Learning From Detailed Maps: Joint 2D-3D Semantic Segmentation for Airborne Data with Selective Label Fusion

Geethanjali Anjanappa[a]*, Sander Oude Elberink[a], George Vosselman[a]

[a]Department of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente

Keywords: Topographic Maps, Deep Learning, Multimodal Semantic Segmentation, 2D-3D Airborne Data, Label Fusion

Abstract

Objects for topographic maps are often extracted manually by interpreting and segmenting airborne data, such as 2D images and 3D point clouds. Deep learning (DL) with semantic segmentation can automate this process using existing maps as ground labels. However, current map-based DL methods are limited to either 2D or 3D, focus on urban regions, segment only a few generic classes, and overlook the effects of abstractions in map-derived labels. To overcome these limitations, we propose a segmentation method that uses maps as ground truth with (i) joint 2D and 3D networks using multi-scale feature learning to capture fine details and segment diverse objects and (ii) a Selective Label Fusion module to refine predictions across both modalities, addressing the effects of map abstractions. Trained and tested in urban, rural, and forested regions, our method segments 11 map-based classes in 2D and 12 classes in 3D. At the class level, we achieve a mean Intersection over Union (mIoU) of 70% for both 2D and 3D, with label fusion improving 3D performance by 15% over non-fused results. Regionally, 4 out of 5 areas achieve mIoU above 60% in both modalities. These results demonstrate the potential of maps and DL to automate the labeling of images and point clouds, helping to create and update maps while also generating valuable labeled datasets for other computer vision tasks.

1. Introduction

Topographic maps provide detailed information on objects in the real world, such as buildings, roads, water bodies, and vegetation. These data are crucial for planning and decision-making in public space management, infrastructure development, and emergency response. In the Netherlands, for example, government agencies are required to use Basisregistratie Grootschalige Topografie (BGT) maps for spatial data in public duties, promoting consistency and coordination among agencies by using standardized geoinformation (PDOK, 2025).

Despite their importance, creating and updating topographic maps is predominantly manual, where cartographers and geoinformation specialists interpret and identify objects from geospatial data. For BGT maps, multiple agencies independently create and manage their domain-specific maps using airborne 2D images and 3D point clouds. Although effective, these methods are time-consuming, labor-intensive, and impractical for large-scale or nationwide mapping (Knudsen and Olsen, 2003).

Recent advances in deep learning (DL) for semantic segmentation classify airborne data, providing an automated solution to manual mapping efforts. However, effective use of these methods requires extensive labeled data covering diverse map-based classes in both 2D and 3D. Existing benchmarks, such as 2D-3D ISPRS, Hessigheim-3D, and Vorarlberg-3D, are limited, either providing labels for only a few generic classes or lacking labels across both 2D and 3D data (Rottensteiner et al., 2014; Kölle et al., 2021; Vorarlberg, 2024). Additionally, manual labeling for detailed classes is costly and time-intensive.

In principle, detailed maps can serve as ground truth for various classes beyond generic categories. Some DL methods have successfully used map-derived labels for the semantic segmentation of airborne data (Kaiser et al., 2017; Yang et al., 2020; Widyaningrum et al., 2021). However, these methods are limited to single modalities (either 2D or 3D), segment only a few generic classes, focus mainly on urban areas, and do not address abstraction in map-derived labels.

Fundamentally, maps are abstract and generalized in nature, which can sometimes lead to inconsistencies between object representations and visual details captured in airborne data. Particularly in 2D images, where only top-surface information is captured, occluding the objects below. For example, the yellow box in Figure 1b shows only trees, while the corresponding map in Figure 1a abstracts this area as roads, water, and shrubs. These limitations can be addressed using multimodal data, where 3D point clouds effectively capture objects beyond occlusions, providing better scene understanding, as seen in Figure 1c.

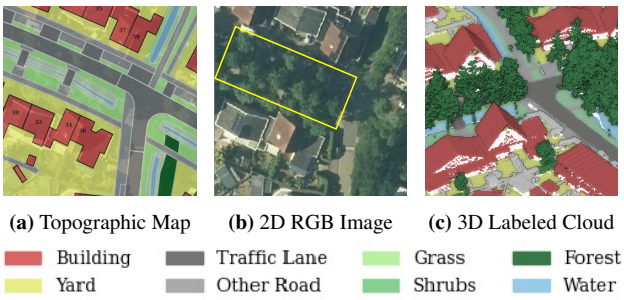


Figure 1. Example of map, 2D image, and 3D point cloud. Objects occluded in the 2D image (yellow box) are captured in the 3D point cloud, aligning with objects on the map.

