

humantech

D3.5 – 3D Point Cloud Semantic Segmentation Algorithm



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement n° 101058236. This document reflects only the author's view, and the EU Commission is not responsible for any use that may be made of the information it contains.



D3.5 – Semantic Segmentation of point clouds

Project Title	Human-Centred Technologies for a Safer and Greener European Construction Industry.
Project Acronym	HumanTech
Grant Agreement No	101058236
Instrument	Research & Innovation Action
Topic	HORIZON-CL4-2021-TWIN-TRANSITION-01-12
Start Date of Project	June 1, 2022
Duration of Project	36 months

Name of the Deliverable	3D point cloud semantic segmentation algorithm
Number of the Deliverable	D3.5 (D14)
Related WP Number and Name	WP3, Dynamic Semantic Twin Generation
Related Task Number and Name	T3.5, Semantic segmentation of point clouds
Deliverable Dissemination Level	Public
Deliverable Due Date	31 st May 2024
Deliverable Submission Date	31 st May 2024
Task Leader/Main Author	Mahdi Chamseddine (DFKI)
Contributing	Suresh Guttikonda (DFKI), Jason Rambach (DFKI), RPTU,



Partners	ZHAW, RICOH, NASKA
Reviewer(s)	Ruprecht Altenburger (ZHAW)

Keywords

Scan-to-BIM, point clouds, semantic segmentation, panoramic RGB-D images, deep learning

Revisions

Version	Submission date	Comments	Author
V1.0	31 st May 2024	Submitted	Mahdi Chamseddine (DFKI)

Disclaimer

This document is provided with no warranties whatsoever, including any warranty of merchantability, non-infringement, fitness for any particular purpose, or any other warranty with respect to any information, result, proposal, specification, or sample contained or referred to herein. Any liability, including liability for infringement of any proprietary rights, regarding the use of this document or any information contained herein is disclaimed. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted by or in connection with this document. This document is subject to change without notice. HumanTech has been financed with support from the European Commission. This document reflects only the view of the author(s) and the European Commission cannot be held responsible for any use which may be made of the information contained.



Acronyms and definitions

Acronym	Meaning
RGB-D	RGB and Depth image
PC	Point Cloud
BIM	Building Information Model
IFC	Industry Foundation Classes
MLP	Multi-Layer Perceptron
mIoU	Mean Intersection over Union



Abstract

This report presents the implementation, training, and evaluation of machine learning algorithms for semantic segmentation of point clouds and panoramic RGB-D images in the context of construction. The developed methods not only meet the project's key performance indicators but also achieve state-of-the-art results in single frame panoramic RGB-D image semantic segmentation. To support model training, a comprehensive annotation guideline for joint 2D-3D data was proposed and published, filling a gap in research standards. The trained models, leveraging a combination of HumanTech, public, and simulated data, will be integrated into the scan-to-BIM pipeline to generate the Semantic Digital Twin.



The HumanTech project

The European construction industry faces three major challenges: increase the safety and wellbeing of its workforce, improve its productivity, and become greener, making efficient use of resources.

To address these challenges, HumanTech proposes to develop **human-centred cutting-edge technologies** such as wearables for workers' safety and support and robots that can harmoniously coexist with human workers while contributing to the ecological transition of the sector.

HumanTech aims to achieve major advances in cutting-edge technologies that will enable a safe, rewarding, and digital work environment for a new generation of highly skilled construction workers and engineers.

These advances will include:

- **Robotic devices equipped with vision and intelligence** that allow them to navigate autonomously and safely in highly unstructured environments, collaborate with humans and dynamically update a semantic digital twin of the construction site in which they are.
- **Smart, unobtrusive workers protection and support equipment.** From exoskeletons activated by body sensors for posture and strain to wearable cameras and XR glasses that provide real-time workers' location and guidance for them to perform their tasks efficiently and accurately.
- An entirely new breed of **Dynamic Semantic Digital Twins (DSDTs) of construction sites** that simulate in detail the current state of a construction site at the geometric and semantic level, based on an extended Building Information Modelling (BIM) formulation that contains all relevant structural and semantic dimensions (BIMxD). BIMxDs will act as a common reference for all human workers, engineers, and autonomous machines.

The **HumanTech consortium** is formed by 22 organisations — leading research institutes and universities, innovative hi-tech SMEs, and large enterprises, construction groups and a construction SME representative — from 10 countries, bringing expertise in 11 different disciplines. The consortium is led by the German Research Center for Artificial Intelligence's Augmented Vision department.



Contents

1. Introduction.....	8
2. Multimodal Data Annotation Guidelines.....	10
2.1. Ontologies in Construction.....	10
2.2. Annotation Guidelines.....	10
3. Data Collection and Annotation.....	12
4. Semantic Segmentation of Point Clouds.....	13
4.1. Point Cloud Segmentation Network.....	13
4.2. Network Architecture.....	13
4.3. Innovations of PTV2.....	14
Grouped Vector Attention.....	15
Position Encoding Multiplier.....	16
Partition-Based Pooling.....	16
4.4. Modifications Introduced for HumanTech.....	17
5. Semantic Segmentation of Equirectangular RGB-D Images.....	18
5.1. Overall Architecture.....	18
5.2. Encoder Design.....	19
5.3. Decoder Design.....	20
6. Experiments and Segmentation Results.....	22
6.1. Using Simulated Data.....	22
6.2. Training and Evaluating Point Cloud Segmentation.....	23
6.3. Training and Evaluating Panoramic RGB-D Segmentation.....	24
6.4. Evaluation Results.....	26
7. Conclusion.....	28
8. References.....	29

1. Introduction

The generation of a geometric model representation of a building can be a labour and time intensive task, requiring definition of individual objects in buildings starting with structural elements such as walls, columns, and slabs and extending to include mechanical installations such as pipes and vents. This model is an essential part of the digital twin of the building.

The as-built digital twins are created from 3D scans of a building or construction site. Thus, to reduce the time and effort required in generating a digital twin, it is essential to employ algorithms to automate the analysis of the 3D data. Automated scene understanding is done using machine learning algorithms for semantic segmentation, that is the task of detecting a semantic class for the different elements in a scanned scene.

In HumanTech, 3D scans used are present in two formats:

1. **Point clouds** generated by terrestrial scanners (RPTU), drones and photogrammetry (ZHAW), and a robotic scanning platform (NASKA/ZHAW). Figure 1 shows an example point cloud recorded by a terrestrial scanner in Kaiserslautern.



