Enhancing 3D Model for Urban Area with Neural Representations

Sitong Li
Student Number: #5683688

1st supervisor: Nail Ibrahimli
2nd supervisor: Ken Arroyo Ohori

July 3, 2024

Type	MedAE[m]	Median[m]	MAE[m]	RMSE[m]	NMAD[m]	Pixels
Overall	1.5199	1.0424	3.5156	6.9758	2.8159	17209506
Building	0.7912	-0.5777	2.5448	7.3884	2.2274	3529615
Forest	3.5613	3.3565	6.1072	9.4555	3.5732	718374
Terrain	2.1896	2.0857	4.2174	6.6949	1.9181	6862569
Terrain_no_vegetation	2.0066	1.8859	3.8383	6.3075	1.7083	4937153

Table 1: Evaluation of statistical metrics for different types of land cover

Table 2: Evaluation of polygons with internal points removed for selected IDs

ID	Area [m^2]	MAE [m]	RMSE [m]	MedAE [m]	Median [m]	NMAD [m]	Pixels
3	40.386	0.146	0.228	0.061	0.000	0.091	1100
18	76.187	1.295	2.119	0.416	0.000	0.617	1950
19	46.613	0.906	1.940	0.112	0.000	0.167	1302

1 Introduction

The accuracy and comprehensiveness of 3D city models have become more and more important in monitoring, sustainability evaluation, disaster management and urban planning (Toschi et al., 2017; Skondras et al., 2022). However, the creation of accurate and detailed 3D models is fraught with challenges. Traditional methods, such as photogrammetry and LiDAR (Light Detection and Ranging), often struggle with issues such as time-consuming processes, occlusion and limitations in capturing fine details of urban landscapes (Zhang and Lin, 2017), (Habib et al., 2005), (Mozas-Calvache et al., 2023). Furthermore, these methods typically produce discrete models that do not represent all the information contained in the original object and restrict the ability to reconstruct small structures.

The potential of neural representation in the field of Geomatics is both promising and yet to be fully explored. Its continuous and generative nature holds the promise of enhancing the quality of 3D models like point cloud and DSM (Digital Surface Model), ensuring there are no blank areas and providing more detailed information. Implicit neural representation involves employing an encoder to learn features and patterns from discrete data, subsequently generating a continuous function that accurately represents the target object (Ran et al., 2022), (Dai and Nießner, 2022). It has been used in many fields like 3D shape and scene reconstruction, signal processing, and fluid dynamics visualization due to its ability to encode complex, high-dimensional data into continuous, lower-dimensional forms.

Convolutional occupancy networks, a novel approach in this domain, leverage the capabilities of convolutional neural networks (CNNs) to process and interpret complex spatial data efficiently. By utilizing these networks, it is possible to create more accurate and detailed models that better represent the intricacies of urban landscapes using its generative feature (Peng et al., 2020). DeepSDF, or Deep Signed Distance Function, is an innovative approach that utilizes a neural network to learn continuous signed distance functions for complex geometries. This technique offers a highly efficient and detailed method for 3D modeling, surpassing traditional mesh or voxel-based approaches in terms of flexibility and scalability (Park et al., 2019).