In this study, we propose a multimodal segmentation method that uses map-derived labels with a two-fold approach: (i) joint 2D and 3D networks with multi-scale feature learning to capture detailed features and segment diverse classes and (ii) a “Selective Label Fusion” (SLF) module that refines multimodal predictions, addressing the effects of map abstractions. By com-

* Corresponding author: g.anjanappa@utwente.nl

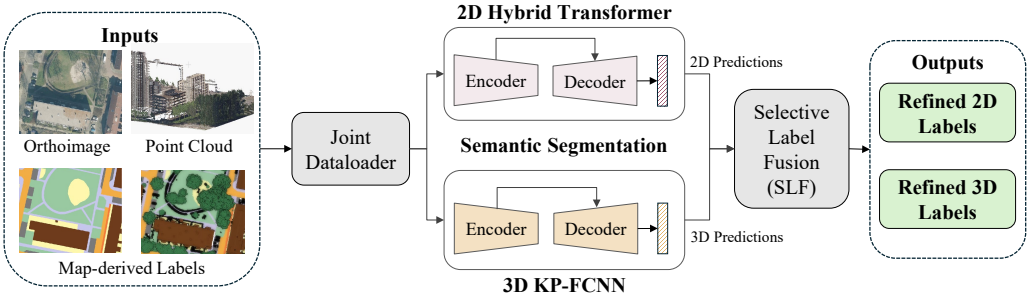


Figure 2. Workflow of proposed method - synchronized inputs from dataloader are processed by 2D hybrid Transformer and 3D KP-FCNN, each including an encoder, decoder, and prediction head to generate 2D and 3D predictions, which are then refined with SLF.

bining multimodal segmentation with SLF, our approach leverages the strengths of each modality to improve segmentation accuracy without adding computational complexity and simultaneously addressing the effects of map abstractions. Unlike existing map-based methods, we use data from urban, rural, and forested regions to demonstrate robustness and adaptability.

For 2D segmentation, we use a Swin Transformer-based architecture (Liu et al., 2021) to capture contextual features to effectively distinguish visually similar but contextually different classes. For 3D, we use the Kernel Point Fully Convolutional Neural Network (KP-FCNN) to learn geometric features from point clouds (Thomas et al., 2019). Further, we use additional features for 3D, such as height above ground to differentiate ground from non-ground classes and reflectance to distinguish roads from other ground-based classes.

The SLF module refines multimodal predictions by selectively transferring class predictions between modalities, allowing 2D predictions to correct 3D predictions as needed and vice versa. Unlike standard output fusion, SLF uses class-specific label transfer, preventing unnecessary refinements and minimizing noise. Classes for label transfer are manually chosen based on a semantic understanding of occlusions and the premise that some classes are better captured in 2D through color and context while others are in 3D through geometry.

2. Related Works

Semantic segmentation involves partitioning images or point clouds into meaningful regions by assigning each pixel or point a semantic label. This section reviews recent developments in 2D, 3D, and multimodal segmentation approaches.

2.1 2D Image Segmentation

Convolutional Neural Networks (CNNs) have been widely used for dense segmentation tasks, including strategies such as skip connections in UNet (Ronneberger et al., 2015) and dilated convolutions in DeepLab (Chen et al., 2017) to improve feature extraction (Long et al., 2015). However, these methods primarily capture local features and do not model the global context.

Transformers address this limitation through attention mechanisms to capture long-range dependencies, with models like Vision Transformer (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021) demonstrating improved performance (Vaswani et al., 2017). More recently, hybrid models integrating Transformers with CNNs have been proposed to integrate global contexts and local features, leading to improved segmentation results (Wang et al., 2022; Ding et al., 2022).

2.2 3D Point Cloud Segmentation

Extending DL methods to 3D data presents challenges due to their irregular and unordered nature (Bello et al., 2020). Early methods converted clouds into grids or voxels, often leading to information loss (Tchapmi et al., 2017; Graham et al., 2017). Later, point-wise methods like PointNet (Charles et al., 2017) and PointNet++ (Qi et al., 2017) directly processed clouds using Multi-Layer Perceptrons (MLPs). However, they required a large number of parameters to capture complex patterns.

Subsequent methods introduced geometric convolutions, such as KP-FCNN (Thomas et al., 2019) and SPNET (Li et al., 2021), which adapt to local geometry and address some limitations of MLP-based models. More recently, attention-based methods have been developed to capture long-range dependencies, like the Point Transformer (Zhao et al., 2021) and Swin3D (Yang et al., 2023). However, these methods continue to rely on MLPs and do not fully exploit geometric relationships, inheriting limitations from earlier approaches (Thomas et al., 2024).

2.3 Multimodal Segmentation with 2D-3D Fusion

Multimodal methods apply three main types of fusion: data fusion at the input level, feature fusion to combine extracted features, and output fusion to merge results (Ramachandram and Taylor, 2017). Early approaches focused on data fusion by converting point clouds into depth maps or Digital Surface Models (DSMs) for 2D segmentation (Wang et al., 2021; Diakogiannis et al., 2020; Cui et al., 2022). For 3D segmentation, the color information was projected through pixel-to-point mapping. However, the effectiveness of color information was proven limited compared to relying solely on geometry (Zhu et al., 2024).

Limited methods have explored the output fusion, primarily due to the limited availability of multimodal data and to avoid error propagation (Zhang et al., 2018). Recently, more methods have focused on feature fusion, where the features of each modality are encoded separately before merging to improve learning. However, most existing methods, such as MVPNet (Jaritz et al., 2019) and TransFusion (Maiti et al., 2023), address 2D or 3D segmentation individually rather than jointly processing them.

Only a few methods exist for joint segmentation, where 2D and 3D data are processed simultaneously by feature fusion, such as SplatNet (Su et al., 2018) and BPNET (Hu et al., 2021). However, these methods mainly use terrestrial and multiview datasets (de Gélis et al., 2021). In addition, data and feature fusion methods are computationally demanding due to larger input sizes and the complexity of processing multimodal features. In contrast, output fusion utilizes multimodal data efficiently, improving results through the strengths of separate learning without added computational cost.