Figure 1 Example of a point cloud recorded by RPTU using the Leica BLK360 G2 at the Prot. Kirchengemeinde in Kaiserslautern

2. **Equirectangular RGB-D images** generated using a 360-degree ToF camera developed by RICOH in T3.2. Figure 2 shows an example of the RGB-D output of the RICOH camera.



Figure 2 Left: An example of a panoramic image captured using the RICOH camera at the Weingarten site. Right: Depth measurements overlapped with the colour information showing the full RGB-D capabilities of the device

Equirectangular RGB-D images can also be used to generate partial point clouds of scenes or fused together to generate more complete point clouds.

The data representations require different semantic segmentation algorithms tailored to each type of data. Successful development and training of machine learning algorithms for semantic segmentation typically requires annotated training data representative and similar to the expected testing data.

While there are multiple annotated datasets for equirectangular RGB-D images such as Stanford2D3DS [21] and Matterport3D [23] and point clouds such as S3DIS [20], they lack the construction domain context as they were not captured with that purpose in mind and thus, they do not fulfil the requirements for creating a Building Information Model (BIM). Furthermore, existing datasets follow different guidelines for annotation which can lead to varied results of semantic segmentation algorithms from one dataset to another.

In WP3, an annotation guideline for multimodal data was developed and published. The guideline was then used to label training data of point clouds and RGB-D images that were subsequently used to train the segmentation algorithms.

The rest of this document is structured as follows: Section 2 explains the annotation guidelines and methodology. Section 3 describes the data collection and annotation for point clouds and RGB-D images. Sections 4 and 5 present the point cloud and image segmentation algorithms followed by their results in Section 6. Finally, the report is concluded and wrapped up in Section 7.



2. Multimodal Data Annotation Guidelines

Deep learning is becoming increasingly popular within the construction industry and supervised semantic segmentation requires annotated datasets for training deep learning models. However, existing datasets do not share a common standard or use formal ontologies (structured knowledge representations) from the construction field.

To mitigate that, a guideline for creating annotated datasets for buildings, using RGB-D and point cloud data, based on construction ontologies was created and published as part of HumanTech [1]. The goal is to make deep learning more applicable to construction problems by providing better datasets that align with construction domain knowledge.

2.1. Ontologies in Construction

Ontologies are structured ways to describe objects and their relationships within a domain. In construction, they define building elements and how they relate, making them useful for organizing and labelling data like images and point clouds.

Various ontologies exist in construction such as Uniclass [17], OmniClass [18], and IFC [19]. For this guideline, the focus was on IFC (Industry Foundation Classes) standard because it is widely used and supports BIM applications.

While IFC defines basic entities, it often lacks detailed descriptions needed for precise labelling of data. For example, the definition of "wall" is provided, but not specifics on where a wall begins and ends in different contexts. Therefore, to use IFC effectively for data annotation, the basic entity definitions need to be supplemented with:

- Clear descriptions of what belongs in each class.
- How to handle boundaries between objects (e.g., between a wall and a slab).
- Specific rules for situations like balconies.

2.2. Annotation Guidelines

These guidelines are meant to provide a standard way of labelling both point clouds and RGB-D data, making things easier and more uniform for data comparison and algorithm training. For that, three main categories are considered:

- Building: Fixed elements of the structure (e.g. walls, doors)



D3.5 – Semantic Segmentation of point clouds

- Construction: Temporary elements used during the building process (e.g. scaffolding)
- Interior: Movable objects within the building (e.g. furniture)

In addition to that, some special labels are defined for some useful categories like "invalid data" (overexposed images, clutter in point clouds) are included to aid in neural network training.

The guidelines address different stages of construction, they are unified for both 2D and 3D data, and conform to IFC classes for compatibility with BIM systems. More details are available in the work by Kaufmann et al. [1] (HumanTech partner RPTU) and on the repository dedicated for it <https://gitlab.rhrk.uni-kl.de/kaufmann/humantech-data-annotation>

3. Data Collection and Annotation

Data collection was done using multiple devices and by different partners at different locations. Table 1 provides a list of the major recordings done within HumanTech and which partners were involved as well as the type of data captured. The data annotation was handled by DFKI and RPTU for the panoramic RGB-D images and point cloud data respectively.

Table 1 A brief overview of the most important locations where data was recorded using multiple sensors within WP3

Nb.	Where / When	Data capture type	Involved partners
1	Weingarten/Germany Oct 2022	RPTU Laserscan (Leica BLK360) NASKA robot RICOH RGB-D camera	RICOH, ZHAW, RPTU, DFKI, NASKA
2	Winterthur Lokstadt/CH (Implenia) Feb. 2023	RICOH RGB-D camera	IMPLENIA, RICOH
3	Aarau/Switzerland (Implenia) July. 2023	RICOH RGB-D camera	IMPLENIA, RICOH
4	Neustadt/Germany ADAC July. 2023 – Dec. 2023	RICOH RGB-D camera RPTU Laserscan (Leica BLK360)	RPTU, RICOH
5	Kaiserslautern/Germany prot. Church	RPTU Leica BLK360 G2	RPTU
6	Kaiserslautern/Germany DFKI building May 2022 – July 2022	RICOH RGB-D camera	RICOH

In addition to the data captured within the project, RPTU also worked on annotating the public data used for the CV4AEC scan-to-BIM challenge [2]. The challenge is intended for building understanding and the data originally provided labels for walls, doors, and columns which were improved and extended by RPTU to include the annotations defined by the HumanTech guidelines described in Section 2.



4. Semantic Segmentation of Point Clouds

Semantic segmentation on point clouds is a machine learning task in which deep neural networks are trained to reason on point cloud scans. There are three main approaches for learning from 3D point clouds: projection-based, voxel-based, and point-based networks:

- Projection-based networks project point clouds onto regular grids (like images) and then process them with 2D convolutional neural networks (CNNs). This approach is intuitive but does not efficiently utilise the sparsity of point clouds and leads to loss in geometric information [9, 10].
- Voxel-based networks convert point clouds into 3D voxels and then apply 3D convolutions. Those networks are computationally expensive and result in the loss of geometric detail due to quantisation [11, 12].
- Point-based networks process point clouds directly as sets using permutation-invariant operators. They are more flexible and can better capture the geometric relationships between points. Some recent work has focused on using self-attention mechanisms in point-based networks, which has shown promise for large-scale 3D scene understanding [13, 14].