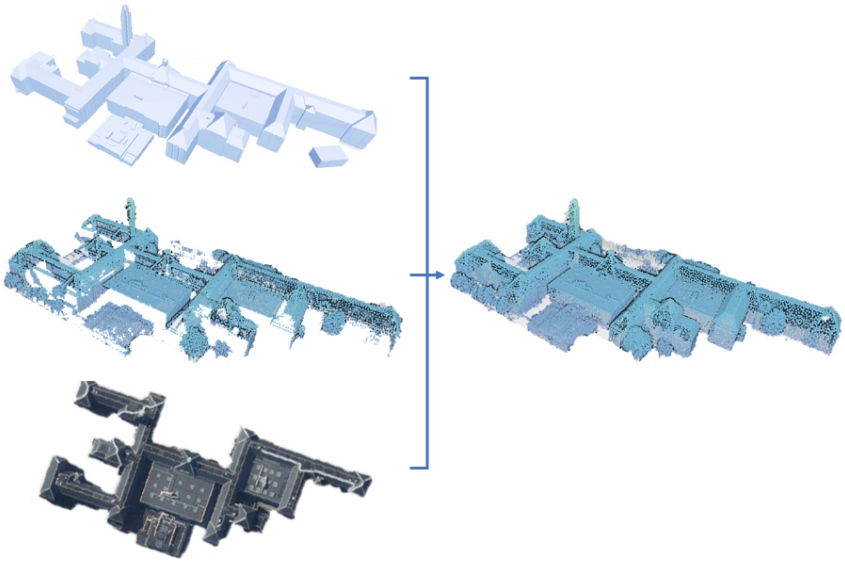


Figure 1: Demonstration of Implicit Neural Representation

This thsis aims to explore the enhancement of 3D models for urban area using implicit neural representation combined with reference data. In this project, the raw point cloud collected by laser scanner and reference data like 3D city model (3D Basisregistratie Adressen en Gebouwen), orthophoto, and BGT (Basisregistratie Grootschalige Topografie) will be encoded to a continuous occupancy field that represents the scene geometry. Location-dependent latent codes are learned to modulate the occupancy probability from the sampled point cloud from the 3D city model and orthophoto. Using a proper decoder, the probability of existence for any point in the 3D space can be estimated. By doing sampling in the space with the help of masks for water, plants, buildings, and terrain, we can get a 3D model (point cloud, DSM, mesh, etc.) with higher quality from the neural network shown in Figure 1.

The project will explore ways to integrate multi-modal data sources effectively with implicit neural representation, enhance the completeness and detail of models and contribute to the field of geomatics by providing a tool that can support 3D real scene reconstruction.

2 Related work

This section explores the relevant literature, underscoring developments in traditional methods, the advent of neural network-based approaches, and the specific advancements in convolutional occupancy networks and their application to urban modeling.

Traditional Methods in 3D City Modeling: Historically, the construction of 3D city models has relied heavily on techniques like photogrammetry and LiDAR. Ground-breaking works by (Ackermann, 1999) and (Vosselman and Maas, 2010) laid the foundation for using aerial imagery and laser scanning for 3D reconstruction. These methods have been used for decades to achieve relatively high accuracy in urban modeling. The processing after data acquisition, such as data fusion and noise cleaning, is very costly. Moreover, the final reconstruction results are often unsatisfactory, frequently featuring issues like

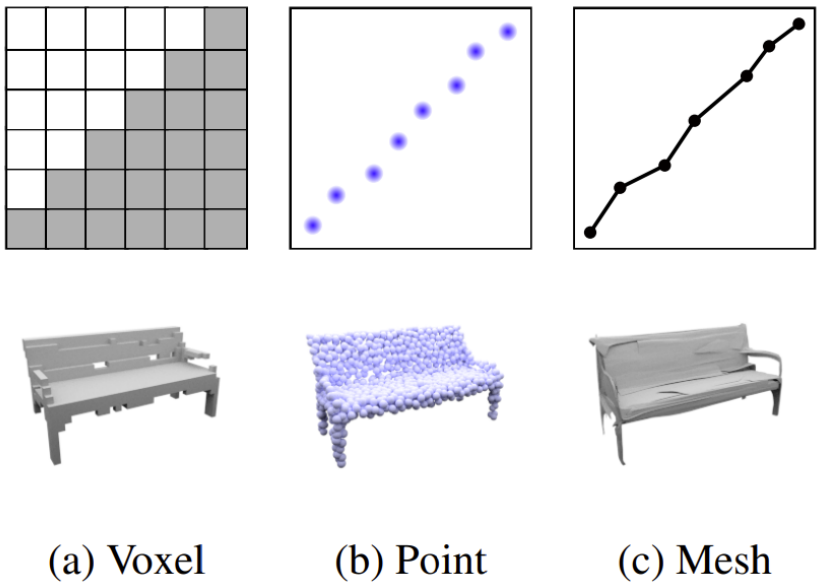


Figure 2: Three traditional representation methods (Mescheder et al., 2018)

low point cloud density due to low reflectance of special planes, data voids due to occlusions from limited scanning angles, and noise that the algorithms cannot recognize. The three most common ways for 3D data representation are voxel, point cloud and mesh as shown in Figure 2.

