Cross Domain Matching for Semantic Point Cloud Segmentation based on Convolutional Neural Networks

Martens J., Blut T., Blankenbach J.
Geodesic Institute and Chair for Computing in Civil Engineering & Geo Information Systems,
RWTH Aachen University, Germany
jan.martens@gia.rwth-aachen.de

Abstract. Together with roads, rails and tunnels, bridges represent ubiquitous infrastructure objects holding an essential role in transportation. Nevertheless, due to the elevated age of many bridges, structural deficiencies are becoming more common. Digital Twins in conjunction with Building Information Modelling (BIM) may significantly support asset management but are only now garnering more attention for infrastructure objects. To ease the transition towards creating digital twins for existing constructions, this work presents a method for automated segmentation of bridge point clouds based on images using convolutional neural networks. For this purpose, semantic segmentation is used for labelling the photographs captured during laser scanning. The classifications masks of this image-based approach are then projected back into 3D, resulting in a labelled point cloud ready for further processing and building component reconstruction.

1. Introduction

The shift towards digitization has led to a notable transformation in the architecture, engineering and construction (AECO) industry. Since its introduction, Building Information Modelling (BIM) is being adopted in civil engineering for digital design and construction and is also seen in the life cycle and asset management due to apparent benefits for data handling and exchange (Blankenbach 2018). While the IFC data model has matured for real estate objects over the past years (Borrmann et al. 2018), an extension of it tailored towards infrastructure objects is only recently being pushed forward (Borrmann et al. 2019). This development makes sense, as infrastructure objects are commonly encountered in everyday life. It also emphasizes the role of BIM, which serves as the origin for the concept of the term “digital twin” in construction and describes data-driven systems for the monitoring, maintenance and management of objects throughout their life cycle (Sacks et al. 2020; Errandonea, Beltran and Arrizabalaga 2020).

However, the creation of semantic-rich 3D models representing the actual (as-is) state of the construction as important basis for digital twins is a notable challenge, as up-to-date plans are rarely present or incomplete as repairs and restructuring may have taken place over the life span of the construction. In consequence, reality capture techniques such as laser scanning or photogrammetry must be used for as-is data acquisition, followed by elaborative manual modeling based on the collected data. This step requires specialized modelling software, trained staff and time. Automated as-is modelling is therefore a hot topic in both, research and the industry, as it cuts down on modelling time and costs. Typically, this process can be broken down into four stages as illustrated in Figure 1.

Capturing and surveying represent the first of these stages and have improved notably over the past years due to unmanned data capturing platforms (drones), laser scanning systems and capturing methods in general having evolved towards high accuracy, low-cost, ease-of-use and fast capturing (Blankenbach 2018). Because of some capturing methods suffering from reduced accuracy, preprocessing aims to rectify the impact of sensor noise and artefacts to clean up the point cloud data and guarantee optimal conditions for semantic segmentation and

https://doi.org/10.7146/aul.455.c209

modelling. Segmentation is typically achieved using geometric algorithms; however methods based on machine learning are becoming an increasingly more popular option. While segmentation can be performed to extract specific regions of interest and planar patches, it can also be used to find and label specific objects and structures. Finally, these are then used to create suitable building components and construct the final 3D model. Depending on whether or not additional information has been gathered during the segmentation step, it may be added to the model alongside semantic attributes. This means that segmentation and modelling are closely intertwined, as the model quality will strongly depend on the segmentation quality.

