Deep Learning Based Methods on 2D Image and 3D Structure Inpainting

Theses of the *Ph.D.* Dissertation

Yahya Ibrahim

Scientific adviser:
Csaba Benedek, D.Sc.

Roska Tamás Doctoral School of Sciences and Technology
Pázmány Péter Catholic University

Budapest, 2023

1 Introduction and aim

Over the past ten years, deep learning networks have yielded exceptional success in a variety of computer vision tasks (segmentation, classification, object detections, etc.) across a wide range of fields, including manufacturing, autonomous driving, healthcare, archaeology, and civil engineering. However, currently available techniques are still far away from human-level performance, due to a variety of reasons, including the low resolution of the used sensors, occlusions, and the limited number of viewpoints used during scanning.
Occlusions may occur under several conditions in machine perception and computer vision applications, whether Lidars, cameras, or multiple sensor technologies are employed. Given that the camera/Lidar – from a fixed viewpoint – can only observe one side of the object being investigated, self-occlusions can occur even in situations with a single object. Moreover, occlusions may be found in a variety of complicated circumstances, for example if a part of an object is outside the field of view or when one or multiple objects occlude each other in the scene.
In object recognition applications, deep neural networks [8, 9] are still behind humans in the presence of occlusion [10] because human minds are adept at predicting the invisible components of the scene by observing the visible ones.
In real-life complex situations, occlusion is a major challenge for many sorts of image processing tasks. We can here mention applications like object detection in Advanced Driving Assistance Systems (ADAS), where it is challenging to accurately detect pedestrians or cars when they are partially occluded, especially in crowded scenes. In remote sensing images, the presence of clouds and their shadows can affect the quality of processing these images in several applications. Hence, to make good use of such images, assuming that the background of the occluded parts follows the same pattern as the visible parts of the image, inpainting algorithms can be trained in such cases to remove the occluded regions and fill them with the expected elements (such as road network, vegetation and built-in structures).

Similar concerns can be seen in the context of archaeology and civil engineering applications, where investigating masonry buildings necessitates the precise outlining of their structural components, which are frequently covered up by objects like decorative elements, wall sculptures, or vegetation.
Therefore, occlusion-aware networks have been thoroughly investigated in a variety of fields, including pedestrian detection [11], object tracking [12], face detection [13], and car detection [14].
In contrast to optical cameras, 3D sensors are less restricted by lighting and illumination, and can get accurate 3D geometric data from the scene. Hence, their usage in environmental perception is expanding rapidly.
However, considering both indoor and outdoor mobile mapping platforms, the scanning platforms equipped with 3D sensors cannot access specific locations to scan certain objects from all sides, resulting in incomplete point cloud representations. Therefore, point cloud completion – the estimation of an object's full shape from point sets that only partially describe its geometry – becomes a fundamental key challenge in numerous computer vision and robotic tasks, such as virtual reality (VR)/ augmented reality (AR) applications, and simultaneous localization and mapping (SLAM).
This thesis deals with two selected tasks from the problem family, in which automated occlusion or missing region detection is applied to either 2D images or 3D point clouds, and inpainting networks are then used to reconstruct the occluded/missing portions based on the observable information from the scene: The *first task* focuses on 2D images of walls that contain occluded or damaged regions. We can expect that the wall's structure follows a regular pattern, while the texture of the occluded wall part significantly differs from the wall's visible regions. In response, our new algorithm can automatically detect the irregular – possibly occluded – regions of the wall and predicts a structure that fits well into the regular mortar-brick pattern of the visible regions. The *second task* aims to tackle the completion of missing regions in 3D models caused by occlusion or by object parts covered form the sensor's viewpoints during the scanning process. The objective is to complete the shape and the

color of 3D point cloud models that represent partial shapes of the scanned items. In particular, we focus in the thesis on the task of completing outdoor objects captured by a car-mounted mobile laser scanning system while maintaining the high resolution of the data measurement.