Voxel representation leads to complex implementations and existing data-adaptive algorithms are still limited to relatively small voxel grids (128^3 or 256^3). What's more, this approach can cause Manhattan World bias because its grid-based structure aligns with a world composed of orthogonal planes and straight lines (Maturana and Scherer, 2015), (Zishu et al., 2020) while the axis of large-scale real scenes is often affected by the curvature of the earth.

Point clouds, while capturing raw 3D data, do not contain the connectivity and topological structures. Consequently, additional postprocessing is needed to derive 3D geometries from the model, as indicated in studies by (Li and Baciu, 2021b), (Zhao et al., 2019) and (He et al., 2019). This requirement limits their direct applicability in computer graphics-related applications, such as shadow estimation. Furthermore, point clouds are often challenged by a limited number of points, which leads to a loss of fine detail. Additionally, point clouds struggle with effectively conveying the global shape of objects. This issue arises because point clouds represent 3D objects with no explicit information about the surface or volume these points belong to, which is crucial in applications requiring a holistic understanding of an object's form (Li and Baciu, 2021a).

Mesh representations are frequently used in discriminative 3D classification or segmentation tasks because they provide a structured and efficient way to represent 3D shapes and surfaces. However, meshes always require class-specific templates (unique structural features) for each class for accurate modeling, which makes it difficult to adapt to objects with significantly different structures. Additionally, deforming these templates to fit specific instances can lead to issues like self-intersection (Maturana and Scherer, 2015), (Reddy et al., 2022).

Advances in Neural Networks for 3D Reconstruction: With the development in deep

learning, neural networks have increasingly been applied to 3D reconstruction tasks. The work of (Qi et al., 2017) on PointNet demonstrated a novel approach to segment and classify point clouds using neural networks, offering a significant boost in processing efficiency and model detail. This shift marked a departure from traditional geometry-based methods, paving the way for more sophisticated model generation techniques. Recent studies have also focused on enhancing the level of detail in 3D models using neural representations. The works of (Park et al., 2019) on DeepSDF and (Chen and Zhang, 2019) on implicit field representations for neural scene synthesis have demonstrated significant advancements in capturing intricate details, which is particularly relevant for urban landscapes where fine geometric features like building façades are crucial.

Convolutional Occupancy Networks in Urban Modeling: The introduction of convolutional occupancy networks represents a paradigm shift in 3D modeling. (Peng et al., 2020) introduces the concept of learning implicit functions for 3D surface reconstruction using deep learning. Their approach features a fully convolutional encoder, which ensures that the implicit representation maintains the translation equivariance characteristic of convolutions. This, in turn, enables the reconstruction on a large scale. Also, this method shows promise in handling complex geometries with higher accuracy and at a lower computational cost. One of the contemporary challenges in 3D city modeling is the effective integration of multi-modal data sources. Works by (Boulch et al., 2017) and (Stucker et al., 2022) have explored the fusion of aerial imagery with point cloud data, demonstrating improvements in both the accuracy and richness of urban models. These studies highlight the potential of using deep learning algorithms to diffuse diverse data types, such as satellite images, LiDAR data, and ground-level photographs, into a cohesive model.

Despite these advancements, challenges persist, particularly in the integration of diverse data sources, ensuring real-time updates, and maintaining high levels of detail. Future research directions suggest more efficient neural architectures, better data fusion techniques, and more robust frameworks for scalability and real-time updating.