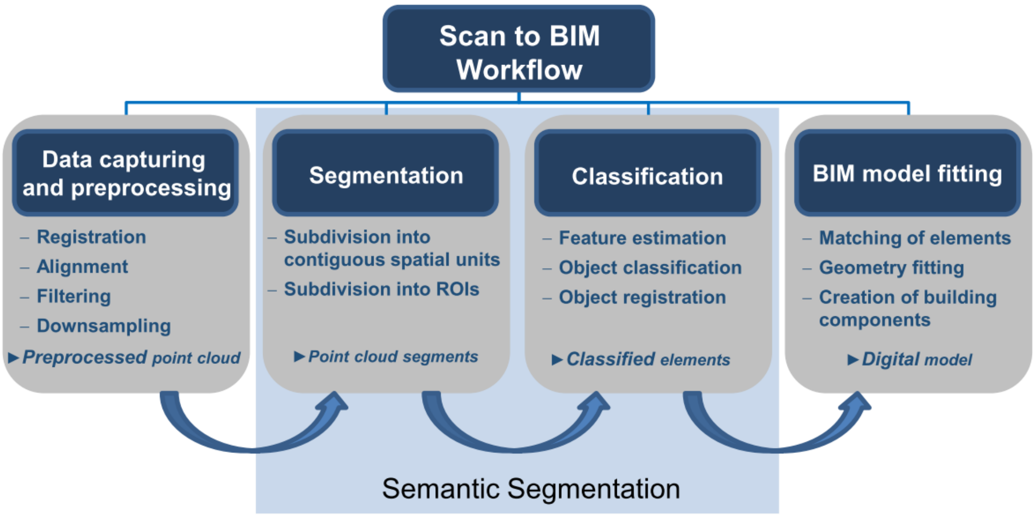

Figure 1: Stages of the Scan-to-BIM process. The presented approach covers the segmentation and classification stages.

Both, segmentation and classification may also be combined into one single step commonly referred to as semantic segmentation. Supervised machine learning methods are already increasingly being used for this purpose, with Deep Learning (DL) in particular holding much potential. However, (supervised) DL requires a large amount of training data with high variability and annotated point clouds are rare or even not available for many construction types like bridges. As part of a larger research project with the overarching goal of automatically deriving digital bridge models from survey data, we try to tackle this problem by a multi-stage semantic segmentation workflow.

Arising from the lack of point cloud training data, this contribution presents an image-based classification approach where 2D semantic segmentation results are transferred to 3D point clouds (cross-domain matching) in order to achieve point cloud segments which can afterwards be refined using prior knowledge. The approach is driven by the fact that photos are usually taken in addition to a point cloud as part of the surveying. The approach is depicted in Figure 2 and shows a neural network performing an initial instance-based pre-segmentation on photos. Segmentation masks then being projected as a coarse segmentation into the point cloud, in order to then carry out a fine segmentation. The advantage of this approach is that on one hand, training data (photos of infrastructure objects) is publicly available (e.g. on the internet) and can also be produced quite easily and quickly. On the other hand, due to neural networks being famous for their performance in image classification tasks, the point cloud data required for training within the fine segmentation step being rather scarce. Although the approach still holds potential for improvement, the presented results prove that it already delivers solid results when combined with suitable post-processing techniques, making it a promising tool for the integration into modelling workflows.

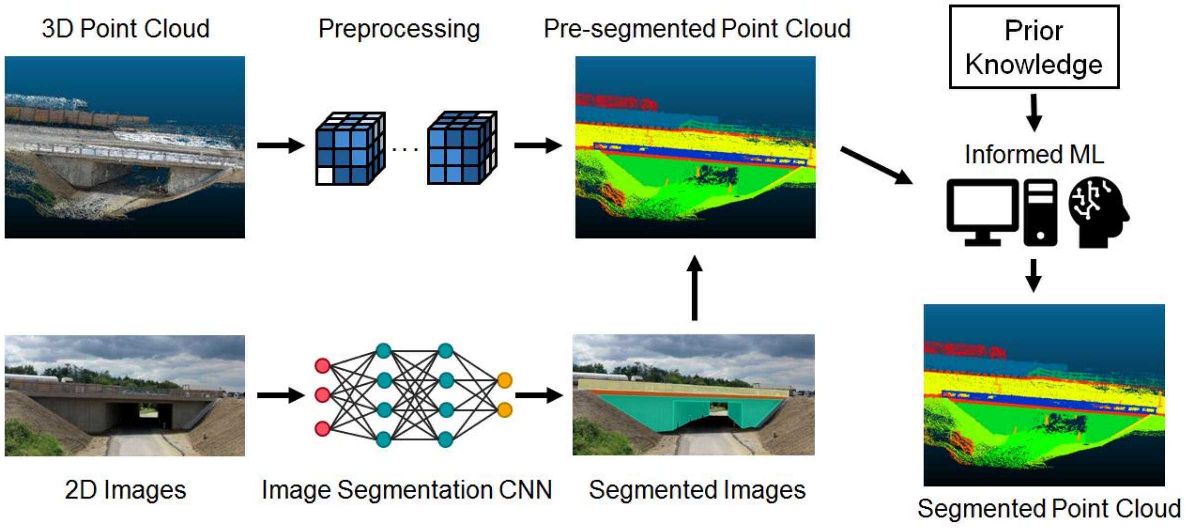

Figure 2: Overall workflow for infrastructure point cloud semantic segmentation. This work deals with the image segmentation part and generation of a pre-segmented point cloud.

