Automated Segmentation of Stone and Mortar in Heritage Structures: A Case Study on the Old Town Bridge Tower in Prague

Jakub Vynikal *[1], Lukáš Běloch [1], Tomáš Bouček [1]

[1] Dept. of Geomatics, Faculty of Civil Engineering, Czech Technical University in Prague, Prague, Czech Republic – (jakub.vynikal, lukas.beloch, tomas.boucek)@fsv.cvut.cz

Keywords: Photogrammetry, U-Net, Heritage Documentation, Segmentation, Deep Learning.

Abstract

Accurate digital documentation of heritage structures is vital for conservation, restoration, and structural analysis. Traditional methods for analyzing masonry are time-consuming and subjective. This study proposes a deep learning-based approach using a U-Net convolutional neural network to automatically segment stone and mortar in heritage masonry, trained on high-resolution imagery of Prague’s Old Town Bridge Tower. Unlike prior studies focused on distinct brick structures, our dataset presents a greater challenge due to the similar textures of stone and mortar. Data was collected using a DJI M300 drone with a P1 camera and an RTC360 laser scanner, capturing the entire tower and its interior. The resulting 3D reconstructions and orthophotos, with a 1 mm ground sampling distance, enabled precise manual segmentation of all stones, excluding non-masonry features. After splitting the available manually annotated data, U-Net models with differing parameters were trained on the train set and evaluated on a test set, achieving a class-averaged F1 score of up to 85.58%. The created segmentation maps can be easily converted to finished vector drawings. Results show that deep learning significantly improves segmentation speed and consistency over manual methods. These maps support conservation tasks such as structural monitoring and damage detection. The trained model will aid future documentation of the Charles Bridge, illustrating the potential of AI in advancing scalable, objective digital heritage conservation.

1. Introduction

Accurate digital documentation of heritage structures is critical for their conservation, restoration, and structural assessment. Advances in photogrammetry and terrestrial laser scanning have made it possible to capture highly detailed representations of historical buildings. However, interpreting this data—particularly segmenting structural components like stones and mortar—remains a manual, time-consuming, and subjective process.

Traditional approaches to masonry analysis often rely on manual annotation or rule-based image segmentation, which are not only labor-intensive but also prone to inconsistencies. Recently, deep learning-based methods have shown promise for automating such tasks (Ibrahim et al., 2019, Dais et al., 2021), offering improved consistency and scalability. Yet, most existing research in this area has focused on relatively regular masonry patterns, such as modern brick walls with uniform materials and clearly distinguishable joints. These approaches are less effective when applied to more complex and heterogeneous historical masonry, where the distinction between stone and mortar is subtle and highly variable.

In this study, we address this gap by proposing a deep learning-based approach for segmenting stone and mortar in historical masonry using high-resolution imagery. Our work focuses on the Old Town Bridge Tower in Prague, a Gothic-era structure characterized by complex stone patterns and only subtle visual contrast between structural elements. Unlike previous datasets that primarily consist of regular, high-contrast bricks, our dataset presents greater visual ambiguity, necessitating a tailored segmentation strategy and robust training data. To this end, we have created a comprehensive, high-resolution dataset of the Bridge Tower, annotated manually to distinguish between stone and mortar while excluding architectural elements irrelevant to structural analysis.

We employed a U-Net (Ronneberger et al., 2015) model, a convolutional neural network (CNN) designed for pixel-wise image segmentation. The U-Net architecture consists of an encoder-decoder structure, where the encoder captures spatial features through progressively deeper convolutional layers, while the decoder reconstructs fine-grained segmentation maps using upsampling and skip connections. This design allows the network to learn both local texture details and broader structural patterns, making it particularly well-suited for segmenting architectural elements with complex spatial arrangements. The U-Net network has consistently outperformed other CNNs in various case studies and is considered versatile and robust (Pešek et al., 2024a, 2024b). We perform some experiments with variable loss function and thresholding value to find the most suitable combination of parameters achieving the best metrics on the test split.

This paper demonstrates the feasibility and effectiveness of using deep learning for detailed masonry segmentation in heritage contexts. Our results suggest that such methods can significantly enhance both the speed and objectivity of documentation workflows, and we discuss their implications for future conservation efforts, including the planned documentation of the Charles Bridge—constructed with materials and methods similar to those of the Bridge Tower.