Point Transformer by Zhao et al. [3] builds upon the foundations of point-based networks and self-attention mechanisms, utilising local self-attention [15], vector attention [16], and appropriate positional encoding.

4.1. Point Cloud Segmentation Network

In HumanTech, the point cloud semantic segmentation was based on from the state-of-the-art deep neural network Point Transformer V2 by Wu et al. [4]. The network was adapted to train on more point cloud features be more robust to differences in sensor measurement characteristics.

4.2. Network Architecture

Similar to the original Point Transformer [3], PTV2 follows a U-Net style encoder-decoder architecture with skip connections. This is a common structure in point cloud processing tasks. Figure 3 shows the network architecture as four encoder stages followed by four

decoder stages with feature dimensions and attention groups increase at each encoder stage.

For the attention blocks, Neighbourhood Attention was used since it focuses on the nearest neighbours of each point. This has been shown to be better for point clouds than shifted-grid attention. Finally, an MLP (Multilayer Perceptron) processes features for each point to generate class predictions for semantic segmentation.

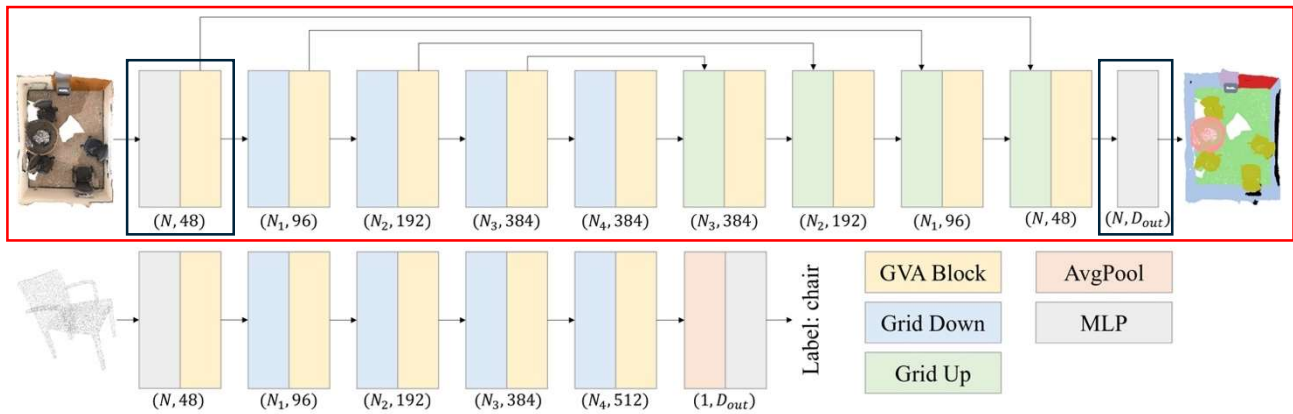


Figure 3 Network architectures for semantic segmentation and classification. The semantic segmentation architecture marked in red with the parts modified for HumanTech in blue.

For point cloud segmentation in HumanTech, the network architecture was modified for the encoder and decoder were modified as indicated in Figure 3.

4.3. Innovations of PTV2

Point Transformer V2 builds upon the success of Point Transformer for 3D point cloud understanding tasks with several improvements, Figure 4 presents the added concepts:

- **Group Vector Attention:** A more effective attention mechanism compared to the previous version. It combines the benefits of learnable weight encoding and multi-head attention with a grouped weight encoding layer.
- **Enhanced Position Encoding:** Strengthens the model's ability to reason about spatial relationships in the point cloud data by incorporating an additional position encoding multiplier.
- **Partition-Based Pooling:** Introduces a new pooling strategy that improves spatial alignment and sampling efficiency compared to previous methods.

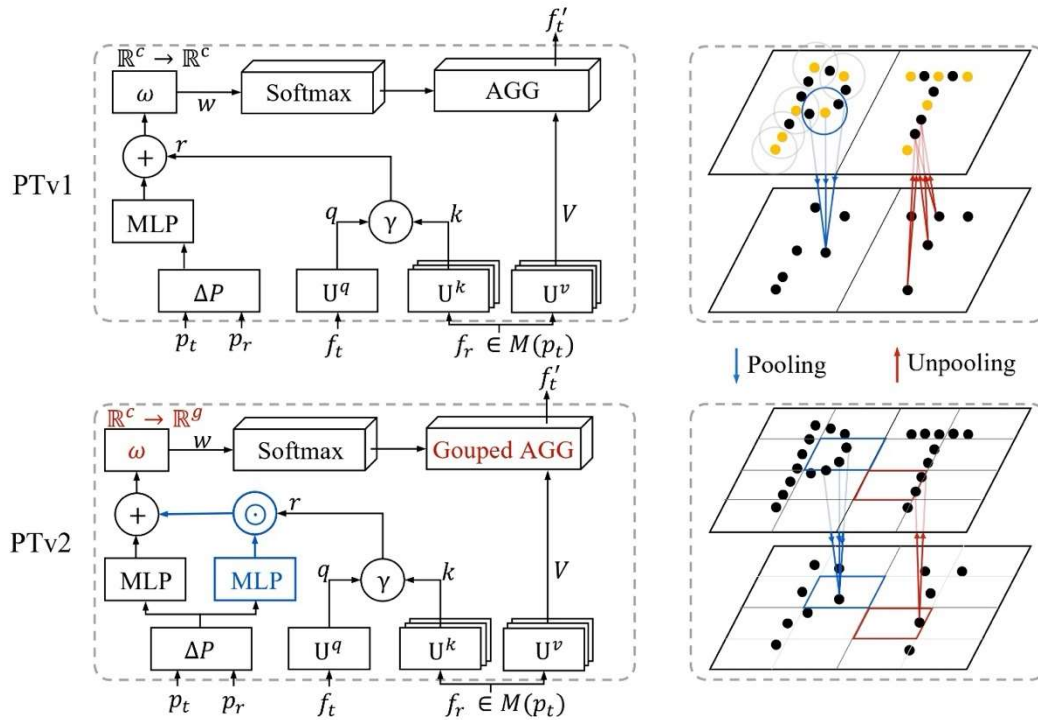


Figure 4 The innovations introduced by Point Transformer V2 as compared to the original Point Transformer network. On the left, both the changes to the positional encoding (MLP) and Grouped Attention are shown. On the right, partition-based pooling is illustrated.

Grouped Vector Attention

The original vector attention, while effective, becomes heavy in parameters as the model deepens and the number of channels increases. This hinders efficiency and the model's ability to generalize.