2 New Scientific Results

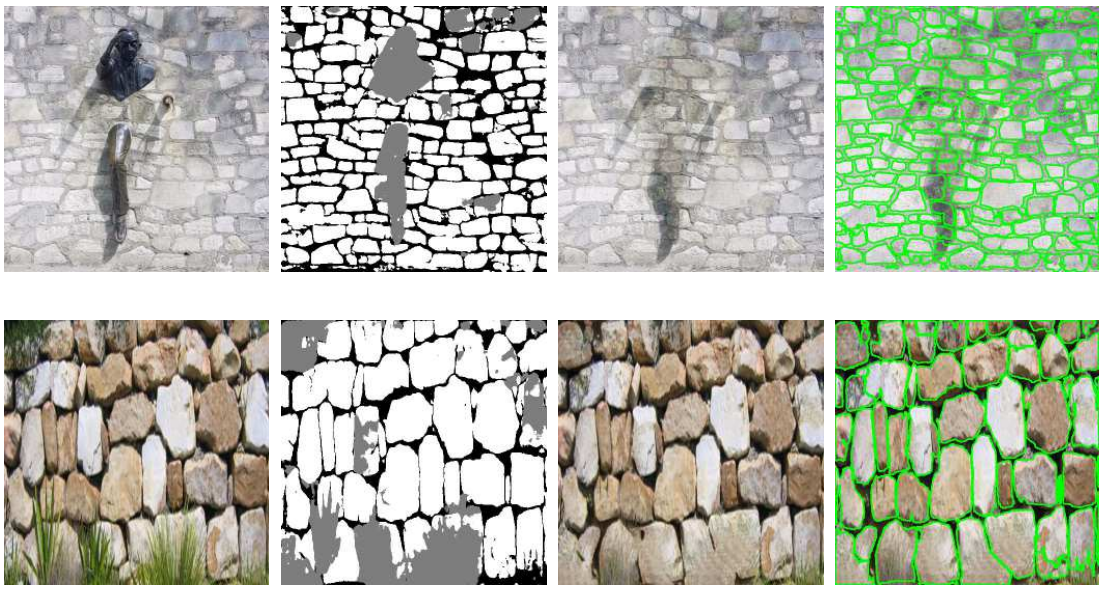
1. Thesis: I have proposed a novel masonry wall image analysis and virtual structure recovery technique. The introduced approach automatically segments the wall structure and inpaints possible wall segments in the observed occluded/damaged regions. I have experimentally demonstrated that the proposed technique outperforms recently published models with the same objectives in terms of various wall structure and visual color metrics.

Published in [1][3][4][6]

This thesis deals with three selected tasks: First robust wall segmentation algorithm is performed to separate various sorts of structural elements (stones, bricks, ashlar, etc.) from the mortar regions, and it can also detect the occluded and damaged wall regions. The second task is predicting and visualizing potential wall segments in the occluded/damaged wall regions. In the third task, a new style transfer technique is proposed between two wall images by filling or modifying the texture style of one wall based on another wall.
All of the aforementioned procedures are publicly available for testing on the following user-friendly website:
http://imgproc.mplab.sztaki.hu/masonrydemo

1.1. I have developed a technique for separating the bricks from the mortar in masonry wall images and obtaining accurate brick structures. The proposed method uses the U-Net-based delineation output as robust markers for the Watershed algorithm. I have shown the importance of employing the marker-based Watershed process rather than a basic connected component

analysis (CCA) approach. Moreover, I have experimentally demonstrated that the proposed technique surpasses the most recent wall segmentation techniques.



(a) Input (b) U-Net output (c) Inpainted im. (d) Seg. output

Figure 1: Results of our proposed method on two selected images: (a) Input: wall image occluded by irregular objects (b) Result of preliminary brick-mortar-occluded region separation (c) Generated inpainted image (d) Final segmentation result for the inpainted image [Thesis 1].