2. Related Works

For years, machine learning methods for point clouds have been an ongoing research topic in the field of robotics with most early approaches using the then widely-used Support Vector Machines (SVM) alongside hand-crafted features (Brodu and Lague 2012). At roughly the same time, neural networks and deep learning have gained more traction in object recognition, as their requirements for large training data sets and computational power were starting to get satisfied. Since the massive success of neural networks for image classification and the subsequent improvements of their architecture in the following years (Krizhevsky, Sutskever and Hinton 2017), early experiments in voxel-based approaches such as VoxNet (Maturana and Scherer 2015). In the face of their limitations, native 3D-based neural networks such as PointNet (Qi et al. 2016) have been developed. Interestingly, hybrid approaches which combine information from the point cloud and image domains have been presented as well (Sindagi, Zhou, and Tuzel 2019).

In an effort of exploiting the high accuracy of image-based neural networks, a handful approaches have already been developed to act as indoor object classifiers (Su et al. 2015, Stojanovic et al. 2019). These methods classify isolated objects or pre-segmented point clouds by rendering them from different viewports and classifying the resulting images. Adaptations for use with automated driving have been made as well (Wolf et al. 2019), proving that this approach can not only be applied to outdoor environments, but even has a better chance of classifying objects in their spatial context than most native 3D approaches.

When it comes to capturing training data, MLS has become widely used for capturing large infrastructure objects, as it benefits from high mobility and capturing speeds (Ma et al. 2018). Due to the application of machine learning in the field of automated driving being pushed into focus recently, urban point cloud data sets captured with MLS such as KITTI (Fritsch, Kuehnl and Geiger 2013) and Paris-Lille-3D (Roynard, Deschaud and Goulette 2018) have become popular benchmarks for the detection of cars, pedestrians and road elements. The crux however lies in the fact that these datasets are not concerned with infrastructure objects such as bridges. Consequently, 3D training data specifically for bridges is highly-limited, rendering 3D-based neural networks unsuitable for bridge component classification.

Given this context, image-based classification approaches like the one employed in this work represent the best option for 3D classification tasks where training is either sparse or non-existent. Image datasets are rather ubiquitous, with the COCO dataset (Lin et al. 2014) being one well-known example and even datasets with bridge data (albeit with labels for cracks and damages rather than components) are available (Bianchi and Hebdon 2021).

This accessibility of existing data makes image-based semantic segmentation a valuable tool for the underlying point cloud classification problem. Due to these techniques being part of the digital twin reconstruction process, their role in capturing the bridge's state and updating the digital model accordingly aligns well with the goal of maintenance and damage tracking of the Industry 4.0 movement (Shim et al. 2019).

3. Methodology

The presented image-based workflow for point clouds can be subdivided into four incremental stages: neural network training, image-based semantic classification, 3D projection and post-processing.

3.1 Neural Network Training

As discussed earlier, the problem of obtaining specific training data for bridges poses a problem due to the lack of labelled bridge images. However, transfer learning represents a solution to this problem (Zhu et al. 2021, Kora et al. 2022). Neural networks for image classification typically consist of a feature detection body where distinct features are recognized and a classification head which associates these features with one of the output classes. Transfer learning describes the process of modifying a pre-trained network such that the features learned in the body are kept, while the classification head is being re-trained and adapted for a new set of object classes. Thus, training is drastically reduced and as an added benefit, the time-consuming and tedious hyper parameter tuning process is omitted.