Vector attention used in the original Point Transformer suffers from a large number of parameters the larger the number of channels increases and the deeper the model becomes; this makes it inefficient and harder to train and generalize.

To overcome the limitations, the authors introduced Grouped Vector Attention (GVA) in which feature channels are divided into groups that share a single attention weight thus reducing the number of parameters compared to vector attention.

The authors further introduce a Grouped Linear Layer within the GVA weight encoding function. This layer independently projects different groups of the input vector with different parameters, further reducing the number of parameters.

Those improvements not only improve the model's efficiency and generalizability, but also reduce its training and inference time.



Position Encoding Multiplier

Unlike 2D images with a regular-grid structure, points in 3D point clouds are distributed unevenly. Transformers typically inject spatial information by adding a position encoding bias to the relation vector. Due to generalization limitations mentioned earlier, simply adding more positional information to the original vector attention would not improve the model's performance.

The authors introduced a position encoding multiplier to the relation vector along with the standard position encoding bias and is meant to focus the learning process on complex point cloud positional relationships.

The position encoding multiplier and GVA work together to provide a balanced approach, giving the model the necessary spatial understanding without excessively increasing its complexity.

Partition-Based Pooling

Existing methods rely on sampling points and then querying neighbour points, for example, Point Transformer uses farthest point sample (FPS) and k-Nearest Neighbours (kNN) for pooling and down sampling the points. This leads to spatial and query overlap misalignment since the queried sets of points may not align well in space and how much neighbourhood queries overlap can be inconsistent.

To overcome those limitations, partition-based pooling was introduced. First, the 3D point clouds are divided into non-overlapping partitions. Then, for each partition, the features are aggregated using max pooling to select the most prominent features from the points within that partition and an average position (mean pooling) is used to average the locations of the points and compute a representative position.

To reverse the pooling stage (up sampling/un-pooling), features from the pooled set are assigned directly to points in the original set, based on which partition they belonged to during the pooling stage.

This pooling approach is not only faster and more efficient than sampling and querying, but also ensures that information is aggregated within well-defined spatial partitions, improving alignment.



4.4. Modifications Introduced for HumanTech

To better address the challenges of construction data, several key modifications have been introduced to the Point Transformer V2 (PTv2) network. The feature encoding has been improved by incorporating normals, enabling the model to capture more detailed geometric information from 3D point cloud data and allow the transformer blocks to extract better features from neighbouring points.

Additionally, changes have been implemented to improve generalizability when training on HumanTech project data. Specifically, an augmentation strategy has been added to the training process, where colour information is randomly dropped from some of the samples. This enhancement enables the model to learn robust features that are less reliant on colour data, ultimately improving its performance on datasets captured by different types of sensors that may not be capable of capturing colour information.

Furthermore, the segmentation classes have been modified to align with the specific requirements of the construction field, allowing for more accurate and relevant results in this domain. These modifications have been added to optimize the performance of PTv2 on construction data, particularly when applied to datasets captured within the HumanTech project.

5. Semantic Segmentation of Equirectangular RGB-D Images

In contrast to 3D point clouds which represent a space by a set of points, panoramic images provide a comprehensive 360-degree view, offering valuable information for scene understanding tasks. However, working with this data involves some challenges mainly because processing panoramic images requires dealing with their distortions. Presently, existing research primarily focuses on semantic segmentation for traditional "pinhole" camera images. This doesn't directly translate well to the unique distortions present in panoramic images.

In HumanTech, a new architecture using transformers to combine information from different modalities has been developed. This architecture specifically addresses the distortions inherent in panoramic formats. In addition to that, multi-modal data (e.g., RGB colour images, depth maps, etc.) were leveraged to improve the robustness of image understanding algorithms.

The neural network architecture developed in HumanTech for equirectangular RGB-D data achieves state-of-the-art performance on several public panoramic image datasets. The work was presented at IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2024 under the title "Single Frame Semantic Segmentation Using Multi-Modal Spherical Images" [8].

5.1. Overall Architecture

The design is inspired by Trans4PASS+ [5] and uses a similar structure. Notably, it incorporates techniques from CMX [6] for feature extraction and fusion across all three data modalities. Figure 5 shows the overall architecture of the network: The input, a panoramic image, is first broken into patches then passed to a **hierarchical encoder**. The encoder processes patches at multiple resolutions to handle panoramic distortions and enables cross-modal interactions between the different data modalities. Finally, the **panoramic decoder** takes the encoder output and generates a segmentation mask the same size as the original image.

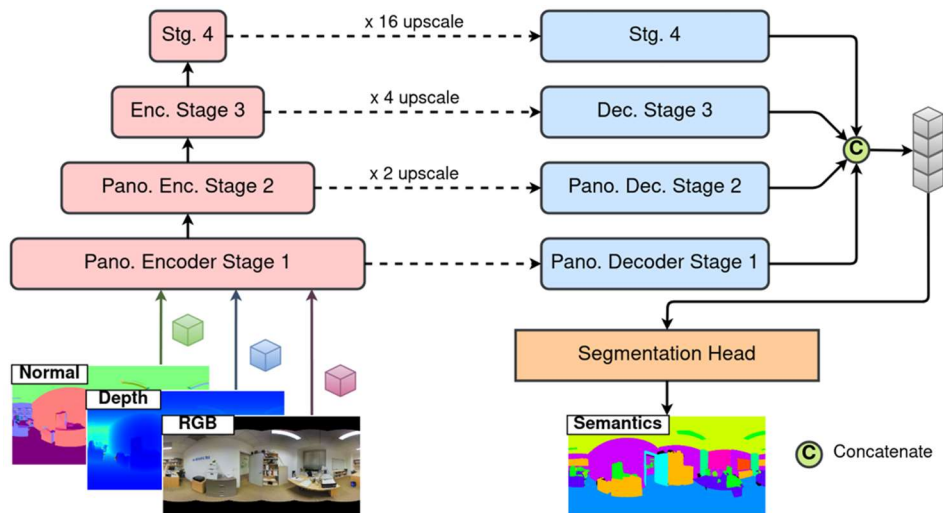


Figure 5 Overview of the multi-modal panoramic segmentation architecture [8]. The inputs are a combination of RGB, Depth, and Normals. The architecture shows 4 encoder stages followed by 4 decoder stages and a segmentation head.