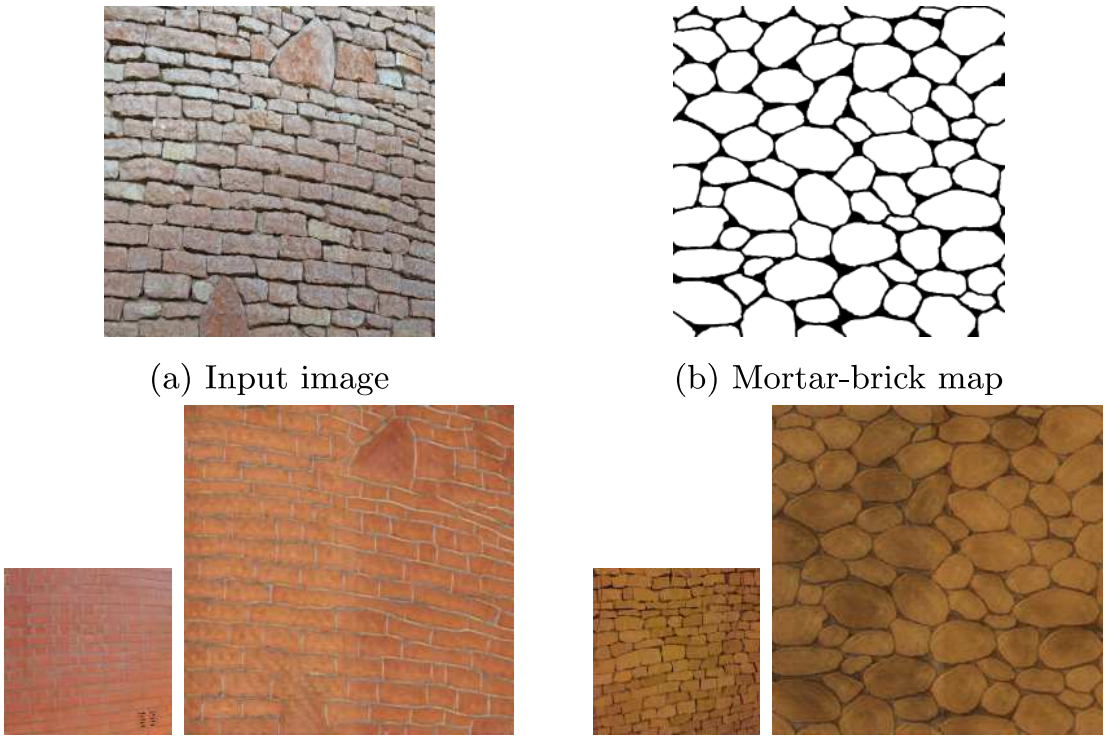
Several wall segmentation approaches exist in the literature [15, 16, 17]; however, the majority of them solely focus on the morphological analysis of quasi-periodic masonry walls, where the geometry of masonry courses follows horizontal rows, a condition that does not hold very often, particularly for ancient walls. I have designed a method that is adaptable to a large variety of wall structures, including both historic wall structures and modern building facades. I have shown that the proposed approach significantly surpasses earlier available solutions, and it is largely robust against various noise effects, different illumination conditions, changes in viewpoints, and varying

masonry types. In addition, I have shown, using both quantitative and qualitative experiments, that the proposed technique can identify a wide variety of possible occluding objects with high accuracy. To train and test the proposed network, I have created a new annotated dataset based on 532 different wall images. I have made the dataset publicly available in the following website:
http://mplab.sztaki.hu/geocomp/masonryWallAnalysis

> *1.2. I have proposed a novel blind masonry wall image inpainting technique, performing the automatic detection and virtual completion of occluded or damaged wall regions. The proposed method works in an end-to-end manner, starting with a segmentation step that detects the occluded regions and the wall structure in the visible areas, it then proceeds by two consecutive inpainting stages: wall feature completion, and color image completion. I have demonstrated the advantages of using domain specific semantic segmentation information (mortar-brick features) over low feature information (such as Canny-based edge information) as a preprocessing step of the inpainting stage for obtaining realistic wall structures in the occluded areas. Moreover, I have experimentally shown that the results of the proposed technique significantly suppress the state-of-the-art inpainting algorithms in terms of FID-score and human visual judgment for masonry wall image inpainting applications.*

The main goal of the proposed algorithm is to efficiently complete missing wall regions with realistic mortar-brick structure and color information. The proposed automatic method works in an end-to-end manner with a U-Net based model for detecting of the occluding segments, followed by two Generative Adversarial Networks (GANs) applied consecutively. The first GAN utilizes as input the segmentation results presented in Thesis 1.1, and completes the missing/broken brick and mortar segments yielding a complete wall structure mask with a connected mortar network. Thereafter, the second GAN estimates the RGB color values of the pixels in the predicted mortar-brick regions. I have demonstrated, through quantitative and qualitative comparisons, that the method is superior to

other state-of-the-art techniques for inpainting realistic wall structures and color textures in terms of FID-score and human visual judgment.