Our key contributions include: (1) creation of a detailed annotated dataset of historic masonry, (2) application of U-Net for stone–mortar segmentation, and (3) evaluation of performance based on variable training losses and thresholding parameters with practical conservation use cases in mind.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-1587-2025 | © Author(s) 2025. CC BY 4.0 License.

2. Materials and Methods

2.1 Location and data collection

The Old Town Bridge Tower is located in Prague and serves as an entry point to Charles Bridge, the most well-known historical bridge in Central Europe. Dating back to 14th century, the bridge is currently undergoing a twenty-year process of structural inspections, restoration, and repairs. A reconstruction undertaken in 2008-2010, necessary after the disastrous 2002 floods, was widely seen as unprofessional and without adequate conservation advice on materials and techniques. Many original stones were lost, damaged or inappropriately positioned and replaced with excessive amount of new masonry. This prompts for better and more detailed documentation of the current masonry in order to avoid such mistakes in the current restoration efforts.

To document the Bridge Tower, extensive photogrammetric documentation was carried out. Data acquisition was conducted using a DJI M300 drone equipped with a P1 camera and an RTC360 terrestrial laser scanner, enabling comprehensive documentation of the structure. Both the interior and exterior were scanned. To capture texture of the interior, a SLR camera was also used. The dataset includes high-resolution imagery and point clouds capturing all four exterior sides of the tower (two of which are depicted in Figure 1), four sides of the interior staircase, and four sides of an interior room. The collected data was processed in RealityCapture to generate detailed 3D reconstructions and high-accuracy orthomosaics, resulting in a ground sampling distance (GSD) of 1 mm, allowing for fine-grained segmentation suitable for structural assessments and conservation planning.

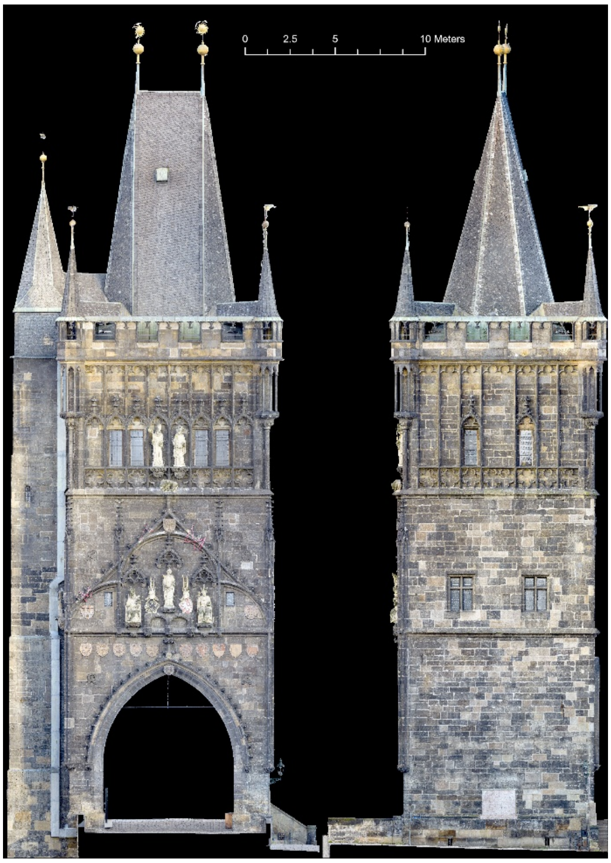

Figure 1. Eastern and northern sides of the Old Town Bridge Tower

2.2 CNNs and U-Net

Convolutional neural networks (CNNs) are a class of deep learning models particularly well-suited for processing grid-structured data such as images. By applying convolutional filters across the spatial dimensions of an image, CNNs are able to learn hierarchical feature representations—from low-level patterns like edges and textures to high-level semantic structures (LeCun et al., 1998; Krizhevsky et al., 2012). While early CNN architectures were designed primarily for image classification, numerous extensions have adapted them for pixel-wise prediction tasks such as semantic segmentation.