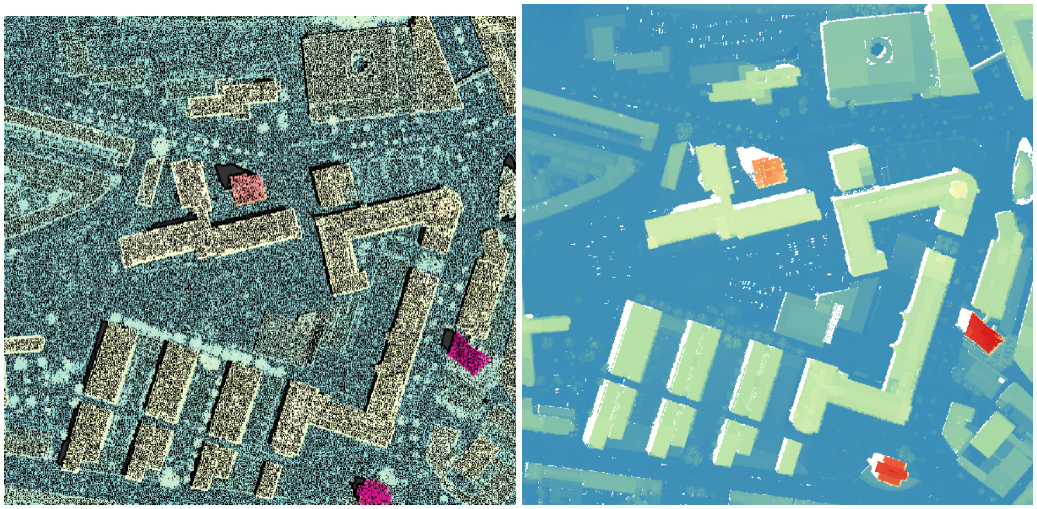
In conclusion, the related work in 3D city modeling illustrates a progressive shift from traditional photogrammetry and LiDAR-based methods to advanced neural network-based approaches. Convolutional occupancy networks and signed distance functions emerge as promising tools in addressing the current limitations of 3D city models, offering a pathway toward more detailed and accurate urban models. This research will build upon these foundations, seeking to further explore and enhance the integration of these technologies in the realm of city modeling.

3 Research questions

Main Question: How to improve the geometric performance of generated 3D models (point clouds, DSM) using implicit neural representation?

1. For point cloud: During data acquisition, obstacles can block the sensor's view of parts of the object, causing gaps and missing parts in the raw point cloud data. However, suppose we model the data as a continuous field of occupancy or signed distance. In that case, it can not only reconstruct clean shapes from noisy and partial point clouds but also generate parts of the object beyond the regions directly captured during the data acquisition phase.

2. For DSM: The generation of a Digital Surface Model (DSM) using implicit neural representations involves converting an implicit function into an explicit surface model through a hierarchical refinement process. This method ensures that the final DSM contains no blank areas, as shown in Figure 3, that typically arise from gaps in the original point cloud data. The generative nature of implicit neural representation effectively compensates for these missing parts, ensuring a complete and continuous surface representation in the final DSM.



(a) AHN4 Point cloud with gaps (b) Corresponding DSM

Figure 3: Gaps in raw point cloud data can cause gaps in DSM result

Subquestion 1: To what extent can the incorporation of reference data and implicit neural representations improve the quality of 3D city models?

In the raw point cloud dataset, there is no connectivity or topology information, which means they can't define the surface or shape of objects and can miss high-frequency details like sharp corners and edges. However, these features are in most cases, visible in aerial or satellite images. In this project, a second latent embedding for image is used as additional input to guide the occupancy prediction. Mask for building, water and plant will also be used for higher accuracy since the point density differs in these areas. What's more, the resolution of the generated 3D model can be infinite theoretically because the result of implicit representation will be a continuous occupancy field and the occupancy status can be estimated at any given coordinate. Its resolution will be bounded only by neuron number and training data quality.

Subquestion 2: What are the limitations and potential biases in current convolutional occupancy network-based methods for 3D city modeling, and how can they be addressed?

1. Scalability: Convolutional occupancy networks currently encounter limitations when attempting to scale up for modeling entire cities, as existing experiments are confined to areas measured in square kilometers. However, this challenge can potentially be overcome by utilizing implicit representation for estimations in regions beyond the trained area. Provided the training dataset is sufficiently extensive and diverse, it opens up the potential for efficiently refining point cloud data at a city-wide scale.

