ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-3-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission III, 4–11 July 2026, Toronto, Canada

Learning from Maps to Update Them: A Deep Learning-Based Approach Using Multimodal Airborne Data

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

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

Keywords: Topographic Maps, Airborne Data, Semantic Segmentation, Change Detection, Multimodal data

Abstract

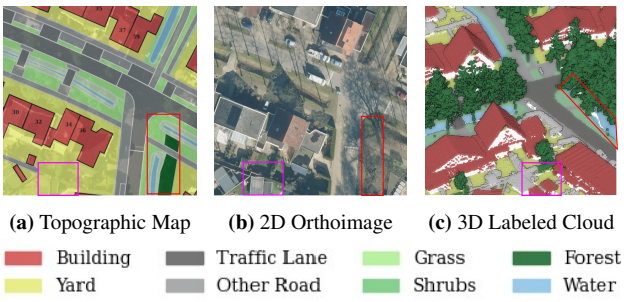
Automatic updating of topographic maps remains a significant challenge, as current workflows still rely heavily on manual interpretation of airborne data. This study proposes a method for identifying topographic changes by learning object representations from existing maps and using them as reference data for change detection. Map-derived labels are used to train independent 2D and 3D segmentation networks that generate semantic predictions from orthoimages and point clouds. Unlike conventional change-detection approaches that require temporally aligned datasets of the same modality, the proposed method directly compares newly acquired airborne data with existing map vectors. Semantic predictions from both modalities are vectorized and selectively fused into polygon geometries, which are subsequently compared with reference map vectors to identify object-level “from–to” changes. The workflow highlights potential change regions and their predicted semantic classes, allowing operators to focus inspection on relevant areas rather than the entire dataset. Detected changes include both real-world developments, such as new construction and demolitions, and inconsistencies in the reference map caused by outdated or inaccurate delineations. To assess the effect of multimodal integration, the workflow is compared with a 2D-only baseline. The results indicate that integrating 3D geometric information can reduce noisy detections and improve the spatial consistency of candidate change objects, particularly for water and bridge classes.

1. Introduction

Topographic maps provide structured information about real-world objects and are essential for urban planning, infrastructure management, and environmental monitoring. In the Netherlands, the Basisregistratie Grootschalige Topografie (BGT) serves as the authoritative nationwide large-scale topographic database (Kadaster, 2025). Although the BGT contains detailed information, it is maintained collaboratively by multiple governmental agencies and updated primarily through manual inspection of airborne images and point clouds. This process is labor-intensive, time-consuming, and prone to subjective interpretation in both semantic labeling and spatial delineation (Knudsen and Olsen, 2003).

Recent advances in deep learning (DL) have demonstrated the potential for automating change detection by learning spectral and spatial patterns directly from geospatial data to identify *from–to* changes. However, most DL-based approaches rely on temporally aligned datasets acquired under similar conditions and from the same modality (Agarwal et al., 2019; Zhu et al., 2022). In addition, they require extensive labeled samples representing all possible *from–to* change categories. These requirements limit the applicability of DL methods for regularly updating large-scale topographic maps (Shafique et al., 2022).

Existing topographic maps provide an alternative source of supervision that can mitigate these limitations. First, they allow DL models to learn classification rules consistent with the cartographic standards used by human mapmakers that can be applied to newly acquired airborne data (Kaiser et al., 2017; Yang et al., 2020; Widyaningrum et al., 2021; Anjanappa et al., 2025). Second, maps naturally represent the historical state of the landscape and can therefore serve as reference data for identifying changes. Previous studies, such as Girard et al. (2019), Niroshan and Carswell (2022), and Zorzi et al. (2020), have used maps to detect simple binary changes, primarily focused on building construction or demolition. However, directly using maps as reference data for multi-class semantic change detection introduces new challenges.



(a) Topographic Map **(b)** 2D Orthoimage **(c)** 3D Labeled Cloud

Figure 1. Ambiguous objects in 2D (red box) are clearly captured in the 3D, providing better correspondence with the map.