5.2. Encoder Design

The four-stage encoder, Figure 5 Left, is specifically designed to handle the distortions attributed to panoramic images. Each encoder stage is made up of three main modules:

- **Deformable Patch Embeddings (DPE)**: Unlike standard patch embeddings, DPE presented by Zhang et al. [7] learns where to sample within image patches to better handle distortions and deformations present in panoramic images by using deformable convolutions to calculate dynamic offsets for each patch.
- **Cross-modal Feature Rectification Module (CM-FRM)**: Based on the work of Liu et al. [6] but expanded to work for additional modalities. It is designed to handle noise and inconsistencies across different modalities (RGB, Depth, Normals) by filtering and calibrating features from one modality using information from the others. The CM-FRM, seen in Figure 6, has channel-wise and spatial-wise stages for both local and global calibration. The output of the CM-FRM module is used as an input to the next encoder stage as well as the feature fusion module.

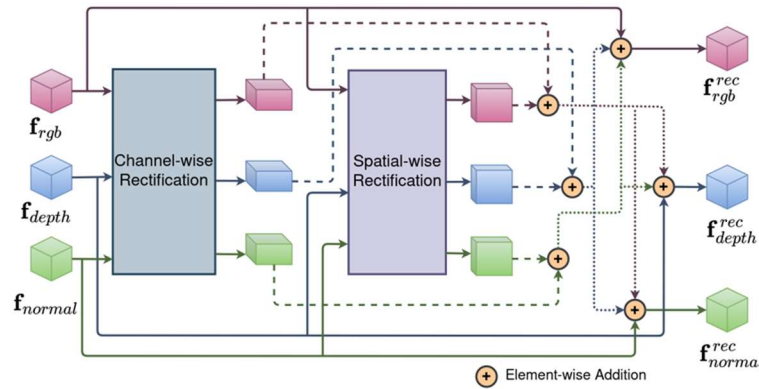


Figure 6 Cross-modal feature rectification module to calibrate RGB, Depth, and Normals features. The rectified features are used for the feature fusion as well as input to the next encoder stage.

- **Cross-modal Feature Fusion Module (FFM):** It allows for information exchange and combination of features from all three modalities. Expands on the work by Liu et al. and uses a multi-head cross-attention mechanism for global interaction. Channel embedding combines the features from different modalities, resulting in a single comprehensive feature map as can be seen in Figure 7. The output of each FFM module is forwarded to its respective decoder stage.

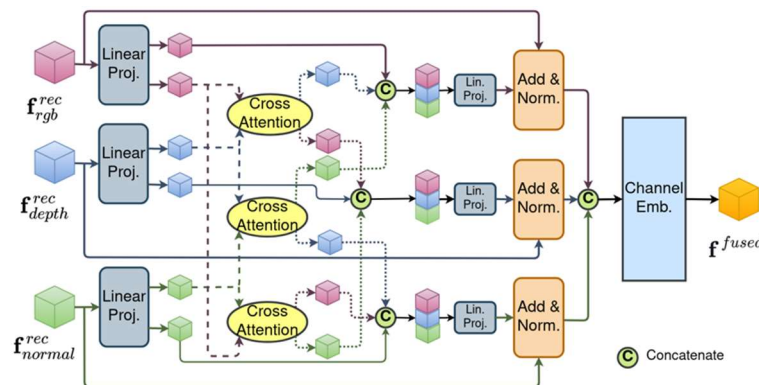


Figure 7 Cross-modal feature fusion module to fuse RGB, Depth, and Normals features. The fused output of each encoder stage is input to its respective decoder stage.

5.3. Decoder Design

Contrary to regular MLP decoder commonly used in semantic segmentation tasks and that have shown to not be effective in the case of panoramic images, a Deformable Token Mixer (DMLPv2) proposed by Zhang et al. [5] and better handles deformations was used. Each encoder stage is coupled with an appropriately designed decoder stage.

Figure 8 shows the architecture of the DMLPv2. It combines several specialized components to improve feature fusion:

- A **Channel Mixer** (CX) which emphasizes important information at the channel level for improved feature representation.
- A **Pooling Mixer** (PX) which uses average pooling for spatial-wise sampling with fixed offsets.
- A **Deformable MLP** (DMLP) that learns adaptive spatial offsets for a more flexible representation.

Finally, the outputs of all decoders are concatenated and passed to a segmentation head.

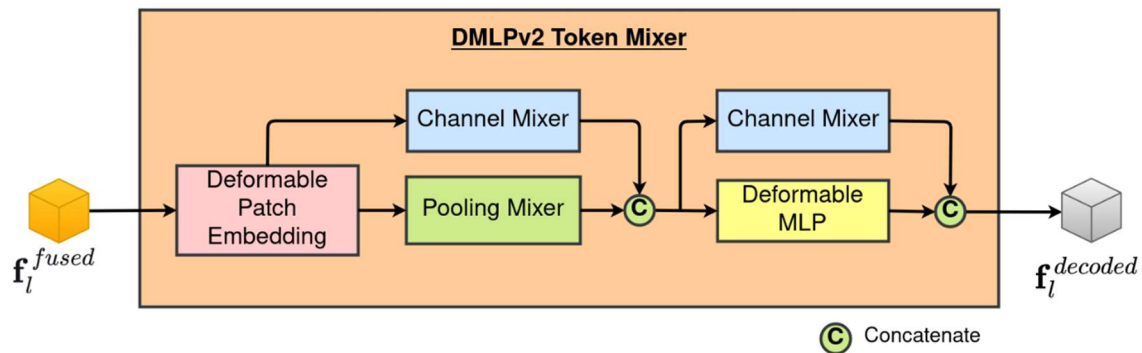


Figure 8 Panoramic decoder stage with fused features from RGB, Depth, and Normals modalities. The outputs of all decoders are concatenated before being passed to the segmentation head.

6. Experiments and Segmentation Results

The algorithms presented in Sections 4 and 5 were evaluated on their respective data to verify their validity. While the point cloud algorithm PTV2 has already been published and evaluated by the authors [4], further changes were required to allow for deployment in the construction domain using the data captured in HumanTech. Meanwhile, the panoramic semantic segmentation algorithm presented in Section 5 and developed as part of the HumanTech project by Guttikonda et al. [8] has been tested on public datasets and managed to achieve the state-of-the-art results.

6.1. Using Simulated Data

As part of Work Package 3 of the project, and as presented in D3.4, simulated data from buildings has been generated to be used for the purpose of training of deep neural networks for point cloud semantic segmentation. This data is used as part of the pool of training data sources: public datasets, simulated data, and data captured and annotated within HumanTech (Section 3) to train and fine tune a point cloud.