Semantic segmentation requires assigning a categorical label to each individual pixel in an image. This is more challenging than image classification, as it necessitates both fine-grained spatial localization and global contextual understanding. Fully convolutional networks (FCNs) pioneered the adaptation of CNNs for segmentation by replacing fully connected layers with convolutional ones, thus preserving spatial dimensions throughout the network (Long et al., 2015). However, FCNs tended to produce coarse outputs due to the loss of spatial resolution in deep network layers.

The U-Net architecture (Ronneberger et al., 2015) shown schematically in Figure 2, originally developed for biomedical image segmentation, addressed these limitations by introducing a symmetric encoder-decoder structure with skip connections. The encoder (contracting path) progressively reduces spatial resolution while capturing increasingly abstract feature representations through repeated applications of convolution and max-pooling operations. The decoder (expanding path) restores the original resolution using up-convolutions (transposed convolutions), guided by skip connections that transfer high-resolution feature maps from corresponding layers of the encoder. These skip connections allow the model to retain spatial precision and recover fine image details lost in the down-sampling process.

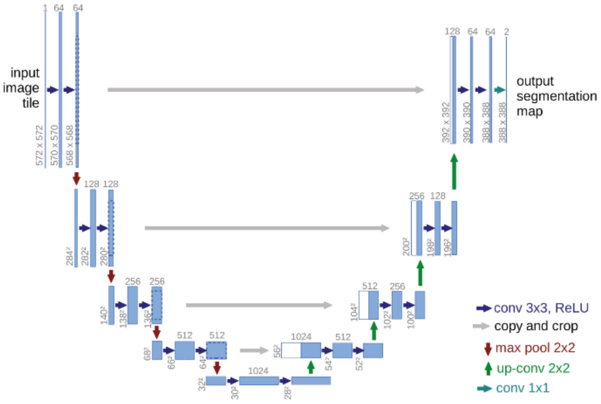

Figure 2. U-Net architecture (Ronneberger et al., 2015)

U-Net has proven effective in domains with limited training data, owing to its efficient use of data augmentation and the ability to learn from relatively small annotated datasets (Ronneberger et al., 2015). At the same time, the network is still relatively cheap in terms of required computational power and can be trained on consumer-grade hardware.

2.3 Data preprocessing

In order to train the network, we need to create an annotated dataset. Several annotators were employed to precisely vectorize the boundaries of the building stones. The orthomosaics were vectorized in a CAD software. While the vectorization of distinct boundaries was very precise, many areas suffered from hardly distinguishable spectral properties, and the boundary wasn't clear even to a human annotator (Figure 3). Thus, even the ground truth data can't be regarded as a hundred percent accurate.

To make the task even harder, the stones often contained holes filled with mortar, isolated from the bonding filler network (Figure 3). Instances of these holes are prone to be detected as false positives, whereas in our task we regard them as part of the stone. The tower mosaics contain clutter objects, such as windows, statues or sculptures obscuring the stones. As the trained network will be used on mosaics of the Bridge, which don't contain such clutter, we decided to mask most of unwanted objects out. Overall, 1,436.2 square meters were marked with polygons as 'stone' class, with 1,031.9 square meters remaining after masking. 149.3 square meters were marked as 'mortar'.

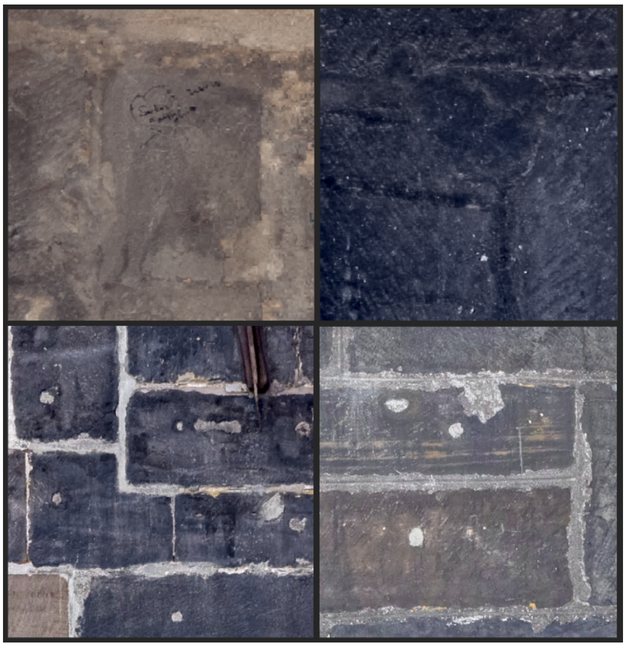

Figure 3. Difficult areas with hardly distinguishable stone/mortar boundaries (top row) and with unconnected pieces of mortar spectrally identical with bonding filler (bottom row)