2. Absolute accuracy: For point cloud, DSM, or other geo-spatial datasets generated from implicit neural representation, during the training and testing process, there

is no promise of absolute accuracy since all features are learned in a black box. The choice of tolerance during data generation needs to be very careful and evaluations like using ground control points, comparing the DSM with a reference model known for its high accuracy and extracting roof ridges from point cloud for comparison. Also, enhancing the quality of training dataset can also be helpful.

3. Real application: The algorithm has not been tested in areas with various land types and different point densities. The estimation result can be affected if no proper weight is given to input points. The masks are used to solve this.

4 Methodology

Convolutional Occupancy Network and ImpliCity network will be used as the guideline for the novel reconstruction method.

4.1 Data preparation

The required input datasets are raw point cloud, point cloud sampled from LOD2.2 city model, orthophoto covering the train and test area, previous DSM as reference data and mask for building, plant and water area.

Raw point cloud is from AHN3, clipped using pdal. LOD2.2 city model is available as open data and can be downloaded from 3DBAG.nl. Orthophoto is also available on PDOK. Masks indicate if there exists a certain object, and can be generated from BGT dataset. What worth mention is that, the building polygon in BGT only exists after it's been built and is measured at ground level. So extracting the roof surface from city model as building mask is more reasonable.

4.2 Implicity Neural Representation Network

The main principle is to encode the real 3D geometric scene into a continuous occupancy field:

$$f_\theta(x, \psi(P, x), \xi(I, x), ...) \rightarrow \hat{o} \in [0, 1]$$

In which $\psi(P, x) \in \mathbb{R}^d$ and $\xi(I, x) \in \mathbb{R}^d$ are latent embeddings for input point cloud and image. More embeddings can be added to increase the representation power of neurons. The use of masks or land-type datasets is still to be explored. This field is parameterized by a neural decoder network. The process is further guided and constrained by a local latent code, which is extracted using a neural encoder network.

The current network architecture employs a fully convolutional encoder, which ensures that the implicit representation adheres to translation equivariance. Additionally, it incorporates local latent codes that align pixel-wise with the supervising image to obtain sharp surface edges that correspond with the image gradients.

During the training phase of the deep learning network, query points are randomly sampled within a specific volume of interest and near the actual surface of the object (3D city model). The training process is guided and supervised using a binary cross-entropy loss function, denoted as L. This loss function plays a critical role in the learning process by comparing the predicted occupancies, represented as $\hat{o}$, with the true occupancies,

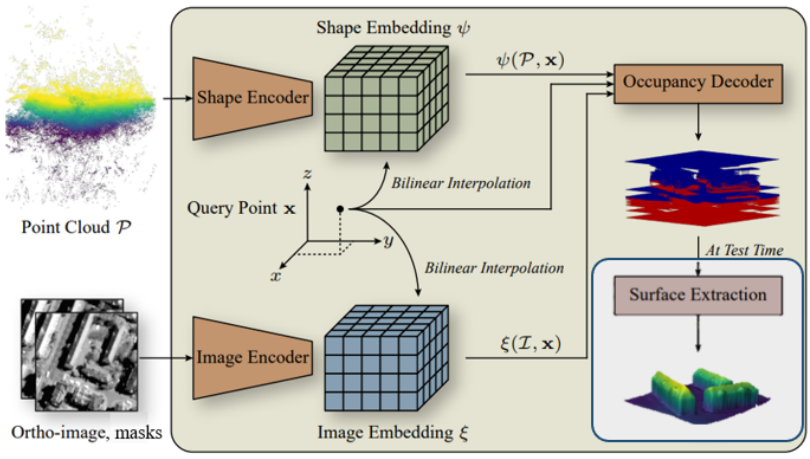


Figure 4: Network architecture ((Stucker et al., 2022))