3.2 Image-based Semantic Segmentation

For the classification of the captured images, Mask R-CNN network (He et al. 2017), an extension of Faster R-CNN was used. Mask R-CNN relies on a feature detection backbone such as ResNet, ResNeXt or SpinNet for feature detection and uses RoIAlign for proposing Regions of Interest (RoI). These regions are subsequently classified by one network branch, while a second branch predicts an object mask for each RoI. Due to Mask R-CNN generating not only object bounding boxes and annotations, but also pixel-accurate classification masks, the resulting regions of interest are well-suited for projection into 3D point clouds, given the camera's extrinsic and intrinsic parameters.

3.3 3D Projection and Postprocessing

Extrinsic camera parameters represent spatial camera parameters in 3D space and encompass location and orientation parameters. They can be captured directly through GNSS/IMU data and tracking of scan positions when using ~~drones, TLS and MLS~~ systems or, if camera systems without GNSS/IMU are used, can additionally be reconstructed through image matching methods. Intrinsic parameters on the other hand represent the focal length, coordinates of the principal point and distortion parameters of the camera and are therefore needed to describe mathematically and physically how captured images are projected onto the image plane. Typically, a camera calibration process is used to estimate intrinsic parameters and the type of distortion present in the images. Using the extrinsic and intrinsic camera parameters, it is possible to project the 3D point cloud into the camera's image plane.

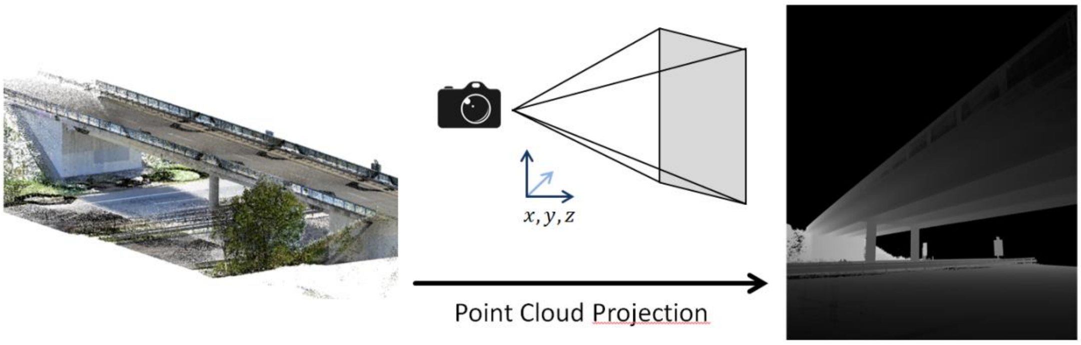

Figure 3: Point cloud projection for the solving occlusion problem through depth map generation. This process requires intrinsic and extrinsic camera parameters to render out each point's distance from a given camera position.

As apparent from their lack of 3D information, a naive projection of 2D images into point clouds will run into the problem of labelling points occluded by the ones in front of them, leading to many points being labelled incorrectly. Regular photographs are missing the required depth information, but through use of the extrinsic and intrinsic camera parameters, point clouds can be aligned with the images, thus bringing them into the camera's view frustum. In a process inspired by computer graphics, this allows for the calculation of each point's distance to the camera. By rendering out these distances into a depth map (also referred to as z buffer or depth map), it is possible to solve the depth occlusion problem. As shown in Figure 3, the depth map is created by constructing a view matrix from the extrinsic parameters and projection matrix from the intrinsic parameters and multiplying all visible points with it. In the case of multiple points falling into the same pixel of the depth map, only the one with the smallest distance to the camera is kept for labelling, due to it occluding the ones behind it.

Once the projection process is complete, candidate labels for each point are post-processed using majority voting. Most points are typically visible from multiple views, which can lead to contradicting labels due to classification masks potentially overlapping, being imprecise or in some cases belonging to the wrong class. The majority voting allows for these problems to be resolved in a simple yet effective way, but shows potential for improvements due to its ignorance towards geometric structures.