(a) Input image (b) Mortar-brick map

(c) Style (d) Our output (e) Style (f) Our output

Figure 2: Wall-to-wall style transfer samples, the first row shows the inputs: (a) wall image. (b) Mortar-brick map. The second row represents the style image and our output side by side [Thesis 1.3].

> *1.3. I have proposed a new technique for style transfer between two different walls. The algorithm replaces the coloring style of a wall image, with another wall's style, while maintaining the wall's original structural integrity. I have provided a comprehensive qualitative evaluation to demonstrate the benefits of our approach in wall-to-wall style transfer.*

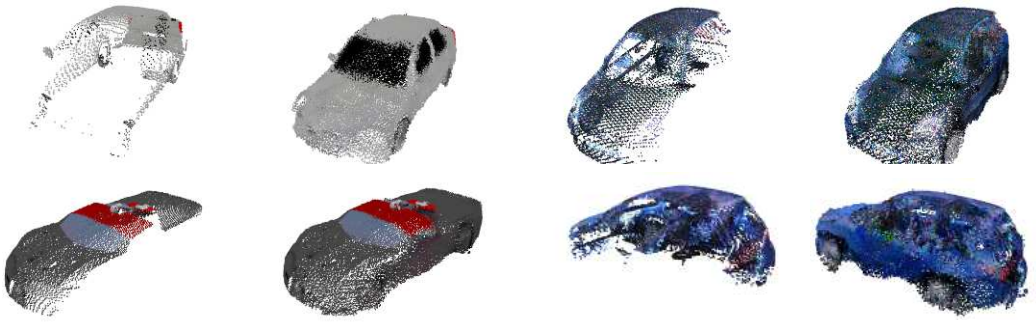
I have modified the second GAN proposed in Thesis 1.2 and adapted it for the wall image style transfer tasks in order to transfer the texture and color style from one image to another wall structure.

This approach requires two images as inputs: the first one is the content image which is a color wall image or a binary image for the wall structure, and the second image is the style image which is a different wall image. The goal is to create a new image that incorporates both the structure of the content image and the texture style of the style image. I have demonstrated, through a number of qualitative experiments, that the proposed method is significantly robust under different circumstances, in various applications.

2. Thesis: I have proposed a novel Multi-View Based Point Cloud Completion Network (MVPCC-Net) for completing colored 3D point clouds representing various incomplete object shapes. I have demonstrated by quantitative and qualitative experiments both on synthetic and real-world MLS data that the proposed method outperforms various state-of-the-art 3D point cloud completion techniques.

Published in [2][5]

Mobile laser scanning (MLS) is an emerging technology for generating extremely dense and very precise 3D point clouds for urban environments; yet, because the scanning vehicle can only operate on roads, many point clouds of field items have incomplete shapes. State-of-the-art point cloud completion methods [18, 19, 20, 21] have demonstrated success in estimating full geometric models of various object shapes. However, 3D point cloud models obtained by previous approaches represent only coarsely detailed object shapes, because the aforementioned methods are limited to providing outputs with a constant fixed number of points.
I have introduced a novel multi view-based approach for completing high-resolution 3D point clouds of partial object shapes obtained by MLS platforms, which is able to preserve the genuine high level of detailedness of the partial point cloud shapes. I have proved through quantitative and qualitative experiments on the provided dataset that our method outperforms state-of-the-art techniques in reconstructing the local fine geometric structures and predicting the overall shape of the objects.