Maps are abstract representations of reality and may not fully capture the visual or geometric complexity observed in airborne data. For example, the red box in Figure 1b highlights an area covered by tree canopies in the 2D image, while the corresponding region in the map (Figure 1a) is labeled as roads, water, and shrubs. A direct comparison between these data sources would incorrectly indicate a change. However, 3D point clouds provide geometric information that helps resolve such ambiguities, as illustrated in Figure 1c. Similarly, when new structures, such as the building highlighted in the pink box, appear in newly acquired airborne data, combining 2D appearance with 3D geometry enables more reliable change detection.

In this study, we adapt a post-classification change detection strategy for vector–to–vector comparison. Unlike conventional post-classification approaches that operate at the pixel (2D) or

* Corresponding author: g.anjanappa@utwente.nl

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-1-2026

point (3D) level, vector-based comparison enables object-level change detection, which is better suited for topographic map updating. Furthermore, instead of relying on temporally aligned datasets of the same modality, the proposed approach compares semantic predictions derived from newly acquired airborne data with an existing topographic map. In this way, existing maps are used both as supervision for training 2D and 3D semantic segmentation models and as reference data for detecting changes.

Following our previous work, we jointly train independent 2D and 3D networks using BGT-derived labels to predict object classes from newly acquired airborne data (Anjanappa et al., 2025). The semantic predictions from both modalities are first vectorized and then selectively fused to generate polygon geometries consistent with the BGT schema. These polygons are then compared with the reference map to identify object-level “from–to” changes. To assess the impact of multimodal information, the proposed workflow is compared with a baseline using only 2D predictions.

The proposed approach aims to identify candidate change regions corresponding to (i) real topographic changes resulting from development activities such as new constructions, demolitions, and land-use changes, and (ii) inconsistencies or outdated features in the reference map. By localizing change regions and their updated categories, the workflow reduces the extent of manual inspection required and supports faster, more consistent updates to large-scale topographic maps.

2. Related Works

In recent years, DL-based semantic segmentation and change detection have been widely applied in remote sensing. This section reviews recent advances in semantic segmentation for airborne data, followed by traditional and map-based change detection approaches relevant to topographic maps.

2.1 Semantic Segmentation

Semantic segmentation assigns semantic labels to objects in airborne images or point clouds, delineating meaningful objects.

2D Image Segmentation: For dense pixel-wise segmentation, Convolutional Neural Networks (CNNs) such as UNet (Ronneberger et al., 2015) and DeepLabV3 (Chen et al., 2017) have been widely used. However, their limited receptive fields restrict the ability to capture long-range contextual information. More recently, Transformer-based models such as ViT (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021) address this limitation using self-attention mechanisms to model global context (Vaswani et al., 2017). Hybrid architectures that combine CNNs and Transformers have further improved segmentation performance in remote sensing applications (Wang et al., 2022; Yamazaki et al., 2023).

3D Point Cloud Segmentation: Early methods converted point clouds into intermediate structures such as grids or voxels for segmentation, often leading to information loss (Tchapmi et al., 2017; Graham et al., 2017). Later, point-based methods such as PointNet and PointNet++ (Qi et al., 2017; Charles et al., 2017) directly processed raw point clouds using Multi-Layer Perceptrons (MLPs) to capture local geometric patterns. Subsequently, point convolution methods such as KPConv (Thomas et al., 2019) improved local feature learning, while transformer-based models such as Point Transformer (Zhao et al., 2021) further improved global scene understanding. Despite these advances, many 3D methods remain computationally intensive and sensitive to variations in point density.

2D–3D Fusion for Segmentation: Multimodal fusion can be performed at the data, feature, or output level (Ramachandram and Taylor, 2017). Input-level approaches either convert 3D point clouds into depth maps or Digital Surface Models (DSMs) for 2D segmentation, or project color information from images onto point clouds to support 3D segmentation (Diakogiannis et al., 2020; Cui et al., 2022). In contrast, output-level fusion combines predictions from independently trained 2D and 3D models using rule-based methods or meta-classifiers (Rudner et al., 2019; Anjanappa et al., 2025). More recent work has focused on feature-level fusion, where 2D and 3D features are encoded separately before being combined for semantic labeling (Hu et al., 2021; Maiti et al., 2023).

2.2 Change Detection

Change detection identifies variations in objects or land cover between data acquired at different times (Singh, 1989). It can be either binary, indicating only the presence of change, or semantic, which further characterizes the type of change.