4. Experiments

For training, a variety of images ranging UAS (Unmanned Aerial System) surve~~y dat~~a taken by the TU Munich to images taken from Internet search engines as well as selfrecorded images and those from bridge inspections were used. Validation was performed on survey data of two different bridges. The first survey was done using a camera-equipped UAS. ~~Based on the~~ captured video footage, a point cloud was reconstructed, meaning that image data and a point cloud suffering from quality degradations characteristic of image reconstruction techniques were available. A second survey was performed using a geodetic terrestrial~~ laser scanner~~ (TLS) equipped with a NIKON D800 camera an~~d based on~~ 20 scan positions 140 pictures and a high-quality point cloud were obtained. In both cases, intrinsic and extrinsic camera parameters were known.

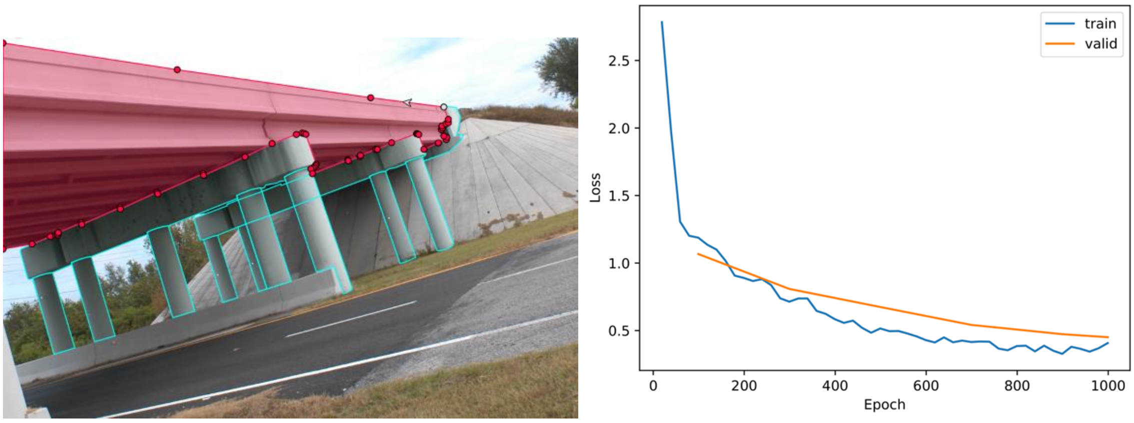

Figure 4: Left: Sample image of the provided training data set. Right: Training and validation loss for Mask R-CNN on the given training dataset.

For the classification of the captured images, a Mask R-CNN network with a backbone network consisting of a ResNet with 50 layers which has been pre-trained on the COCO dataset (Lin et al. 2014). To improve classification accuracy, transfer learning was applied, where the original network was modified and retrained on a hand-annotated bridge dataset to detect and classify bridge abutment, railing and deck components. The corresponding data set for retraining was created from around 600 hand-annotated images (a sample image is depicted in Figure 4, left) and expanded using typical data augmentation operations such as cropping, rescaling and mirroring to make the resulting classifier more robust. Training and validation loss dropped sharply during the first 600 iterations, before slowly converging as indicated by Figure 4 (right).

As shown in Figure 5, results for classification show that the trained network performs favourably for most images and can reliably detect the three object classes (abutment, railing and deck) within the images regardless of camera distance and angle to the object. Detection masks, however, have a tendency towards displaying a jagged rather than smooth border, making these regions somewhat unreliable. Masks for the railing class are occasionally slightly imprecise, which may be a result of their fine structure which allows for objects behind them to still be visible. Projections of the classification masks into the point cloud result in a decent segmentation of the bridge components alongside some unpleasant artefacts (see Figure 5, right column). While the majority voting algorithm and depth occlusion are capable of resolving glaring issues, additional post-processing would improve result quality even further.

Despite different camera parameters, TLS survey data was processed analogously to the UAS~~ data.~~ Results for this dataset are similar in terms of robustness but also present different challenges. Unlike the UA~~S sur~~vey data, where the bridge is always in frame, it also contains images in which the bridge not visible. One should note that as a form of negative control, the resulting classification masks of these images are empty and thus do not contribute to the overall point cloud classification process. Additionally, some scans were acquired underneath the bridge and show components which are typically occluded from view when looking at it from an outside perspective and were not included in the training data set. Reliability of classifications masks in these regions is therefore lower, such that components are occasionally correctly recognized despite the lack of training data but also oftentimes incorrectly labelled or not recognized. This issue can be resolved through retraining with the appropriate images though.