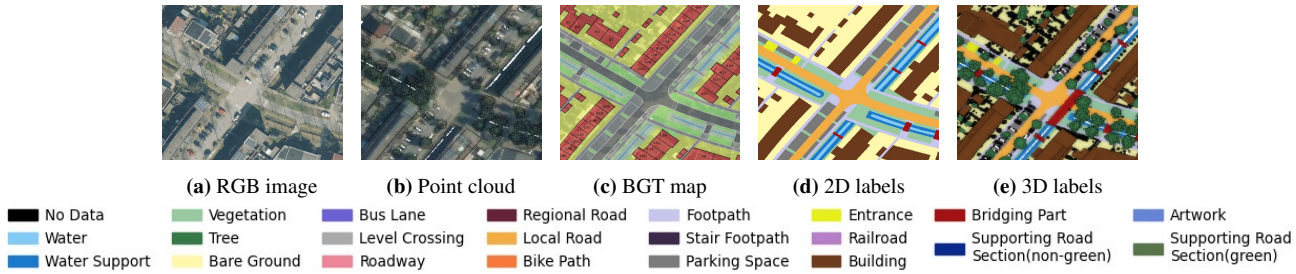


Figure 3. Example of RGB image, point cloud, BGT map, and derived 2D and 3D labels. The dataset contains labels for 22 semantic classes listed in the legend. The colors in the BGT map here are illustrative (for its legend, see Figure 1).

3. Dataset

As shown in Figures 3a-3c, we use airborne data, including 2D images and 3D point clouds with BGT maps as ground truth.

3.1 BGT Maps

These detailed topographic maps have scales ranging from 1:500 to 1:5000 and provide information for more than 30 physical objects (Geonovum, 2025). Each category is subdivided based on physical appearance or function, such as roads sub-categorized into motorways, provincial roads, bike paths, and railroads. The maps are available in vector formats and can be downloaded through the Web Feature Service (WFS) provided by the Publieke Dienstverlening Op de Kaart (PDOK). In addition, older versions of the maps can be generated retrospectively.

3.2 Airborne Data

The dataset includes georeferenced true orthoimages and point clouds across five regions, as summarized in Table 1. The orthoimages, generated by ESRI from data collected by Beeldmateriaal during the leafless season, contain RGB bands. The point clouds, sourced from Actueel Hoogtebestand Nederland (AHN) (Rijkswaterstaat, 2024), include attributes such as reflectance and class labels for ground, buildings, water, and civil structures. The dataset consists of 1200 tiles, with 2D images at 4000×4000 pixels and a 7.5 cm ground resolution. The 3D point cloud tiles cover 300m×300m, with an approximate density of 10 points per m^2. Additionally, height above ground was computed for the point clouds using PDAL (Contributors, 2024).

Region	Type	Coverage (%)
Deventer, Enschede	Urban, Suburban	70
Giethoorn, Wijhe	Rural, Forested	20
Sallandse Heuvelrug	Forested	10

Table 1. Overview of dataset coverage across study regions.

Using BGT information, airborne data are labeled for 22 semantic classes, as indicated in the legend of Figure 3. For 2D labels, maps are rasterized in a specific order, while rule-based PDAL pipelines are used for 3D to update the initial AHN labels. This process is consistently applied across all classes, except for trees, which are not labeled in 2D but are included in 3D. Since 3D labels combine AHN and map labels, they differ slightly for a few classes in 2D (see the bridging part class in Figures 3d and 3e).

4. Method

Figure 2 illustrates the workflow of our proposed joint 2D-3D semantic segmentation method. Synchronized 2D and 3D airborne data are processed in parallel through their respective segmentation networks for multimodal predictions. The 2D and 3D predictions obtained are then refined using the SLF module described in the upcoming Section 4.5.

4.1 Semantic Classes

Out of the 22 semantic categories in the dataset, several contextually similar classes were merged for this study. As a result, the final segmentation includes 11 classes for 2D and 12 for 3D, excluding the “No Data” class containing unlabeled objects. The classes **Bare Ground**, **Bike Path**, **Parking Space**, **Railroad**, **Building**, and **Bridge Part** remain the same.

The merged classes include **Water** (Water + Water Support), **Road** (Bus Lane + Level Crossing + Roadway + Regional Road + Local Road + Entrance), **Footpath** (Footpath + Stair Footpath), **Supporting Road Section** (Sup Road Sec - Green + Non-Green), and **No Data** (No Data + Artwork). As described in Section 3.2, the 2D data was acquired during the leafless season. Therefore, the tree and vegetation classes are merged into **Vegetation** in 2D, whereas in 3D, the **Tree** and **Vegetation** classes remain separate.

4.2 Data Preparation

The dataset is randomly divided into training, validation, and testing with a 60:20:20 ratio, and the data samples are chosen for each set. During training, a potential-based class sampling method is used to address class imbalances, as proposed by Thomas et al. (2019). Minor classes, like bridges and supporting road sections, are oversampled, while major classes, such as vegetation and buildings, are undersampled. For validation and testing, we use systematic grid sampling for evenly distributed samples to ensure complete tile coverage.