To train a convolutional neural network, it is necessary to export the raster as square image chips of given size. The number needs to reflect available VRAM and desired context (receptive field) of the network. We opted for a size of 512 pixels. Due to limited training data, padding was set to half the size, exporting roughly four times more chips. Overall, 11,293 image chips were exported, of which 8,236 were used for training, 1,359 for validating and 1,698 for testing. The upper and lower halves of the southern tower side were set as validation and testing sets, respectively. The rest of the available data constitutes the training set. A training chip consists of the cropped RGB image and its equivalent ground truth binary raster. Examples of some training chips are in Figure 4.

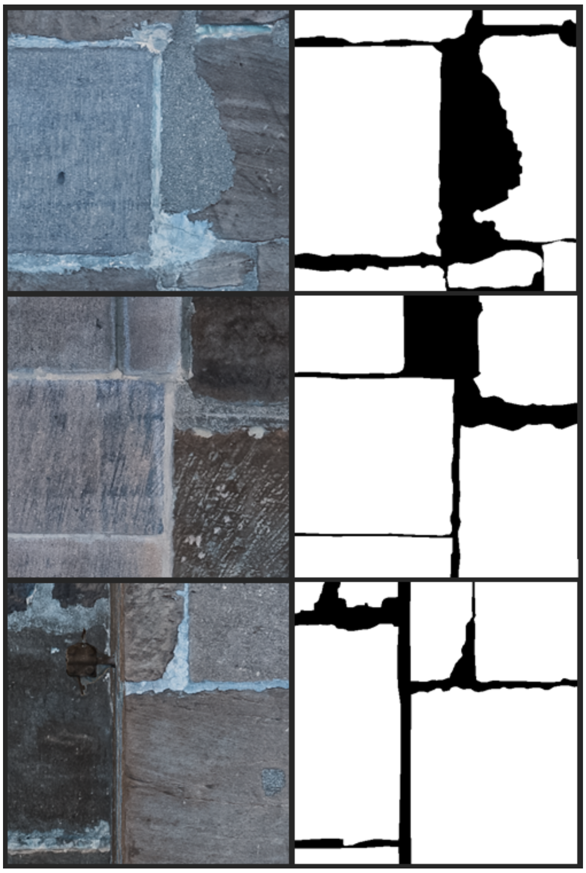

Figure 4. Examples of training chips (before spectral/geometric augmentation)

To improve generalizability, we employed data augmentation of the training set. First spectral, randomly modifying contrast with a coefficient between 0.8 and 1.25, and geometric, with random affine transformations with constraints. This should improve robustness of the network across diverse areas.

2.4 Training

We conducted several experiments, training three U-Net models. One supervised by binary cross-entropy (BCE) loss, one with dice loss, and one with the combination of both losses. BCE loss is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)], \quad (1)$$

where y_i is the ground truth label and $\hat{y}_i$ the predicted probability. This metric is summed and averaged for all N pixels. This metric punishes pixels with high false probabilities, and this gets optimized with gradient descent. The other loss function, Dice loss, is defined as:

$$L_{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN}, \quad (2)$$

where TP, FP and FN stand for true positives, false positives and false negatives, respectively.

Dice loss directly optimizes for segmentation quality and generates sharper edges, while BCE predictions are blurrier. The BCE and Dice loss can be combined to leverage their combined strengths (Wang et al., 2022); this can be done with addition weighted with a coefficient. For the third model, we train with both losses simply added together without any coefficient.

Given the significant class imbalance, it was necessary to account for it in the training, weighing the input data accordingly based on their representation ratio in the batch. For training, we used a batch size of 4, as this should ensure robust gradient descent, while being able to fit the samples into an available VRAM of 12 GB. We set the initial training rate to 1e-4 with decay to 1e-5. The model is trained for one hundred epochs, while being continuously evaluated on the validation split to avoid overfitting.