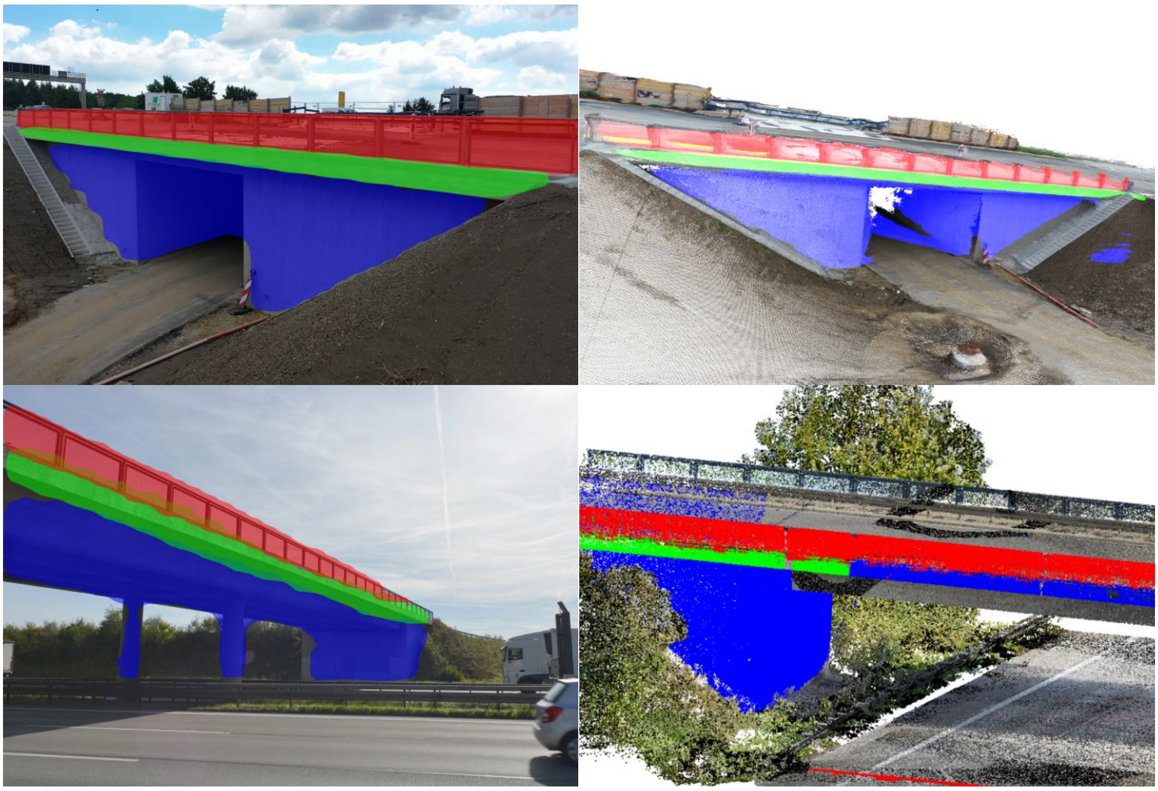

Figure 5: Results of image-based segmentation for two different bridges. Top left: Image captured by a UAS ~~with~~ superimposed segmentation masks. Top right: Point cloud with labels from multiple viewpoints projected onto it. Bottom left: Image captured during TLS survey with superimposed classification masks. Bottom right: TLS point clouds with labels from multiple viewpoints projected onto it.

Another challenge occurs for images where the bridge is visible from an oblique angle (shown in Figure 5, bottom left) and where the perspective results in a severely reduced resolution of the deck and railing components. Classification masks have a tendency to be less accurate in both cases, as the lower resolution of these areas makes it harder to make out defining features, such that deck and railing components are more likely to be confused for one another as apparent in Figure 5 (bottom right). A projection of these masks into the point cloud labels the abutment consistently correctly, but segmentation of deck and railing can suffer from degraded quality in aforementioned cases. While the method already works admirably well in most cases, this factor proves that retraining and further post-processing e.g. using spatial and geometric reasoning is required to achieve a more consistent result quality.