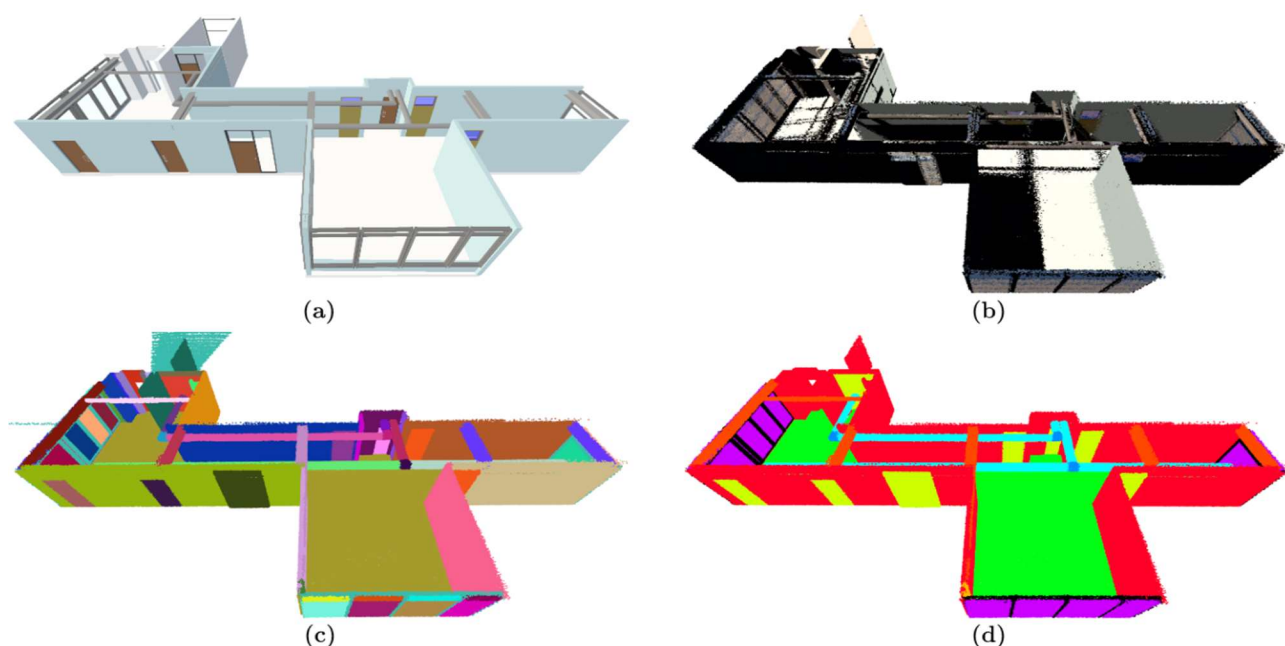


Figure 9 (a): BIM model of DFKI office floor obtained manually and/or from T3.6 scan-to-bim pipeline, (b): Synthetic RGB point cloud generated using BIM model in UnrealEngine5, (c) and (d): Object instances and semantics ground truth annotations for generated synthetic scan.

Figure 9 shows the results of generating synthetic point clouds from BIM using the Unreal Engine. The synthetic data is similar to the real data available from point cloud datasets [20] where both instance and class annotations are provided.

6.2. Training and Evaluating Point Cloud Segmentation

In order to make use of all available data for training the model, a modified version of Point Transformer V2 was first trained using public point cloud data (S3DIS). S3DIS is a dataset published by Armeni et al. at Stanford University [20]. The data is coloured point clouds and comprised of five areas representing different buildings with multiple rooms each.

When first training the model, point cloud normals are used as an extra modality input in addition to cartesian coordinates and colour. More feature inputs allow for better scene understanding especially in the field of construction where normals are of significant value for plane detection (eg. Walls, floors, ceilings). Furthermore, as part of the data augmentation, the colour is randomly dropped from the data to simulate different types of point cloud measuring sensors that only record intensity instead of colour. This is essential for the fine-tuning step as the data used for that step is a mixture between colour and intensity data.

During the fine-tuning step, the detection head of the network that corresponds to the classes present in the S3DIS dataset is replaced by a new detection head that is relevant to the semantic segmentation task for HumanTech with a different subset of classes. The data used in fine-tuning is another public dataset published within the CV4AEC challenge [2] and intended for building understanding. The training done with a lower learning rate and evaluated on the data captured within HumanTech. Figure 10 shows an example of a single floor from the Weingarten ICU building captured by RPTU. The walls floors and ceiling as well as other classes such as doors and windows are well segmented, the ceiling has been partially removed for visualization purposes and to show the quality of the segmentation.