Fine-grained details are not essential for this task, so the images are downsampled by a factor of 2. For 3D, along with the XYZ geometry, reflectance and height above ground are used as additional attributes, which are clipped to a maximum of 95^{th}% of regional global values and normalized to the range [0, 1].

4.3 Joint Dataloader

This custom module handles the loading, augmentation, and batch creation of airborne data for 2D and 3D networks. First, for each sampled data point, a window is defined around its center to crop the corresponding image and point cloud. Then, augmentation techniques, such as flipping, scaling, and rotation, are applied to the cropped data to enhance the variability of the training data. Finally, the augmented data are organized into batches suitable for the corresponding network. The 3D network uses variable batch sizes to process points efficiently,

which is handled in the batch creation stage. For testing, tile and pixel indices are tracked to combine individual patch-wise predictions for tile-wise results.

4.4 Segmentation Networks

Both 2D and 3D networks follow the same architectural design, consisting of four encoder layers, four decoder layers, and a prediction head. In each network, the encoder and decoder are connected through skip connections in intermediate stages.

2D Network: We use a hybrid Transformer network with a pre-trained Swin Transformer as the encoder and a custom CNN-based decoder. In Stage 1 of the encoder, the input image is divided into non-overlapping patches, which are then tokenized and processed with linear embedding for feature extraction. In the subsequent three stages, the patch merging block downsamples the feature maps, and a series of Swin blocks increases the channel dimensions by a scale factor of 2 for hierarchical representations. Furthermore, multi-head self-attention modules within shifted windows enable the model to effectively capture local and global features (Liu et al., 2021).

Each decoder stage includes a weighted fusion block that combines skip connections with decoder features, followed by a decoder block that refines features through convolutions. The first decoder stage consists of only the decoder block. In the final stage, dynamically weighted spatial and channel attention is applied using a Feature Refinement Head, adapted from UNetFormer (Wang et al., 2022). A convolutional head produces dense patch-wise predictions for the image.

3D Network: We adopt KP-FCNN architecture using Kernel Point Convolutions (KPConv) to process point clouds directly (Thomas et al., 2019). Each encoder layer consists of two convolutional blocks, each containing a KPConv layer, followed by batch normalization and leaky ReLU activation. We used standard KPConv blocks instead of deformable ones to maintain simplicity, as outdoor point clouds do not exhibit complex geometries typically found in indoor settings.

In the decoder, the point-wise features are extracted using the nearest-upsampling method. These features are concatenated with intermediate encoder features via skip connections and refined with unary convolution. The segmentation head then applies unary convolution to predict per-point labels.

Loss Function: Compound loss functions have proven effective and robust for class-imbalanced segmentation (Ma et al., 2021). Therefore, we adopt an equally weighted combination of cross-entropy ($\mathcal{L}_{ce}$) and dice ($\mathcal{L}_{dice}$) losses for both 2D and 3D segmentation, to ensure a balance between pixel-wise or point-wise accuracy while addressing class imbalance. The total loss is formulated as follows:

$$\mathcal{L} = 0.5 \cdot \mathcal{L}_{ce} + 0.5 \cdot \mathcal{L}_{dice} \quad (1)$$

4.5 Selective Label Fusion (SLF)

Unlike conventional output fusion, which merges predictions indiscriminately, we use a semantic-driven approach to selectively transfer labels based on the strengths of each modality (2D or 3D). Classes are chosen based on the assumption that some are better represented in 2D due to color and contextual cues, while others are more effectively captured in 3D through geometric properties. Table 2 lists the selected classes and their label transfer directions.

Fusion	Semantic Classes
2D-to-3D	Road, Bike path, Footpath, Parking space, Supporting road section
3D-to-2D	Buildings
Both	Water, Bridging parts

Table 2. Selected classes for SLF categorized by fusion direction.

In this process, water and bridge classes benefit from bidirectional fusion: 2D-to-3D transfer enhances bridge segmentation using contextual learning, while 3D-to-2D transfer improves the spatial extent of water bodies. Additionally, 3D-to-2D transfer helps resolve occlusion issues where tree cover obstructs ground objects in 2D (Figure 1). Although trees are well-defined in 3D, they are not transferred to 2D for two main reasons. First, since the 2D data is acquired during the leafless season, delineating trees in 2D is ambiguous. Second, orthorectification introduces distortions, complicating their representation in 2D. Therefore, trees are retained exclusively in 3D.

We follow the workflow illustrated in Figure 4 for SLF. For each selected class, we generate georeferenced and topologically valid polygons that outline the boundaries of class segments in 2D and 3D predictions. Although 2D and 3D data are typically acquired around the same time, temporal discrepancies may still exist. To address this, geoprocessing operations, such as intersection and overlap, are applied to validate the generated polygons before transferring predictions.

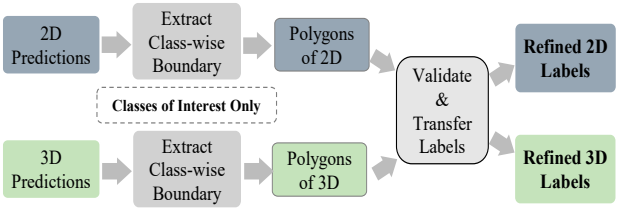
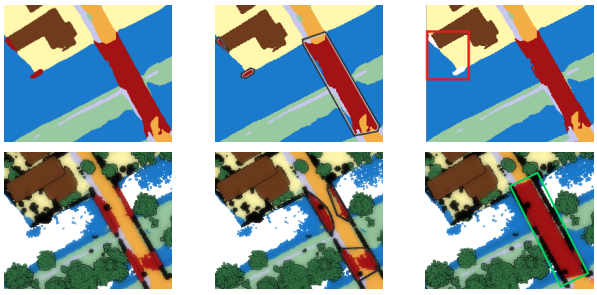


Figure 4. Workflow for Selective Label Fusion (SLF).