5. Discussion and Outlook

As shown by the early results, the presented approach provides a viable way of semantic segmentation for point clouds using supervised machine learning and is capable of bypassing the current scarcity of point cloud training data. Semantic segmentation is used to generate object masks, which are projected onto the point cloud. Occlusions specific to camera perspectives are being simulated by rendering out a depth map using the underlying camera parameters. Afterwards, a post-processing cleans up point labels to improve result quality by means of a majority voting for point with multiple overlapping regions. Among the observed challenges are inaccurate region borders, misclassifications of low-resolution regions and occasional outliers. Further investigations into the performance of image classification are currently being carried out, involving commonly-used metric such as Intersection over Union (IoU), precision and recall. An investigation of the same metrics for the labelled point clouds are of major interest as well.

Aside from the image segmentation quality which can be improved by adding more data to the training process, the most apparent way of dealing with inaccurate results and improving segmentation quality lies in the post-processing step. The success of earlier works in image classification, where post-processing has led to improvements of the overall results, points towards the same conclusion (Tivive and Bouzerdoum 2006). In its current state, the overall process does not take the spatial context into account and is thus limited when it comes to dealing with wrong or incomplete labels. Methods such as 3D-nearest neighbour labelling for unlabelled points are obvious methods to include, but more sophisticated methods which pre-segment the 3D point cloud and then propagate labels inside these segments are more promising. In a similar vein, the use of spatial reasoning for correcting labels based on object height, surface structure and orientation should yield improved results and will help clean up undesirable classification artefacts. This sort of prior knowledge can also be integrated into the machine learning process itself to further lessen the impact of limited training data (von Rueden et al. 2021) and improve results further, as indicated in Figure 2.

Extensions for post-processing and an introduction of new classes for a more granular segmentation will be another issue worth investigating, as bridges consist of more components than the ones represented in this work. Especially automated reconstruction will benefit from this factor, as the presented image-based segmentation can help inform the choice of reconstruction algorithm for each individual building component. After all, with the current scarcity in labelled point cloud data for infrastructure objects, the presented approach thus represents a key element automated reconstruction workflows.

Acknowledgements

This work has been supported and funded by the German Ministry of Transportation (BMDV) as part of the “TwinGen” research project. The authors thank their project partners from the Chair of Computational Modelling and Simulation at the Technical University of Munich, Computing in Engineering at Ruhr University Bochum, Design Computation at RWTH Aachen University and ZPP Ingenieure AG.

References

He, K., Gkioxari, G., Dollár, P., Girshick, R. (2018). Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988.

Qi, C.R., Su, H., Mo, K., Guibas, L.J. (2016). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. arXiv preprint arXiv: 1612.00593.

Maturana, D., Scherer, S., (2015). VoxNet: A 3D Convolutional Neural Network for real-time object recognition, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 922-928, doi: 10.1109/IROS.2015.7353481.

Sindagi, V.A., Zhou, Y., Tuzel, O. (2019). MVX-Net: Multimodal VoxelNet for 3D Object Detection, 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 7276-7282, doi: 10.1109/ICRA.2019.8794195.

Wolf, J., Richter, R., Discher, S., Döllner, J. (2019). Applicability of Neural Networks for Image Classification on Object Detection in Mobile Mapping 3D Point Clouds. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLII-4/W15. 111-115. 10.5194/isprs-archives-XLII-4-W15-111-2019.

Stojanovic, V., Trapp, M., Richter, R., Döllner, J. (2019). Classification of Indoor Point Clouds Using Multiviews. In The 24th International Conference on 3D Web Technology (Web3D '19). Association for Computing Machinery, New York, NY, USA, 1–9. DOI:https://doi.org/10.1145/3329714.3338129

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E. (2015). Multi-view Convolutional Neural Networks for 3D Shape Recognition, 2015 IEEE International Conference on Computer Vision (ICCV), pp. 945-953, doi: 10.1109/ICCV.2015.114.