Existing DL-based methods can be broadly categorized into two groups. Metric-based approaches detect changes by comparing features extracted from multi-temporal data using distance measures (Zhan et al., 2017). Classification-based methods instead treat change detection as a dense classification task, either by comparing independently generated classification maps (Ji et al., 2019; Shafique et al., 2022) or by jointly learning from multi-temporal inputs using labeled from–to samples (Daudt et al., 2018; Rahman et al., 2018).

More recent studies have improved robustness using Siamese or dual-branch architectures (Chen et al., 2021; de Gélis et al., 2023). However, most DL-based methods are limited to comparisons within the same modality (image–image or cloud–cloud) and require extensive labeled *from-to* change samples.

Using Maps and Airborne Data: Only a few studies have explored change detection using existing maps as prior reference data. Girard et al. (2019) used CNNs to refine building boundaries and detect new or missing buildings from airborne imagery. Similarly, generative model-based methods such as OSM-GAN (Niroshan and Carswell, 2022) and MapRepair (Zorzi et al., 2020) generate map-like representations from images before detecting binary building changes. Although effective, these approaches are computationally demanding and inefficient for complex geometries (Albrecht et al., 2020).

Since maps and airborne data differ in modality, scale, and abstraction, directly applying conventional DL-based change detection methods remains challenging. Furthermore, existing approaches are largely limited to building detection and do not generalize to multiple semantic classes. These studies, however, demonstrate the potential of using existing maps as both supervision and reference data for multi-class semantic change detection, enabling the identification of “from–to” changes.

3. Dataset

The dataset used in this study includes two components: (1) BGT maps, which serve as both ground truth for semantic seg-

mentation and reference data for change detection; and (2) airborne 2D orthoimages and 3D point clouds. The data are derived from the Map2ImLas dataset (Anjanappa et al., 2026).

BGT Maps: BGT provides standardized large-scale topographic maps that delineate more than 30 object types, which are further subdivided according to physical characteristics or functional categories (Geonovum, 2025). For example, roads are subdivided into highways, regional roads, footpaths, and bicycle paths. The maps are available in vector format and can be accessed retrospectively for the year of interest through Publieke Dienstverlening Op de Kaart (PDOK).

Airborne Data: Obtained in 2022, the airborne data include 2D orthoimages and 3D point clouds from two regions in the Netherlands: Deventer and Enschede. In total, the study area covers approximately 75 km^2 and includes urban, suburban, and industrial environments. The datasets are georeferenced and spatially aligned to ensure consistent analysis across modalities.

True orthoimages with RGB bands are provided by Esri based on aerial images collected by Beeldmateriaal. The point clouds are obtained from the Actueel Hoogtebestand Nederland (AHN) (Rijkswaterstaat, 2024). They include reflectance and intensity with additional Height-Above-Ground (HAG) features derived using the Point Data Abstraction Library (PDAL) (Contributors, 2024). Each 2D tile covers 4000×4000 pixels with a ground sampling distance of 7.5cm, while the corresponding point cloud tile covers 300m×300m with a density of approximately 10 points per m^2. Both 2D and 3D data are collected during the leafless season (January-March) to minimize vegetation occlusion.

Semantic Classes: The dataset includes 10 semantic classes derived from BGT object types. Nine classes are commonly used for both 2D and 3D segmentation, while the **Tree** class is exclusive to 3D due to its distinct geometric features in point clouds. The selected classes represent natural surfaces and built structures commonly found in urban and rural environments. Table 1 summarizes the class definitions.

Following the approach of Anjanappa et al. (2026), semantic labels were generated for all classes using BGT vector data. For this study, BGT maps were retrospectively rendered for March, 2022 to approximately match the date of airborne data acquisition. For 2D labeling, vector maps were rasterized using a predefined class hierarchy. For 3D labeling, rule-based PDAL pipelines were used to transfer class labels to the point clouds.

Class	Description
Water	Rivers, canals, and other water bodies
Vegetation	Low vegetation such as grass or shrubs
Tree*	Tall vegetation with a distinct canopy
Bare Ground	Unpaved open areas, such as soil or gravel
Road	Paved vehicular routes, streets, bicycle paths, and sidewalks
Parking Space	Areas specifically for vehicle parking
Railroad	Railway tracks and associated paved corridors
Building	Roofed man-made structures
Bridge	Elevated road or walkway segments
No Label	Any object not belonging to the above categories

Table 1. Semantic classes and their definitions. The **Tree*** class appears only in 3D segmentation.