(a) Initial predictions **(b)** Bridge polygons **(c)** Refined labels

Figure 5. SLF for bridges in 2D (top) and 3D (bottom). The 2D and 3D polygons from (b) are intersected. Here, in 2D-to-3D transfer, overlapping polygons are used for label transfer (green box), while in 3D-to-2D transfer, segments appearing only in 2D without a 3D counterpart are removed (red box).

The 3D-to-2D transfer is carefully regulated to preserve the spatial boundaries of the initial 2D predictions. For buildings, only segments entirely missing in 2D predictions are transferred from 3D. For bridges, we intersect the generated 2D and 3D polygons, transferring only the missing 3D segments to 2D and

Data	SLF	Semantic Classes (IoU %)												mIoU (%)	mPrec (%)	mRec (%)
		Water	Veg	Tree	Bare Grnd	Road	Bike Path	Foot Path	Parking	Rail road	Sup Road	Building	Bridge			
2D	✗	79.4	90.3	-	69.0	75.1	62.4	56.1	50.4	93.5	47.8	85.1	62.4	70.1	83.9	79.5
	✓	80.2	90.4	-	68.9	75.1	62.7	56.1	50.4	93.5	47.8	84.9	67.2	70.7	84.1	80.1
3D	✗	52.2	85.0	98.4	49.9	67.5	34.1	36.9	40.5	68.3	6.5	95.7	26.3	55.1	60.5	83.6
	✓	67.5	87.3	98.4	51.5	72.4	62.6	58.0	45.9	92.0	48.9	96.2	60.1	70.1	79.2	84.2

Table 3. Class-level evaluation metrics across 2D and 3D domains with and without Selective Label Fusion (SLF) module, presenting IoU across all classes with mean IoU, precision, and recall for entire test data.

removing any non-overlapping 2D bridges as false positives (see Figure 5). For water bodies, overlap and intersection methods are inefficient due to possible occlusion by trees. Therefore, we directly transfer 3D water polygons to 2D but use a 100 m^2 area threshold to minimize noise and false positives.

For 2D-to-3D transfer, since most selected classes are ground-based (except bridges), filtering is applied to ensure that labels are transferred only to identified ground points. For bridges, validated polygons from the 3D-to-2D transfer workflow are used to refine labels accurately. PDAL pipelines are used to filter ground points and transfer labels from validated polygons.

5. Experiments and Results

We present joint segmentation results for 11 classes in both 2D and 3D, along with an additional tree class in 3D only. To evaluate the impact of label fusion, we compare results across two experiments: with and without the SLF module.

5.1 Implementation Details

For the 2D network, we adopt the standard window size, embedding dimension, depths, and attention heads across four stages of the Swin Transformer encoder, as proposed by Liu et al. (2021). The input patch size was set to 512x512, with three input channels (RGB) for both experiments. For the 3D network, grid subsampling with a grid size of 0.4m and input radius of 20m is applied to handle varying point cloud densities. Additionally, we use square sub-cloud tiles rather than spherical ones to achieve better alignment with 2D data.

Both networks were trained jointly on a dual GPU setup using NVIDIA A10, with the AdamW optimizer and the CosineAnnealingWarmRestarts learning rate scheduler (Loshchilov and Hutter, 2017, 2019). The 2D network was trained with a learning rate (Lr) of $5e^{-4}$, a weight decay of $2.5e^{-4}$, and a dropout rate of 0.2. For the 3D network, we used a Lr of $1e^{-3}$ and a weight decay of $1e^{-3}$, with gradient clipping. Both networks were trained for 400 epochs, using a step size of 400 and a batch size of 4. For the 2D network, gradient accumulation with a step size of 2 was implemented to simulate a batch size of 8 and Lr was adjusted every alternate epoch.

5.2 Evaluation Metrics

Intersection over Union (IoU) is used as the primary metric to measure the overlap between predictions and ground truth at both class and regional levels. Class-level IoUs evaluate per-class performance across the entire dataset, aggregating labels from all regions. In contrast, regional-level IoUs are computed separately for each region to assess the model's consistency across different geographical settings.

Given the abstract nature of the map-based test labels, precision and recall are also used for class-level evaluation. Precision quantifies the model's ability to minimize false positives, reducing misclassification errors. In contrast, recall measures the model's effectiveness in detecting all relevant features, even when the labels are abstract or simplified. The “No Data” class is excluded from all computed metrics.

5.3 Class-level Results

Table 3 presents class-level metrics for both experiments, with qualitative results shown in Figure 6. Without fusion, both 2D and 3D models perform well on distinct classes like buildings and vegetation, achieving IoU scores above 85%. In contrast, classes such as parking spaces and supporting road sections perform poorly, with around 45% IoU in 2D and even worse in 3D. In addition, the 2D network effectively differentiates visually similar classes like roads and bike paths, achieving an IoU greater than 60%. However, there is a 15% difference in performance between 2D and 3D without label fusion.