Model is then used on the test split. This results in a probability raster, which assigns each pixel a $\hat{y}_i$ probability. After inference, we reclassify the predicted raster at a chosen threshold to obtain binary representation. Then we compute the metrics: accuracy, recall and F1 score, for both the classes of 'stone' and 'mortar'. F1 score, being harmonic mean of precision and recall is frequently used as the benchmark metric and in binary segmentation task is identical to Dice loss.

To further widen the comparison and find the best performing network, we test various threshold values. This will modify the results, potentially skewing the metrics in favor of one of the two classes, but ultimately helping to find the equilibrium with the best training parameters.

3. Results

We trained three networks and evaluated nine various thresholds, resulting in twenty-seven combinations. While variable thresholding is mostly redundant for the model trained only with Dice loss, we include the results anyway for direct comparison. We report the F1 score for both the 'stone' and 'mortar' classes in Table 1, as they are both important for later evaluation. The precision and recall values are reported in Appendix.

F1 score	Loss function		
Threshold	BCE	Dice	Both
0.1	96.59 / 65.98	96.56 / 74.15	**96.73** / 73.47
0.2	96.34 / 68.21	96.54 / 74.19	96.67 / 74.16
0.3	96.08 / 69.63	96.52 / 74.20	96.63 / 74.44
0.4	95.81 / 70.74	96.51 / 74.20	96.58 / 74.57
0.5	95.53 / 71.69	96.50 / 74.21	**96.52 / 74.63**
0.6	95.22 / 72.53	96.49 / 74.21	96.46 / **74.66**
0.7	94.87 / 73.28	96.47 / 74.21	96.38 / 74.62
0.8	94.40 / 73.82	96.46 / 74.21	96.26 / 74.48
0.9	93.64 / 73.56	96.43 / 74.19	96.01 / 74.08

Table 1. F1 scores reported on the test split for each combination of training loss and thresholding parameter. The left value represents the 'stone' class result, the right one belongs to the 'mortar' class

The test raster contained some non-class clutter which was masked before calculating the metrics. Throughout the trainings, the validation loss generally decreased and remained plateaued. A segment of qualitative results is shown in Figure 5.

The network trained with the combined loss generally outperformed networks trained with individual losses across most threshold values. The best result was obtained with the network using combined loss and thresholding parameter of exactly 0.5. The average F1 score of both classes in this case is 85.58. The best results for individual classes are also recorded by the network trained with combined loss.

Maybe a bit surprisingly, the networks supervised solely with BCE loss always demonstrate inferior performance. This can be attributed to the better handled class imbalance in Dice loss. Even though with BCE we weigh the inputs according to their quantities in the training dataset, the imbalance is inherently better handled with Dice loss, which directly optimizes region overlaps. Some losses attempting to improve upon the existing losses were already developed, such as unified focal loss (Yeung et al., 2022).

Generally, it is not advisable to set the thresholding parameter too high, as F1 scores for both classes decrease in these cases. The higher the threshold, the more it favors the 'mortar' class, which can be useful for later polygon extraction of stones.