Brodu, N., Lague, D. (2012). 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 68, 2012, pp. 121-134, ISSN 0924-2716 doi: https://doi.org/10.1016/j.isprsjprs.2012.01.006

Krizhevsky, A., Sutskever, I., Hinton, G., E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (June 2017), 84–90. doi: https://doi.org/10.1145/3065386

Lin, T.Y., Maire, M. Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. Processings of Computer Vision ECCV 2014, Springer International Publishing. pp. 740-755, ISBN: 978-3-319-10602-1

Fritsch, J., Kuehnl, T., Geiger, A. (2013). A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms. In: International Conference on Intelligent Transportation Systems (ITSC), 2013.

Roynard, X., Deschaud, J., Goulette, F. (2018). Paris-Lille-3D: A Point Cloud Dataset for Urban Scene Segmentation and Classification, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2108-21083, doi: 10.1109/CVPRW.2018.00272.

Bianchi, E., Hebdon, M. (2021). COCO-Bridge 2021+ Dataset. doi: 10.7294/16624495.v1, https://data.lib.vt.edu/articles/dataset/COCO-Bridge_2021_Dataset/16624495

Borrmann A., Beetz J., Koch C., Liebich T., Muhic S. (2018) Industry Foundation Classes: A Standardized Data Model for the Vendor-Neutral Exchange of Digital Building Models. In: Borrmann A., König M., Koch C., Beetz J. (eds) Building Information Modeling. Springer, Cham. https://doi.org/10.1007/978-3-319-92862-3_5

Blankenbach J. (2018). Building Surveying for As-Built Modeling. In: Borrmann A., König M., Koch C., Beetz J. (eds) Building Information Modeling. Springer, Cham. https://doi.org/10.1007/978-3-319-92862-3_24

Borrmann, A., Muhic, S., Hyvarinen, J., Chipman, T., Jaud, S., Castaing, C., Dumoulin, C., Liebich, T., Mol, L. (2019). The IFC-Bridge Project – Extending the IFC standard to enable high-quality exchange of bridge information models. In: Proceedings of the 2019 European Conference on Computing in Construction, pp. 377-386, ISBN: 978-1-910963-37-3

Tivive, F. H. C., Bouzerdoum, A., (2006). Texture Classification using Convolutional Neural Networks. TENCON 2006 - 2006 IEEE Region 10 Conference, 2006, pp. 1-4, doi: 10.1109/TENCON.2006.343944.

Zhu, W., Braun, B., Chiang, L.H., Romagnoli, J.A. (2021). Investigation of transfer learning for image classification and impact on training sample size, Chemometrics and Intelligent Laboratory Systems, Volume 211, 2021, ISSN 0169-7439, doi: https://doi.org/10.1016/j.chemolab.2021.104269.

Kora, P., Ooi, C.P., Faust, O., Raghavendra, U., Gudigar, A., Chan, W.Y., Meenakshi, K., Swaraja, K., Plawiak, P., Rajendra Acharya, U. (2022). Transfer learning techniques for medical image analysis: A review, Biocybernetics and Biomedical Engineering, Volume 42, Issue 1, 2022, pp. 79-107, ISSN 0208-5216, doi: https://doi.org/10.1016/j.bbe.2021.11.004.

Ma L, Li Y, Li J, Wang C, Wang R, Chapman MA. Mobile Laser Scanned Point-Clouds for Road Object Detection and Extraction: A Review. *Remote Sensing*. 2018; 10(10):1531. https://doi.org/10.3390/rs10101531

Sacks, R., Brilakis, I., Pikas, E., Xie, H., & Girolami, M. (2020). Construction with digital twin information systems. Data-Centric Engineering, 1, E14. doi:10.1017/dce.2020.16

Shim, C.-S., Dang, N.-S., Lon, S., Jeon, C.-H. (2019) Development of a bridge maintenance system for prestressed concrete bridges using 3D digital twin model, Structure and Infrastructure Engineering, 15:10, 1319-1332, doi: 10.1080/15732479.2019.1620789

Errandonea, I., Beltran, S., Arrizabalaga, S. (2020). Digital Twin for maintenance: A literature review, Computers in Industry, Volume 123, 2020, ISSN 0166-3615, https://doi.org/10.1016/j.compind.2020.103316.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramanmurthy, R., Garcke, J., Bauckhage, C., Schuecker, J. (2021). Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems, in IEEE Transactions on Knowledge and Data Engineering, May 2021, doi: 10.1109/TKDE.2021.3079836.