With the SLF module, the overall 3D performance improves substantially, with the fused mIoU exceeding 70%. In particular, supporting road sections and bridges show an approximately 40% increase in IoU. At the same time, bike paths and footpaths see improvements of around 20% and 30%, respectively, which can be noticed in Figure 6j. For 2D, the overall mIoU improves slightly, with a 5% increase in IoU for bridges; however, there is a minor 0.2% decrease in IoU for buildings.

Furthermore, recall rates remain high in both 2D and 3D, even without fusion, though 3D precision is nearly 20% lower than in 2D without fusion. However, with the SLF module, the high precision of 2D is effectively transferred to 3D, significantly improving 3D precision to 79% and reducing false positives.

Region	Count	No SLF mIoU (%)		SLF mIoU (%)	
		2D	3D	2D	3D
Deventer	104	72.0	56.5	72.7	71.3
Enschede	62	63.5	54.8	63.5	64.1
Giethoorn	38	60.5	43.9	60.1	60.5
Wijhe	12	69.1	56.4	69.2	67.5
S Heuvelrug	24	46.9	36.1	47.0	47.4

Table 4. Region-level 2D and 3D mIoU without and with SLF.

5.4 Region-level Results

Table 4 presents region-level IoU metrics with the number of tiles used for testing per region. As listed in Table 1, Deventer and Enschede represent urban and suburban regions, Giethoorn

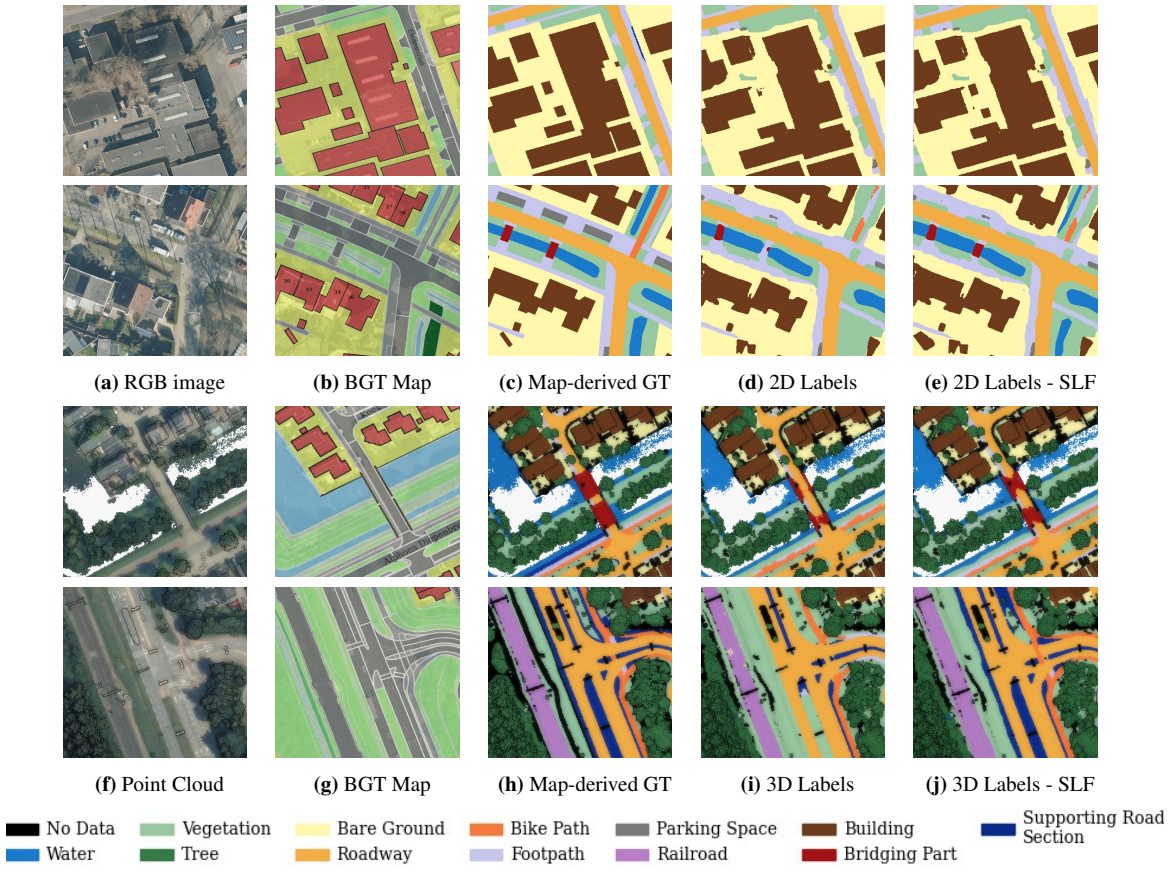


Figure 6. Qualitative results of 2D (a-e) and 3D (f-j) semantic segmentation with and without Selective Label Fusion (SLF) module. The BGT maps from which ground truth is derived are also compared.

and Wijhe are rural and marginally forested areas, and S Heuvelrug is predominantly forested.