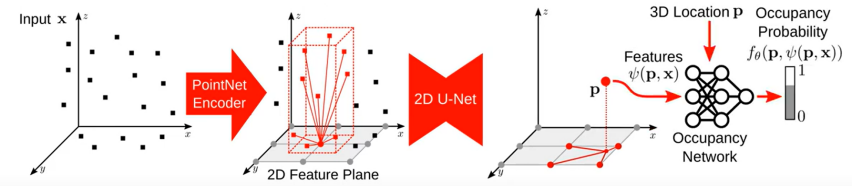


Figure 5: Convolutional occupancy network ((Peng et al., 2020))

denoted as o, at these sampled points. The goal of this process is to minimize the discrepancy between the predicted and actual occupancies, thereby enhancing the accuracy and effectiveness of the network in modeling the given 3D structures.

$$L(\hat{o}, o) = \sum_i (o_i \cdot log(\hat{o_i}) + (1 - o_i) \cdot log(1 - \hat{o_i}))$$

The architecture of the network will based on Figure 4: The fundamental operating principles of the encoder and decoder for the point cloud are illustrated inFigure 5: The fundamental operating principles of the encoder and decoder for the image are illustrated in Figure 6

4.3 Expecting Result

Generate 3D models like point cloud and DSM with no gap and with fine detail, especially on buildings.

4.4 Evaluation

The application is designed for real-world scene. To assess the precision of the generated point clouds and Digital Surface Models (DSM), it is necessary to develop specific algorithms that can measure both the absolute and relative accuracy of these models.

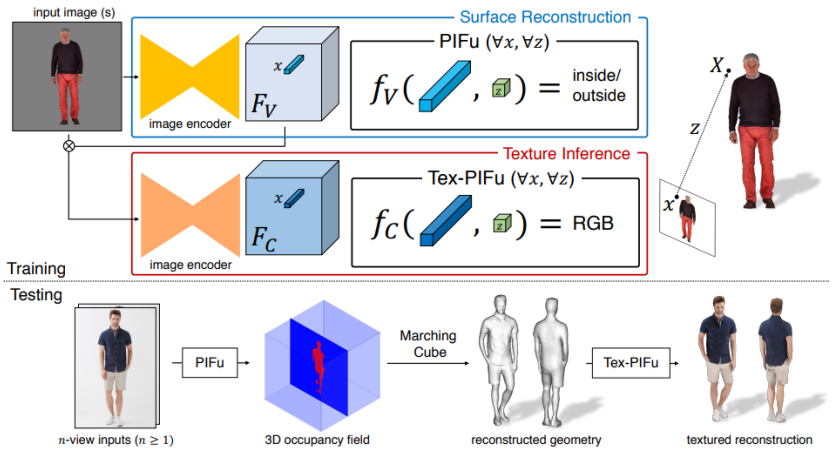


Figure 6: Overview of image in implicit representation(Saito et al. (2019))

5 Time planning

The schedule for the thesis is outlined in Table 3.

Time	period		Tasks
Oct. 13	–	Nov. 17	Literature study and P1 Progress review Graduation Plan.
Nov. 20	–	Dec. 15	Study on 3D deep learning.
Dec. 15	–	Dec. 29	Reproduce the open-source code of related paper and gain a thorough understanding of it.
Jan. 2	–	Jan. 19	Generate the input data required to run the code using Dutch public data and adjust the code to obtain preliminary results. Complete the formal assessment graduation plan for P2.
Jan. 19	–	Jan. 26	Prepare for P2 presentation and evaluate the preliminary result.
Jan. 29	–	Feb. 23	Get ideal point cloud and DSM result in no water area.
Feb. 26	–	Mar. 29	Modify the network architecture, using masks as strong constraint conditions for training and data generation. Prepare for P3 colloquium midterm.
Apr. 1	–	Apr. 19	Write algorithms to evaluate the absolute and relative accuracy of generated dataset.
Apr. 22	–	May. 10	Prepare for P4 formal process assessment.
May. 13	–	Jun. 14	Writing the thesis document and prepare for final defense.

Table 3: Time line

6 Tools and datasets used

Tools:

1. QGIS for generation of masks.
2. Pycharm for sample points from 3D city model, generating test data.
3. Colab for deep learning network.

Dataset:

AHN3,4,5, 3D BAG, BGT, 8cm orthophoto photo.