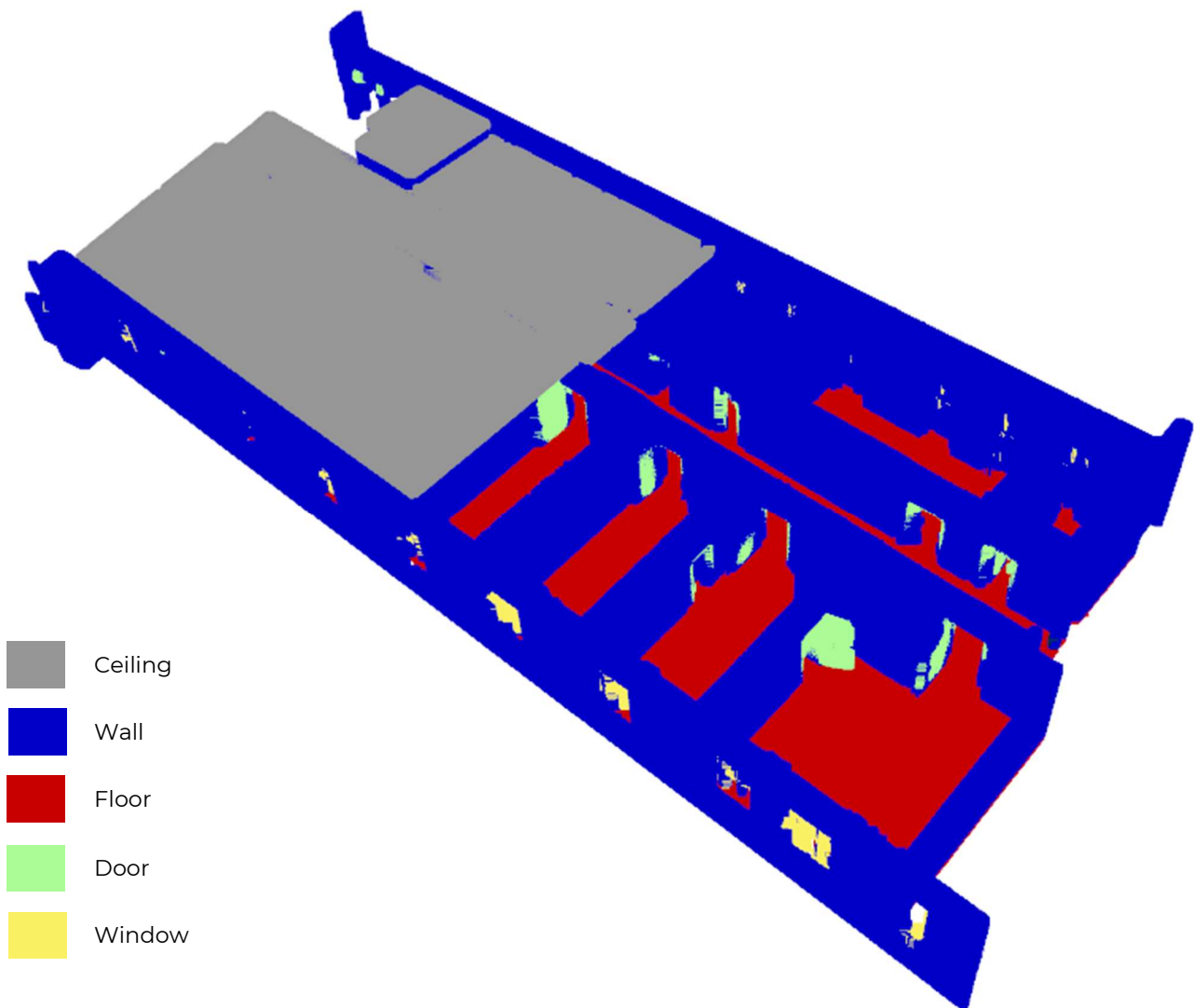


Figure 10 Semantic segmentation results of a sample point cloud captured by RPTU representing the ICU building at the Weingarten Hospital. The ceiling has been removed to show the results of the segmentation on the classes inside the scan.

6.3. Training and Evaluating Panoramic RGB-D Segmentation

The panoramic RGB-D segmentation algorithm was evaluated using multiple modality inputs and using different publicly available datasets: Stanford2D3DS [21], Structured3D [22], and Matterport3D [23].

The results showed a significant improvement over existing algorithms and state-of-the-art segmentation on the mentioned datasets.

Figure 11 shows examples of the semantic segmentation as compared to the ground truth on a myriad of classes. Details of the evaluation are presented in the published research paper by Guttikonda et al. [8].

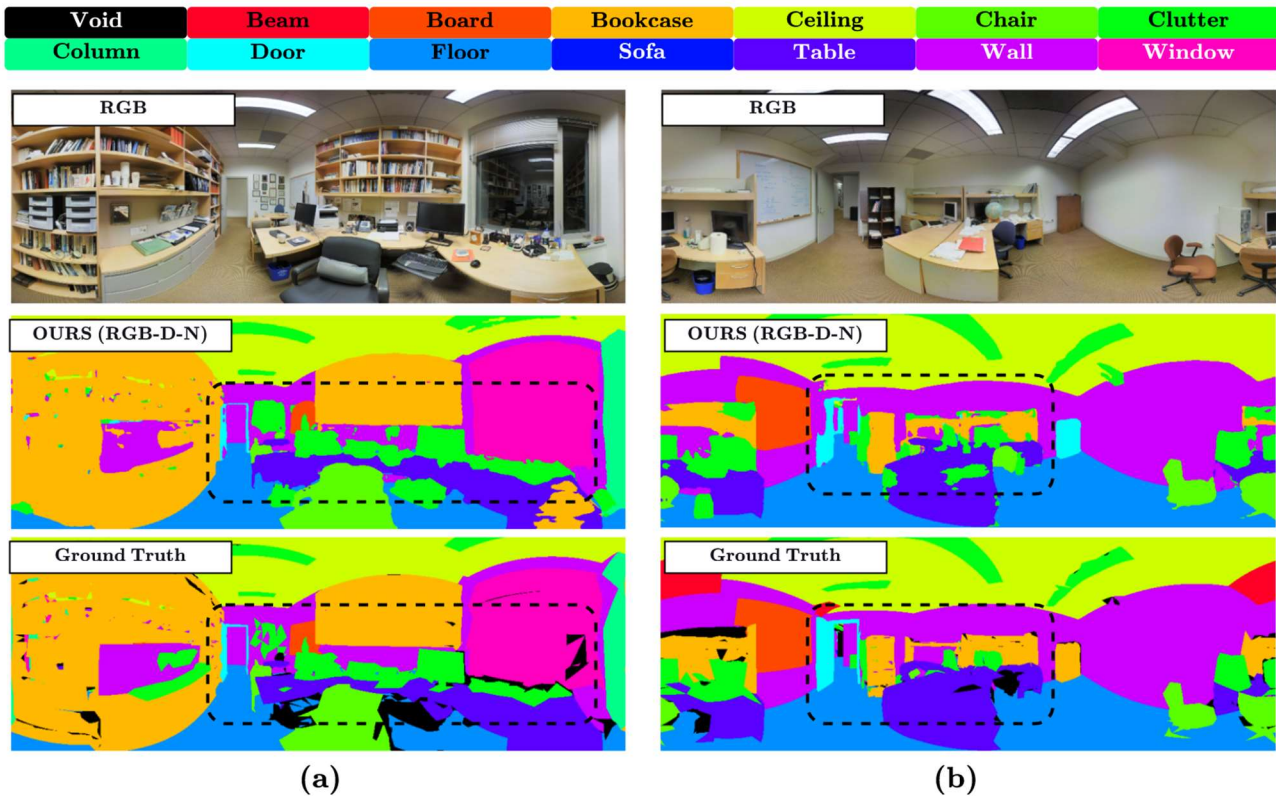


Figure 11 Samples from the Stanford2D3DS [21] semantic segmentation visualizations.

6.4. Qualitative Evaluation

Figures 10 and 11 show high quality segmentation results of point clouds and panoramic RGB-D images.