Without fusion, the 2D model performs well in urban regions, achieving mIoU above 63%. Although 3D performance in these regions is lower than 2D, it remains the highest compared to other regions. Performance varies in rural and forested regions, with Heuvelrug achieving the lowest IoU in both 2D and 3D. With the SLF module, the 2D gains are minimal, with a slight increase in Deventer and Wijhe, but the 3D IoU improves by almost 10% in all regions, consistent with class-level results.

6. Discussion

The results in Table 3 show that the proposed method effectively identifies detailed map objects, achieving nearly 70% mIoU for 2D and 3D data. The proposed SLF module improves overall performance, especially in 3D, with a 15% improvement using 2D results. Label fusion for 3D is particularly beneficial for geometrically similar classes like bike paths and footpaths along with context-based classes such as supporting road sections and bridges. The higher recall rates in 2D and 3D indicate each model's effectiveness in capturing relevant objects within its modality, regardless of fusion. However, 3D precision is lower without fusion, likely due to the sparse nature of point clouds and the similarity among ground-based classes.

In general, both 2D and 3D networks struggle with classes that share overlapping spectral and geometric features, such as bike paths, footpaths, and roads, or vegetation and supporting road sections. Misclassifications often occur interchangeably among these classes. For example, the missing segments of the bike paths in Figures 6d and 6i are misclassified as footpaths or roads due to their overlapping features. Adopting a multi-label prediction approach could offer improved insight into these overlapping classes, unlike the current fixed single-label method.

Furthermore, while bridges possess distinct 3D geometries and are well represented in training data, their performance remains low in 3D. Most bridges in the dataset are medium-sized over narrow canals, but larger bridges that extend over larger water bodies are rare and often misclassified by the 3D model (see Figure 7). In our case, the 2D model accurately identifies larger bridges using contextual information, which is then used to refine 3D predictions via the SLF module. Future work could explore integrating similar contextual cues directly for the 3D network to improve performance for context-based cases.

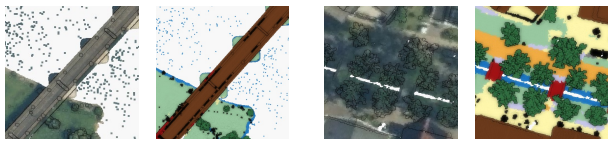


Figure 7. Example of point cloud and 3D predictions where (a) large bridge is misclassified as building and (b) small bridges are correctly identified.

Qualitative results further show that generalized map-derived test labels also contribute to reduced performance. For example, map-derived ground truth boundaries for buildings or water bodies are abstracted in Figures 6c and 6h. Although the 2D and

3D networks identify these objects, their prediction boundaries may not align exactly with the test labels. Such cases affect overlap-dependent IoU and precision metrics, potentially underestimating the true performance of 2D and 3D models.

In addition, the regional analysis in Section 5.4 shows that both models perform well in urban areas, moderately in rural areas, and poorly in forested regions. Urban areas, which make up 70% of the total dataset, would naturally have the highest visibility during training, allowing models to capture their features effectively. This suggests that a more strategic training approach could improve overall performance. For example, initial training on urban data and then fine-tuning on less-represented rural and forested regions may achieve better results than training on all regions simultaneously.

Despite the effectiveness of the SLF module for 3D, its contribution to 2D remains limited. In addition, the current fusion method is complex and rule-based, with multiple steps that require careful tuning. If not managed properly, this can lead to error propagation and the addition of false positives or negatives, as seen in the slight reduction in IoU for buildings in 2D.

Exploring alternative fusion methods, such as data and feature fusion, may be beneficial. However, data fusion is unlikely to improve segmentation in our case due to similar inter-class spectral and height features. In contrast, feature fusion using high-level features from both modalities can potentially improve the model’s ability to differentiate between similar classes, leading to more accurate predictions.

7. Conclusion

This study demonstrates the efficacy of BGT maps in automating multimodal segmentation for detailed map-based classes, performing well for 2D and 3D. The SLF module’s selective label transfer between 2D and 3D proves resourceful in addressing challenges like abstraction and occlusion in map-based labels, significantly improving 3D performance with minimal error propagation. Trained and tested on urban, rural, and forested regions, our method performs well in 4 out of 5 areas in both modalities, proving its robustness and adaptability.

Despite SLF’s benefits, the proposed fusion method remains complex, with limited impact on 2D segmentation and prone to error propagation. Future work could explore more integrated fusion techniques, such as using the insights of the SLF module for a custom feature fusion method and incorporating contextual modeling in 3D for better feature learning. Additionally, multi-label predictions may better handle overlapping classes than the strict single-label method. Our findings indicate that generalized map-derived labels can limit precision, highlighting the need for refined evaluation methods like soft metrics.

These findings highlight the potential of existing maps with DL to automate 2D-3D labeling across different geographic settings, offering a foundation to automate map creation and updating tasks that can be adapted for other systems like OpenStreetMap. In addition, this method can also be valuable for generating labeled datasets for other computer vision tasks involving airborne data beyond topographic mapping.

References

Beeldmateriaal, 2024. Beeldmateriaal. https://www.beeldmateriaal.nl/. Accessed: 2024-09-30.

Bello, S. A., Yu, S., Wang, C., 2020. Review: deep learning on 3D point clouds. arXiv:2001.06280.

Charles, R. Q., Su, H., Kaichun, M., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017-January, IEEE, 77–85. DOI:10.1109/CVPR.2017.16.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587.