References

Ackermann, F. (1999). Airborne laser scanning present status and future expectations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54:64–67.

Boulch, A., Saux, B. L., and Audebert, N. (2017). Unstructured Point Cloud Semantic Labeling Using Deep Segmentation Networks. In Pratikakis, I., Dupont, F., and Ovsjanikov, M., editors, *Eurographics Workshop on 3D Object Retrieval*. The Eurographics Association.

Chen, Z. and Zhang, H. (2019). Learning implicit fields for generative shape modeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941.

Dai, A. and Nießner, M. (2022). Neural poisson: Indicator functions for neural fields. *ArXiv*, abs/2211.14249.

Habib, A., Ghanma, M., Morgan, M., and Al-Ruzouq, R. (2005). Photogrammetric and lidar data registration using linear features. *Photogrammetric Engineering and Remote Sensing*, 71:699–707.

He, T., Huang, H., Yi, L., Zhou, Y., Wu, C., Wang, J., and Soatto, S. (2019). Geonet: Deep geodesic networks for point cloud analysis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890.

Li, Y. and Baciu, G. (2021a). Hsgan: Hierarchical graph learning for point cloud generation. *IEEE Transactions on Image Processing*, 30:4540–4554.

Li, Y. and Baciu, G. (2021b). Sg-gan: Adversarial self-attention gcn for point cloud topological parts generation. *IEEE Transactions on Visualization and Computer Graphics*, 28:3499–3512.

Maturana, D. and 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)*, pages 922–928.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2018). Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465.

Mozas-Calvache, A., Gómez-López, J. M., and Pérez-García, J. L. (2023). Multitemporal and multiscale applications of geomatic techniques to medium-sized archaeological sites—case study of marroquíes bajos (jaén, spain). *Remote Sensing*.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174.

Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., and Geiger, A. (2020). Convolutional occupancy networks. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision – ECCV 2020*, volume 12348 of *Lecture Notes in Computer Science*, Cham. Springer.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ran, Y., Zeng, J., He, S., Chen, J., Li, L., Chen, Y., Lee, G., and Ye, Q. (2022). Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. *IEEE Robotics and Automation Letters*, 8:1125–1132.

Reddy, B. R., Uttarakumari, M., S, S. S., Bency, M., D, S., Patil, K. R., and Holla, P. (2022). Machine learning based voxelnet and lunet architectures for object detection using lidar cloud points. *2022 IEEE Delhi Section Conference (DELCON)*, pages 1–6.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314. IEEE.

Skondras, A., Karachaliou, E., Tavantzis, I., Tokas, N., Valari, E., Skalidi, I., Bouvet, G. A., and Stylianidis, E. (2022). Uav mapping and 3d modeling as a tool for promotion and management of the urban space. *Drones*.

Stucker, C., Ke, B., Yue, Y., Huang, S., and Armeni, I. (2022). Implicity: City modeling from satellite images with deep implicit occupancy fields. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2022:193–201.

Toschi, I., Ramos, M. M., Nocerino, E., Menna, F., Remondino, F., Moe, K., Poli, D., Legat, K., and Fassi, F. (2017). Oblique photogrammetry supporting 3d urban reconstruction of complex scenarios. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:519–526.

Vosselman, G. and Maas, H., editors (2010). *Airborne and terrestrial laser scanning*. CRC Press (Taylor & Francis).

Zhang, J. and Lin, X. (2017). Advances in fusion of optical imagery and lidar point cloud applied to photogrammetry and remote sensing. *International Journal of Image and Data Fusion*, 8:1 – 31.

Zhao, C., Yang, J., Xiong, X., Zhu, A., CAO, Z., and Li, X. (2019). Rotation invariant point cloud classification: Where local geometry meets global topology. *ArXiv*, abs/1911.00195.

Zishu, L., Song, W., Tian, Y., Ji, S., Sung, Y., Wen, L., Zhang, T., Song, L., and Gozho, A. (2020). Vb-net: Voxel-based broad learning network for 3d object classification. *Applied Sciences*.