(a) Input (b) Our result (c) Input (d) Our result

Figure 3: Results of the proposed method on a synthetic dataset (a-b) and real MLS point clouds acquired using a Riegl VMX Mobile Laser Scanner (c-d) [Thesis 2]

2.1. I have proposed a new approach for encoding the unstructured 3D point cloud data from several surrounding perspectives into a set of regular multi-channel 2D images comprising both geometry and color information. This representation permits the use of 2D Convolutional Neural Networks (CNNs) to fill in the missing structural and color information in the image domain, and afterward it enables the creation of a dense colored 3D point cloud representing the full object's shape. To experimentally validate the proposed approach, I have constructed a new database that consists of both synthetic and real MLS data and I have made it publicly available.

Previous point cloud completion methods employ various techniques to deal with the fundamental unorganized character of the point clouds, such as voxelization [18], intermediary 3D grids [19], or directly processing the point cloud [20, 21], with the PointNet encoder [22]. In order to apply the aforementioned data representations to MLS data, the input point cloud must be spatially downsampled, resulting in simplified object shape models with substantially lower point density and less geometric information than the original MLS measurements. To complete the missing regions of MLS measurement data and to keep its high resolution, I have proposed a new representation of the point cloud data in which the input point

cloud is represented as a collection of sparsely filled multi-channel images that capture from multiple angles and convey the geometry and color information. This representation facilitates the usage of 2D Convolutional Neural Networks (CNNs) and simplifies the fusion of color and geometry information. The proposed deep neural network is trained on this new data representation and the results are then reprojected to a 3D point cloud. I have demonstrated the effectiveness of this representation by presenting quantitative and qualitative results that verify the accuracy and density of the output point cloud.
For training and quantitative evaluation of the proposed method, I have provided a new point cloud dataset consisting of both synthetic point clouds of four different street object classes with accurate ground truth, and real MLS measurements of partially or fully scanned vehicles. I have demonstrated by quantitative and qualitative experiments both on synthetic and real-world MLS data that the proposed method is applicable and it outperforms various state-of-the-art techniques in terms of geometric shape accuracy, realistic RGB coloring, and preserving high resolution.

> *2.2. I have proposed a late fusion-based technique for fusing the view-level features generated from several perspectives around the object. The proposed method provides a robust global feature for transmitting shared characteristics between distinct viewpoints. I have demonstrated by quantitative and qualitative results that the proposed approach outperforms the early fusion baseline technique both for pure geometric data samples (XYZ), and for colored point clouds (XYZRGB).*

Existing multi-view based methods use an early fusion technique, where all projected images from all views are concatenated into a single input data for the encoder, while the decoder generates all output views in one step. On the contrary, my proposed method uses a late fusion strategy, whereby a shared encoder is used to build a view-level feature representation for each view separately, and then a feature fusion network component is used to create a global feature from the view-level features. The shared decoder

receives the global feature and the view-level features in a certain order to generate the view-level output images that constitute the whole 3D point cloud.
I have quantitatively and qualitatively demonstrated that adding RGB information to the early fusion based method reduces the geometric accuracy notably more than the late fusion based method, where we notice just a minor decrease in the accuracy of the predicted object shape.

3 Application of the Results

All the developed algorithms can be used by various up-to-date or future computer vision systems. The first thesis can be applied to various image-based documentation and survey applications in archeology, architecture, or civil engineering, where brick segmentation is considered as an important initial step in the analysis of masonry wall images. Image-based analysis of man-built structures is considered as a core step of many applications, such as stability analysis in civil engineering, condition estimation and damage detection of buildings in architecture, digital documentation in archeology or maintenance and restoration in cultural heritage preservation.
The 3D point cloud completion method presented in the second thesis is useful in a wide variety of computer vision and robotic activities where full scene representation is needed for improved scene visualization, such as VR/AR applications, self-driving, and surveillance applications.

4 Datasets and Implementation Details

For training and evaluation of the proposed method in the first thesis, I created a new annotated dataset containing images of masonry walls from various locations, including both ancient walls and facades of new-fashioned modern buildings. The dataset is based on 532 different wall images of size 512×512, which are divided among the training set (310 images), and three test sets (222 images).