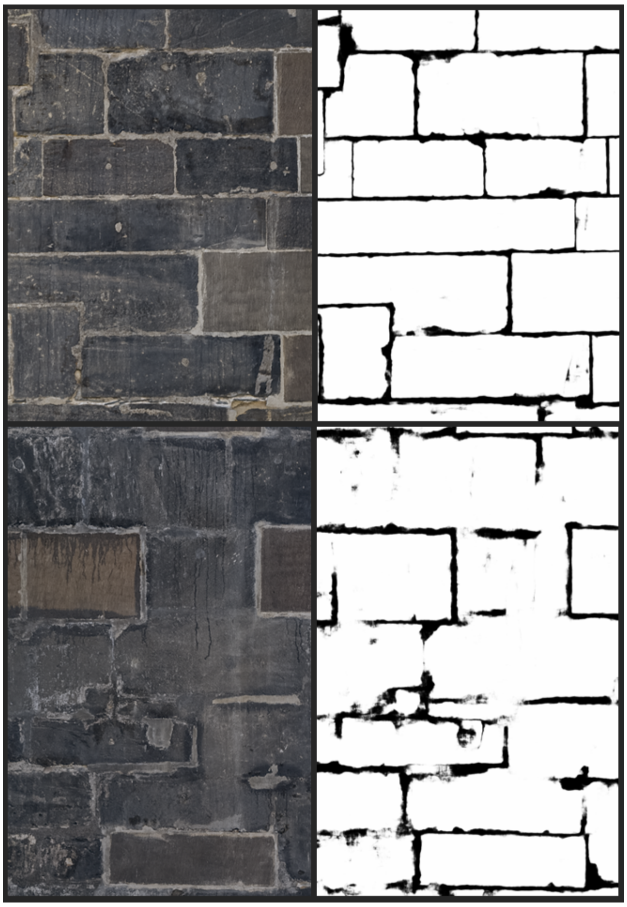

Figure 5. Results on the test split with BCE loss. Upper pair shows very good segmentation with distinguishable area, bottom picture shows slightly worse results on a more problematic area, which even human operators struggle with

4. Conclusion

This study measures pixel-level segmentation of the neural network and quantifies the results in this way. In this view, results are encouraging and will form the basis for further segmentation of the whole of Charles Bridge. Some fine-tuning may be required should the new dataset differ qualitatively.

An additional desired outcome would be to extract closed shapes of the stone outlines. While the results obtained with the best approach are often already sufficient for closed shape vectorization (Figure 6), in order to extract closed shapes of stones to precisely count them and consider topology constraints, additional work would have to be done. Weak mortar connections create gaps, which result in wrongly merged stones. To strengthen the mortar topology, fitting loss functions could be utilized, such as BALoss (Ngoc et al., 2021). To extract closed shapes, leveraging even weak edge connections which would vanish in thresholding, a watershed extraction can be of use (Meyer, 1994). This would require the use of BCE loss, whether combined or sole, as the results obtained solely with Dice loss are already almost binary, without any weak connections to recover.

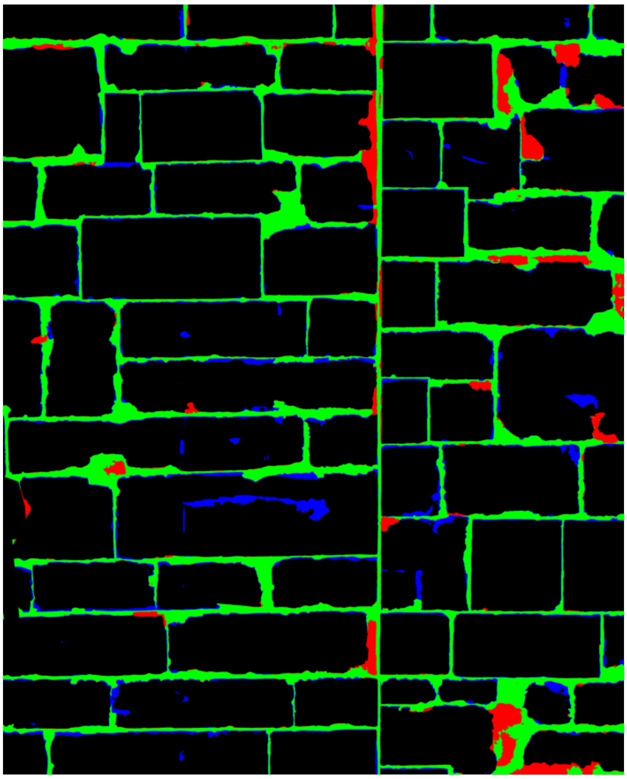

Figure 6. An image of an area of the test split produced with the best parameters (combined loss and 0.5 threshold). From the 'mortar' class perspective, the true positives are shown in green, false positives are blue and false negatives red.

Chen attempted to find the best performing combination of parameters for deep edge filtering and watershed extraction in the realm of scanned topographic maps (Chen et al., 2024). We believe the same approach would greatly benefit our use-case. The same can be said about the final metric COCO-PQ (Kirillov et al., 2019), which measures panoptic segmentation quality, accounting for both classification and intersection over union of individual polygons. This would directly translate to the necessity of extracting closed shapes.

To strengthen the topology of mortar connections, it could also be useful to dilate the cracks in the training data to widen the connections and make them more resilient to topological breaks. Given the relative sparsity of the cracks compared to the stones, the dilation parameter could be relatively high, even in order of centimeters. This would however need to be adjusted to avoid merging of neighboring cracks.