Contributors, C., 2024. Point Data Abstraction Library (PDAL). https://github.com/PDAL/PDAL.

Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y., Cao, D., 2022. Deep Learning for Image and Point Cloud Fusion in Autonomous Driving: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 23, 722-739. DOI:10.1109/TITS.2020.3023541.

de Gélis, I., Lefèvre, S., Corpetti, T., 2021. Change Detection in Urban Point Clouds: An Experimental Comparison with Simulated 3D Datasets. *Remote Sensing*, 13, 2629. DOI:10.3390/rs13132629.

Diakogiannis, F. I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162, 94-114. DOI:10.1016/j.isprsjprs.2020.01.013.

Ding, L., Lin, D., Lin, S., Zhang, J., Cui, X., Wang, Y., Tang, H., Bruzzone, L., 2022. Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–13. DOI:10.1109/tgrs.2022.3168697.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Geonovum, 2025. Basisregistratie Grootschalige Topografie Gegegevenscatalogus. https://docs.geostandaarden.nl/.

Graham, B., Engelcke, M., van der Maaten, L., 2017. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. arXiv:1711.10275.

Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.-T., 2021. Bidirectional Projection Network for Cross Dimension Scene Understanding. arXiv:2103.14326.

Jaritz, M., Gu, J., Su, H., 2019. Multi-view pointnet for 3d scene understanding. *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, 3995–4003. DOI:10.1109/ICCVW.2019.00494.

Kaiser, P., Wegner, J. D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11), 6054–6068. DOI:10.1109/tgrs.2017.2719738.

Knudsen, T., Olsen, B. P., 2003. Automated Change Detection for Updates of Digital Map Databases. *Photogrammetric Engineering & Remote Sensing*, 69, 1289-1296. DOI:10.14358/PERS.69.11.1289.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J. D., Ledoux, H., 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 1, 11. DOI:10.1016/j.ophoto.2021.100001.

Li, Y., Fan, C., Wang, X., Duan, Y., 2021. SPNet: Multi-Shell Kernel Convolution for Point Cloud Semantic Segmentation. arXiv:2109.11610.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440. DOI:10.1109/CVPR.2015.7298965.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv:1608.03983.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A. L., 2021. Loss Odyssey in Medical Image Segmentation. *Medical Image Analysis*, 71, 102035. DOI:10.1016/j.media.2021.102035.

Maiti, A., Elberink, S. O., Vosselman, G., 2023. Transfusion: Multi-modal fusion network for semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6537–6547. DOI:10.1109/CVPRW59228.2023.00695.

PDOK, 2025. Geobasic registrations. https://www.geobasisregistraties.nl/.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. arXiv:1706.02413.

Ramachandram, D., Taylor, G. W., 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34, 96-108. DOI:10.1109/MSP.2017.2738401.

Rijkswaterstaat, 2024. Actueel Hoogtebestand Nederland (AHN). https://www.ahn.nl/. Accessed: 2024-09-30.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J. D., 2014. Theme section Urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93, 143-144. DOI:10.1016/j.isprsjprs.2014.04.009.

Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.-H., Kautz, J., 2018. SPLATNet: Sparse Lattice Networks for Point Cloud Processing. arXiv:1802.08275.

Tchapmi, L. P., Choy, C. B., Armeni, I., Gwak, J., Savarese, S., 2017. SEGCloud: Semantic Segmentation of 3D Point Clouds. arXiv:1710.07563.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L., 2019. Kpconv: Flexible and deformable convolution for point clouds. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 6410–6419. DOI:10.1109/ICCV.2019.00651.

Thomas, H., Tsai, Y.-H. H., Barfoot, T. D., Zhang, J., 2024. KPConvX: Modernizing Kernel Point Convolution with Kernel Attention. arXiv:2405.13194.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Vorarlberg, S. G., 2024. Vorarlberg-3D. https://vorarlberg.at/.

Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X., 2021. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sensing*, 13(16). DOI:10.3390/rs13163065.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P. M., 2022. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 196-214. DOI:10.1016/j.isprsjprs.2022.06.008.

Widyaningrum, E., Bai, Q., Fajari, M. K., Lindenbergh, R. C., 2021. Airborne Laser Scanning Point Cloud Classification Using the DGCNN Deep Learning Method. *Remote Sensing*, 13(5). DOI:10.3390/rs13050859.

Yang, Y.-Q., Guo, Y.-X., Xiong, J.-Y., Liu, Y., Pan, H., Wang, P.-S., Tong, X., Guo, B., 2023. Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding. arXiv:2304.06906.

Yang, Z., Jiang, W., Lin, Y., Elberink, S. O., 2020. Using Training Samples Retrieved from a Topographic Map and Unsupervised Segmentation for the Classification of Airborne Laser Scanning Data. *Remote Sensing*, 12(5). DOI:10.3390/rs12050877.

Zhang, R., Li, G., Li, M., Wang, L., 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 143, 85-96. DOI:10.1016/j.isprsjprs.2018.04.022.

Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V., 2021. Point transformer. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 16239–16248. DOI:10.1109/ICCV48922.2021.01595.

Zhu, H., Wang, Y., Huang, D., Ye, W., Ouyang, W., He, T., 2024. Point Cloud Matters: Rethinking the Impact of Different Observation Spaces on Robot Learning. arXiv:2402.02500.