For the second thesis, I trained and quantitatively evaluated the proposed model using synthetic data, which consists of pairs of partial and complete point cloud models of four street object shapes derived from ShapeNet [23]. In addition, I extensively tested our trained shape completion network on real-world (incomplete) Mobile Laser Scanning (MLS) measurements. The synthetic dataset contains in total 4918 distinct models, of which 4580 objects are used to train our model and 338 ones are used for evaluation. The real MLS data collection consists of 424 object samples in total, 370 point clouds represent partial vehicle shapes, 54 samples depict almost entire vehicle shapes. All the proposed datasets have been made available to the academic community.
The main platform for point cloud handling and processing was implemented in Python3 and Open3D while the neural network models were implemented and trained in Python3 with Pytorch/ Keras frameworks. The hardware set up for training contains two Nvidia Geforce RTX 3060 Ti GPU with 16 GB device memory and 64 GB main memory.

5 Acknowledgements

First of all, I owe my deepest gratitude to my supervisor *Prof. Csaba Benedek* for his continuous support, motivation and patience during my Ph.D. study and work.
I would like to express my sincere gratitude to all the Pazmany Peter Catholic University (PPCU) members, thanks to *Prof. Péter Szolgay* who provided me the opportunity to study here, and *Prof. Gábor Szederkényi* for providing me with all the necessary facilities for the research, thanks to *Mrs. Vida Tivadarné* for her continuous assistance with administrative concerns. I deeply thank *Prof. Árpád Csurgay* for the motivation and encouragement. I would like to thank all colleagues at PPCU with whom I spent these past few years *Sam Khozama*, *Ward Fadel*, *Jalal Alafadi*, *Nawar Al.hemeary*. Special thanks go to *Marwan Hassan* for his tremendous support since the first day of my international journey.

I thank the reviewers of my thesis for their work and valuable comments.
The support of the Institute for Computer Science and Control (SZTAKI) is also gratefully acknowledged for employing me to proceed with my research. My sincere thanks go to *Prof. Támas Szirányi* for offering me the job. I thank my closest colleagues from SZTAKI, Machine Perception Research Laboratory for their help: *Lóránt Kovács*, *Balázs Nagy*, *Örkény Zováthi*, *Zsolt Jankó*, *Marcell Kégl*.
For further financial support, my research work was supported by various projects and grants: by Stipendium Hungaricum scholarship program; by the National Research, Development and Innovation (NRDI) Office of Hungary within the frameworks of the National Laboratory for Autonomous Systems (NLAS), and the Artificial Intelligence National Laboratory (MILAB); by the NRDI Fund (OTKA) grants K-120233, KH-125681 and K-143274; by the Hungarian R&D grants EFOP-3.6.2-16-2017-00013, and TKP2021-NVA-01.
To all my friends, thank you for your kindness and assistance throughout my times of need. Your friendship strengthens me to face difficulties. I cannot list all the names here, but you are always on my mind.
Last but not least, I would like to express my sincere gratitude To: the kind loving simple people of my small village "*Balghounes*", my big family *"Ibrahim"* and *"Omran"*. To my first mentors and forever guardian angels my grandfathers *Yahya* and *Mohammed*. To my grandmothers *Radiah*, *Manera*, and *Hajar* for their neverending blessings. To my godfather *Nasr* for his kind heart. To my father *Ahmed*, whose wise words have guided me through the hard times. To my mother *Zainab* and her unconditional love that I will never be able to pay back even if I live for a thousand years. To my brothers: *Mohammed* whose feelings were always beside me, *Yousef* who has always been here for me whenever and wherever needed, and to the hope of our small family *Sharaf* who has a bright future ahead of him. To *Ghaith* my brother from another mother.
To those whom I may have not mentioned by name but who supported me directly or indirectly in accomplishing my Ph.D. research.

6 Publications

6.1 The Author's Journal Publications

[1] **Y. Ibrahim**, B. Nagy, and C. Benedek, "Deep Learning-Based Masonry Wall Image Analysis," *Remote Sensing*, vol. 12, no. 23, 2020. IF=4.8.

[2] **Y. Ibrahim** and C. Benedek, "MVPCC-Net: Multi-View Based Point Cloud Completion Network for MLS Data," *Image and Vision Computing*, vol. 134, no. article 104675, 2023. IF=3.860.