The results could be improved by paying more attention to detail when annotating this high-resolution dataset. As already mentioned, many borders between classes were barely recognizable, which requires skilled annotators and careful work. Given that the dataset wasn't initially annotated for the purpose of training a neural network and the annotations are slightly inconsistent depending on the annotator, there's definitely room to improve. The model could be then fine-tuned to a specific area of highly detailed ground truth data.

Acknowledgements

This research was funded by the Grant Agency of Czech Technical University in Prague, grant SGS25/046/OHK1/1T/11.

References

Chen, Y., Chazalon, J., Carlinet, E., Ngoc, M., Mallet, C., Perret, J., 2024: Automatic vectorization of historical maps: A benchmark. *PLOS ONE*, 19(2). doi.org/10.1371/journal.pone.0298217.

Dais, D., Bal, İ. E., Smyrou, E., Sarhosis, V., 2021: Automatic Crack Classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. *Automation in Construction*, 125, p. 103606. doi.org/10.1016/j.autcon.2021.103606.

Ibrahim, Y., Nagy, B., Benedek, C., 2019: CNN-based watershed marker extraction for brick segmentation in masonry walls. *Lecture Notes in Computer Science*, pp. 332–344. doi.org/10.1007/978-3-030-27202-9_30.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P., 2019: Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9404-9413). doi.org/10.48550/arXiv.1801.00868.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012: ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998: Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Long, J., Shelhamer, E., Darrell, T., 2015: Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Meyer, F., 1994: Topographic distance and watershed lines. *Signal Processing*, 38(1), pp. 113–125. doi.org/10.1016/0165-1684(94)90060-4.

Ngoc, M., Chen, Y., Boutry, N., Chazalon, J., Carlinet, E., Fabrizio, J., Mallet, C., Géraud, T., 2021: Introducing the Boundary-Aware loss for deep image segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, Virtual, 23–25 November 2021; Volume 17.

Pešek, O., Krisztian, L., Landa, M., Metz, M., Neteler, M., 2024a: Convolutional neural networks for road surface classification on aerial imagery. *PeerJ Computer Science*, 10. doi.org/10.7717/peerj-cs.2571.

Pešek, O., Brodský, L., Halounová, L., Landa, M., Bouček, T., 2024b: Convolutional Neural Networks for urban green areas semantic segmentation on sentinel-2 data. *Remote Sensing Applications: Society and Environment*, 36, p. 101238. doi.org/10.1016/j.rsase.2024.101238.

Ronneberger, O., Fischer, P., Brox, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*, pp. 234–241. doi.org/10.1007/978-3-319-24574-4.

Yeung, M., Sala, E., Schönlieb, C., Rundo, L., 2022: Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics* 95 (2022): 102026. doi.org/10.48550/arXiv.2102.04525.

Wang, H., Shen, Y., Liang, L., Yuan, Y., Yan, Y., Liu, G., 2022: River extraction from remote sensing images in cold and arid regions based on attention mechanism. *Wireless Communications and Mobile Computing*, 2022, pp. 1–18. doi.org/10.1155/2022/9410381.

Appendix

Here we provide precision and recall tables in a similar fashion as Table 1 reports F1 score. The experiments with the best F1 score are kept marked bold, and the best individual metrics for individual classes are written in italics. These are unsurprisingly located in extreme thresholding values.

Model trained with BCE loss generally achieves low precision on ‘mortar’ class, while recall is comparably better. The BCE loss generally detects more false positives.

We also provide the complete image of the test set along with the best result in Figure 6.

Precision	Loss function		
Threshold	BCE	Dice	Both
0.1	96.24 / 52.58	96.59 / 71.79	**96.14** / 66.76
0.2	96.91 / 56.10	96.64 / 72.21	96.44 / 69.25
0.3	97.26 / 58.61	96.68 / 72.49	96.61 / 70.69
0.4	94.49 / 60.82	96.74 / 72.71	96.75 / 71.78
0.5	97.67 / 62.94	96.74 / 72.92	**96.87 / 72.75**
0.6	97.83 / 65.16	96.76 / 73.12	96.99 / **72.75**
0.7	97.98 / 67.70	96.78 / 73.34	97.11 / 74.70
0.8	98.13 / 70.86	96.82 / 73.61	97.26 / 75.90
0.9	*98.32* / 75.65	96.87 / 74.00	97.49 / *77.84*