For Point Clouds, the trained model was able to properly segment main sections of the building, more specifically, floors, walls, ceilings, and doors. The windows were also segmented albeit with less success, this can be attributed to the higher variety of window structures between different datasets and buildings. Furthermore, there was an attempt to segment columns that was not successful in some cases because columns are often embedded into walls which makes it very difficult to differentiate between columns and walls.

As for the panoramic RGB-D images, the model has claimed state-of-the-art results on multiple benchmarks, such as the Stanford2D3DS [21] dataset. The figure shows very

good segmentation results compared to the ground truth with some shortcomings in image patches with clutter as seen in Figure 11 (b).

6.5. Quantitative Evaluation

The quantitative results for both algorithms are measured as mean Intersection over Union between the predicted and expected segmentation results. The Intersection over Union (IoU), also referred to as the Jaccard Index, is a common metric used for semantic segmentation tasks as is defined as the: number of true positives divided by the sum of false positives, false negatives, and true positives:

$$IoU = \frac{TP}{TP + FP + FN}$$

The mean Intersection over Union (mIoU) is thus defined as the weighted sum of the IoUs for all the classes available in a dataset:

$$mIoU = \frac{\sum n_i \times IoU_i}{n}$$

Where n is the number of instances, n_i and IoU_i are the number of instances per class and IoU per class respectively, and i is the class index.

Table 2 presents the results of both point cloud and panoramic RGB-D segmentation algorithms on various datasets. The table shows that the KPIs defined in the project were achieved.

The HumanTech dataset corresponds to measurements done by RPTU at the Weingarten hospital, two floors, ICU and Offices were annotated by RPTU and used in the evaluation. As explained in Section 6.4, the difficult task of segmenting columns and windows led to lower segmentation results on this dataset. Due to the small size of dataset, the model was first pretrained on the S3DIS dataset [20] then fine-tuned on the CV4AEC dataset [2] while the HumanTech data was used for evaluation.



D3.5 – Semantic Segmentation of point clouds

Table 2 Results of the semantic segmentation models on different public and project data. The results show that KPIs defined in the project were achieved.

KPI	Task	Target mIoU		Dataset	mIoU
		M18	M36		
K3.08	[T3.5] Semantic Segmentation Accuracy (Point Cloud scans dataset)	0.55	0.65	HumanTech	0.50
				S3DIS [20]	0.69
K3.09	[T3.5] Semantic Segmentation Accuracy (RGB-D Single Frame)	0.4	0.5	Stanford2D3DS [21]	0.55
				Structured3D [22]	0.73
				Matterport3D [23]	0.39

7. Conclusion

This report covers the implementation and training of machine learning algorithms for semantic segmentation of point clouds and panoramic RGB-D images for construction. Not only did the trained methods achieve the required KPIs set in the project, but also a state-of-the-art method was developed for Single Frame Panoramic RGB-D Image Semantic Segmentation within this work package.

Since training those models requires big amounts of training data, an effort was also made to collect and annotate this data and thus a proper annotation guideline was proposed for joint 2D-3D data annotation inspired by the construction ontologies. This guideline was also published to fill a gap in the research for such a standard.

The data collected as part of HumanTech, in addition to public and simulated data are used to deliver the models presented in Sections 4 and 5. Finally, the fruits of the work done can be reaped as part of the upcoming deliverable D3.6 where the semantic segmentation is integrated into the scan-to-BIM pipeline that finally generates the Semantic Digital Twin.

It is expected that with more data becoming available and incremental improvements achieved in the development and training of deep neural networks, the results of the semantic segmentation are set to improve and subsequently introduce more improvements to the scan-to-BIM pipeline.

8. References

- [1] Kaufmann, Fabian, et al. "Ontology-based semantic labeling for RGB-D and point cloud datasets." EC3 Conference 2023. Vol. 4. European Council on Computing in Construction, 2023.
- [2] Armeni, Iro, et al. "Computer Vision in the Built Environment." GitHub Pages, cv4aec.github.io, 2024.
- [3] Zhao, Hengshuang, et al. "Point transformer." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
- [4] Wu, Xiaoyang, et al. "Point transformer v2: Grouped vector attention and partition-based pooling." Advances in Neural Information Processing Systems 35 (2022): 33330-33342.
- [5] Zhang, Jiaming, et al. "Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation." arXiv preprint arXiv:2207.11860 (2022).
- [6] Zhang, Jiaming, et al. "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers." IEEE Transactions on Intelligent Transportation Systems (2023).
- [7] Zhang, Jiaming, et al. "Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [8] Guttikonda, Suresh, and Jason Rambach. "Single Frame Semantic Segmentation Using Multi-Modal Spherical Images." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
- [9] Chen, Xiaozhi, et al. "Multi-view 3d object detection network for autonomous driving." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.
- [10] Lang, Alex H., et al. "Pointpillars: Fast encoders for object detection from point clouds." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- [11] Graham, Benjamin, Martin Engelcke, and Laurens Van Der Maaten. "3d semantic segmentation with submanifold sparse convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.



- [12] Choy, Christopher, JunYoung Gwak, and Silvio Savarese. "4d spatio-temporal convnets: Minkowski convolutional neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- [13] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [14] Wang, Yue, et al. "Dynamic graph cnn for learning on point clouds." ACM Transactions on Graphics (tog) 38.5 (2019): 1-12.
- [15] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
- [16] Zhao, Hengshuang, Jiaya Jia, and Vladlen Koltun. "Exploring self-attention for image recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- [17] National Building Specification. "Uniclass - Unified Construction Classification." National Building Specification, 2022, <https://www.thenbs.com/our-tools/uniclass>. Accessed 30 Apr. 2024.
- [18] Construction Specification Institute. "OmniClass." Construction Specification Institute, 2023, <https://www.csiresources.org/standards/omniclass>. Accessed 30 Apr. 2024.
- [19] Buildingsmart. "Industry Foundation Classes 4.0.2.1." Buildingsmart, 2020, https://standards.buildingsmart.org/IFC/RELEASE/IFC4/ADD2_TC1/HTML/. Accessed 30 Apr. 2024.
- [20] Armeni, Iro, et al. "3d semantic parsing of large-scale indoor spaces." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [21] Armeni, Iro, et al. "Joint 2d-3d-semantic data for indoor scene understanding." arXiv preprint arXiv:1702.01105 (2017).
- [22] Zheng, Jia, et al. "Structured3d: A large photo-realistic dataset for structured 3d modeling." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020.
- [23] Chang, Angel, et al. "Matterport3d: Learning from rgb-d data in indoor environments." arXiv preprint arXiv:1709.06158 (2017).