6.2 The Author's International Conference Publications

[3] **Y. Ibrahim**, B. Nagy, and C. Benedek, "CNN-Based Watershed Marker Extraction for Brick Segmentation in Masonry Walls," in *16th International Conference on Image Analysis and Recognition*, (Waterloo, Canada), pp. 332–344, Springer International Publishing, 2019.

[4] **Y. Ibrahim**, B. Nagy, and C. Benedek, "A GAN-based Blind Inpainting Method for Masonry Wall Images," in *25th International Conference on Pattern Recognition (ICPR)*, (Milan, Italy (Virtual)), pp. 3178–3185, 2021.

[5] **Y. Ibrahim**, B. Nagy, and C. Benedek, "Multi-view Based 3D Point Cloud Completion Algorithm for Vehicles," in *26th International Conference on Pattern Recognition (ICPR)*, (Montreal, Canada), pp. 2121–2127, 2022.

[6] **Y. Ibrahim**, P. Szulovszky, and C. Benedek, "Masonry Structure Analysis, Completion and Style Transfer Using a Deep Neural Network," in *International Workshop on Pattern Recognition for Cultural Heritage, to appear in the ICPR 2022 Workshop Proceedings, Lecture Notes in Computer Science*, (Montreal, Canada), Springer, August 21, 2022.

[7] W. Fadel, C. Kollod, M. Wahdow, **Y. Ibrahim**, and I. Ulbert, "Multi-Class Classification of Motor Imagery EEG Signals Using Image-Based Deep Recurrent Convolutional Neural Network," in *8th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4, 2020.

6.3 Selected Publications Connected to the Dissertation

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 770–778, IEEE Computer Society, jun 2016.

[9] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, 2015.

[10] H. Zhu, P. Tang, and A. L. Yuille, "Robustness of Object Recognition under Extreme Occlusion in Humans and Computational Models," in *Annual Meeting of the Cognitive Science Society*, 2019.

[11] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd," in *European Conference on Computer Vision (ECCV)* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 657–674, Springer International Publishing, 2018.

[12] Y. Liu, X.-Y. Jing, J. Nie, H. Gao, J. Liu, and G.-P. Jiang, "Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 664–677, 2019.

[13] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recog-

nition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] N. D. Reddy, M. Vo, and S. G. Narasimhan, "Occlusion-net: 2d/3d occluded keypoint localization using graph networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7318–7327, 2019.

[15] B. Riveiro, B. Conde, H. Gonzalez, P. Arias, and J. CaamaĂao, "AUTOMATIC CREATION OF STRUCTURAL MODELS FROM POINT CLOUD DATA: THE CASE OF MASONRY STRUCTURES," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-3/W5, pp. 3–9, 8 2015.

[16] N. Oses, F. Dornaika, and A. Moujahid, "Image-Based Delineation and Classification of Built Heritage Masonry," *Remote Sensing*, vol. 6, p. 1863â1889, Feb 2014.

[17] M. Hemmleb, F. Weritz A, A. Schiemenz B, A. Grote C, and C. Maierhofer, "Multi-spectral data acquisition and processing techniques for damage detection on building surfaces," in *ISPRS Commission V Symposium*, pp. 1–6, 1 2006.

[18] A. Dai, C. R. Qi, and M. Nießner, "Shape Completion using 3D-Encoder-Predictor CNNs and Shape Synthesis," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] H. Xie, H. Yao, S. Zhou, J. Mao, S. Zhang, and W. Sun, "GRNet: Gridding Residual Network for Dense Point Cloud Completion," in *European Conference on Computer Vision (ECCV)*, pp. 365–381, Springer International Publishing, 2020.

[20] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "PCN: Point Completion Network," in *International Conference on 3D Vision (3DV)*, pp. 728–737, 2018.

[21] L. P. Tchapmi, V. Kosaraju, H. Rezatofighi, I. Reid, and S. Savarese, "TopNet: Structural Point Cloud Decoder,"

in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 383–392, 2019.

[22] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2017.

[23] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.