). arXiv: 1804.02767. URL: http://arxiv.org/abs/1804.02767.

[18] Chien-Yao Wang, I.-Hau Yeh, and Hong-Yuan Mark Liao. "You Only Learn One Representation: Unified Network for Multiple Tasks". In: *arXiv:2105.04206 [cs]* (May 2021). arXiv: 2105.04206. URL: http://arxiv.org/abs/2105.04206.

[19] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Ultralytics YOLOv8*. 2023. URL: https://github.com/ultralytics/ultralytics.

[20] Alexander Kirillov et al. "Segment Anything". In: (2023). DOI: 10.48550/ARXIV.2304.02643.

[21] Wenhui Wang et al. "Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks". In: arXiv:2208.10442 (Aug. 2022). arXiv:2208.10442 [cs]. URL: http://arxiv.org/abs/2208.10442.

[22] Ben Mildenhall et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *CoRR* abs/2003.08934 (2020).

[23] Thomas Müller et al. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". In: *CoRR* abs/2201.05989 (2022). arXiv: 2201.05989. URL: https://arxiv.org/abs/2201.05989.

[24] Kunlong Hong, Hongguang Wang, and Bingbing Yuan. "Inspection-Nerf: Rendering Multi-Type Local Images for Dam Surface Inspection Task Using Climbing Robot and Neural Radiance Field". In: *Buildings* 13.1 (2023). ISSN: 2075-5309.

[25] Martin Fisher Shun Hachisuka Alberto Tono. "Harbingers of NeRF-to-BIM: a case study of semantic segmentation on building structure with neural radiance fields," in: 2023.

[26] Johannes L. Schönberger et al. "Pixelwise View Selection for Unstructured Multi-View Stereo". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 501–518. ISBN: 978-3-319-46487-9.

[27] Dimitrios Katsatos et al. "Comparative Study of Surface 3D Reconstruction Methods Applied in Construction Sites". In: *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE. 2023, pp. 1–6.

[28] Zhexiong Shang and Zhigang Shen. "Real-Time 3D Reconstruction on Construction Site Using Visual SLAM and UAV". In: 2018.

[29] Richard A. Newcombe et al. "KinectFusion: Real-time dense surface mapping and tracking". In: *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. 2011, pp. 127–136. DOI: 10.1109/ISMAR.2011.6092378.

[30] Victor Adrian Prisacariu et al. "InfiniTAM v3: A Framework for Large-Scale 3D Reconstruction with Loop Closure". In: *CoRR* abs/1708.00783 (2017). arXiv: 1708.00783.

[31] Helen Oleynikova et al. "Voxblox: Building 3D Signed Distance Fields for Planning". In: *CoRR* abs/1611.03631 (2016). arXiv: 1611.03631. URL: http://arxiv.org/abs/1611.03631.

[32] Mathieu Labbé and François Michaud. "RTAB-Map as an open-source lidar and visual SLAM library for large-scale and long-term online operation". In: *Journal of Field Robotics* 36.2 (2019), pp. 416–446. DOI: https://doi.org/10.1002/rob.21831.

[33] Bo Xiao and Shih-Chung Kang. "Development of an Image Data Set of Construction Machines for Deep Learning Object Detection". en. In: *Journal of Computing in Civil Engineering* 35.2 (Mar. 2021), p. 05020005. ISSN: 0887-3801, 1943-5487. DOI: 10.1061/(ASCE)CP.1943-5487.0000945.

[34] URL: https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection.

[35] Liangbin Xie. *Hardhat*. 2019. DOI: 10.7910/DVN/7CBGOS.

[36] Rui Duan et al. "SODA: A large-scale open site object detection dataset for deep learning in construction". In: *Automation in Construction* 142 (2022). Citation Key: DUAN2022104499, p. 104499. ISSN: 0926-5805. DOI: https://doi.org/10.1016/j.autcon.2022.104499.

[37] Martin Mundt et al. "Meta-Learning Convolutional Neural Architectures for Multi-Target Concrete Defect Classification With the COncrete DEfect BRidge IMage Dataset". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 11188–11197. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.01145.

[38] Jiesheng Yang et al. "Computer Vision for Construction Progress Monitoring: A Real-Time Object Detection Approach". In: arXiv:2305.15097 (May 2023). arXiv:2305.15097 [cs]. URL: http://arxiv.org/abs/2305.15097.

[39] Haoyu Wang et al. "Synthetic Datasets for Rebar Instance Segmentation Using Mask R-CNN". en. In: *Buildings* 13.3 (Feb. 2023), p. 585. ISSN: 2075-5309. DOI: 10.3390/buildings13030585.

[40] Qiang Wang et al. "Fast Online Object Tracking and Segmentation: A Unifying Approach". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 1328–1338. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00142.

[41] F. Perazzi et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, June 2016, pp. 724–732. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.85.

[42] Patrick Schmidt. *RobetArme Stabilized SiamMask output*. Youtube. 2024. URL: https://www.youtube.com/watch?v=In6euNHw_Sk.