Table 2. Precision scores reported on the test split for each combination of training loss and thresholding parameter. The left value represents the ‘stone’ class result, the right one belongs to the ‘mortar’ class

Recall	Loss function		
Threshold	BCE	Dice	Both
0.1	96.94 / *88.55*	96.52 / 76.69	***97.32*** / 81.67
0.2	95.79 / 86.99	96.43 / 76.27	96.91 / 79.83
0.3	94.93 / 85.75	96.37 / 75.99	96.64 / 78.61
0.4	94.18 / 84.54	96.32 / 75.76	96.58 / 77.59
0.5	93.48 / 83.26	96.27 / 75.54	**96.52 / 76.63**
0.6	92.75 / 81.78	96.22 / 75.33	96.46 / **75.65**
0.7	91.95 / 79.87	96.16 / 75.10	95.66 / 74.54
0.8	90.94 / 77.03	96.10 / 74.82	95.28 / 73.11
0.9	89.38 / 71.56	95.99 / 74.37	94.59 / 70.67

Table 3. Recall scores reported on the test split for each combination of training loss and thresholding parameter. The left value represents the ‘stone’ class result, the right one belongs to the ‘mortar’ class.

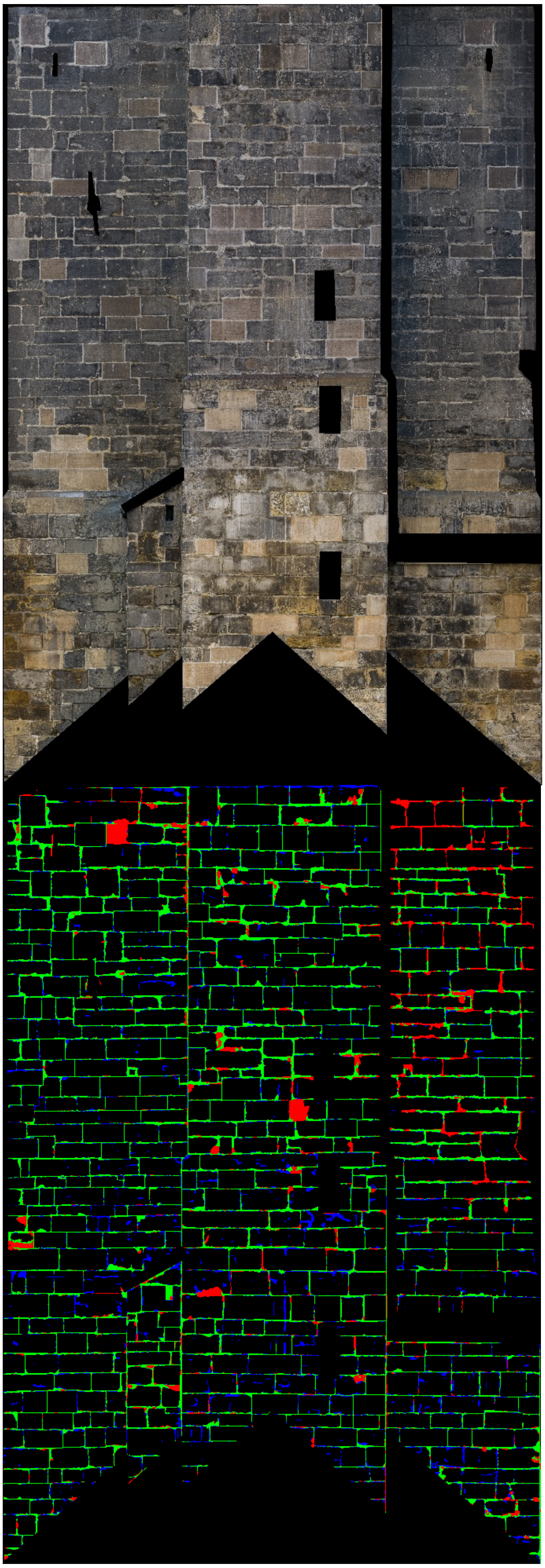

Figure 6. Complete area of the test split and the resulting map with the best parameters (combined loss and 0.5 threshold).

True positives are shown in green, false positives are blue and false negatives red. The rest are stones and some masked areas.