4. Method

Figure 2 illustrates the proposed workflow for map-based change detection. Airborne data are processed using independent 2D and 3D semantic segmentation networks to generate their respective predictions. The resulting prediction maps are vectorized and subsequently compared with reference map vectors using a rule-based procedure described in Section 4.3.

4.1 Data Preparation

The dataset contains 835 tiles, which were randomly split into training, validation, and test subsets at a 3:1:1 ratio. For semantic segmentation, the 2D network uses RGB values as input, while the 3D network uses geometric coordinates, reflectance, and Height-Above-Ground (HAG) features. Reflectance and HAG values were clipped at the 95th percentile and normalized to the range [0, 1]. The **No Label** class was excluded from evaluation.

Since no ground truth for map changes was available, the test set was manually inspected to identify tiles that likely showed changes. These include both real topographic modifications,

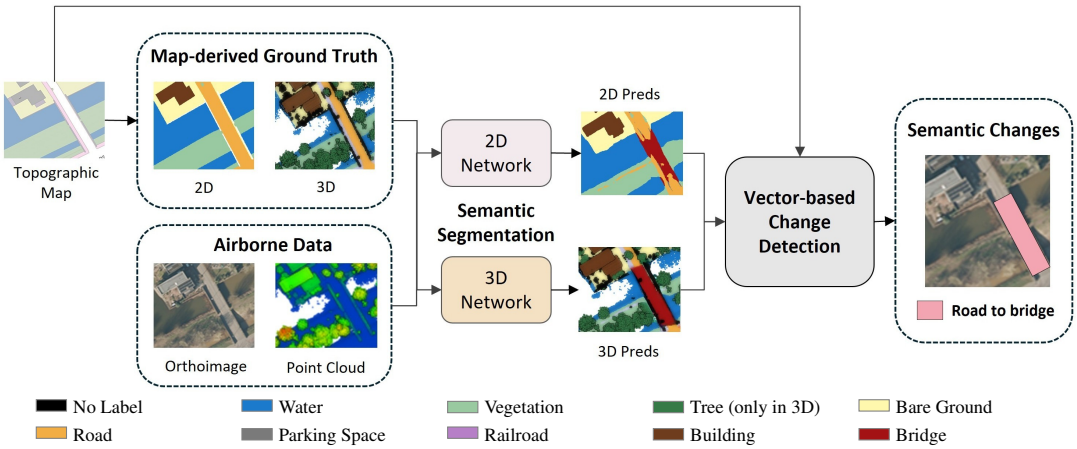


Figure 2. Overview of the proposed workflow. 2D orthoimages and 3D point clouds are processed independently using dedicated segmentation networks. Their predictions are fused and vectorized for change detection. In this example, areas mapped as roads in the BGT are predicted as bridges, indicating a probable topographic change.

such as new constructions or demolitions, and inconsistencies in the BGT reference map. Because the 2D and 3D datasets are already co-registered with the BGT polygons, no additional spatial alignment was required (Anjanappa et al., 2026).

4.2 Semantic Segmentation

Both the 2D and 3D networks follow an encoder–decoder architecture with skip connections linking intermediate stages.

2D Image Segmentation: For 2D segmentation, AerialFormer (Yamazaki et al., 2023), a hybrid Transformer–CNN architecture is used. The model includes three components: a CNN stem, a Transformer encoder, and a Multi-Dilated CNN (MDC) decoder. The CNN stem preserves fine-grained spatial details from the input image, while the Transformer encoder captures long-range contextual dependencies through self-attention. The MDC decoder then fuses multi-scale features from the encoder and skip connections using parallel dilated convolutions, effectively combining local detail and global context for accurate semantic segmentation.

3D Point Cloud Segmentation: For 3D segmentation, the Kernel Point Convolution–based fully convolutional network (KP-FCNN) is used (Thomas et al., 2019). Each encoder layer consists of two KPConv blocks, each followed by batch normalization and a leaky ReLU activation. Since outdoor point clouds typically exhibit simpler geometric structures, only standard KPConv blocks are used for computational efficiency. Decoder features are reconstructed through nearest-neighbor upsampling and fused with encoder features via skip connections. A final unary convolution layer predicts per-point semantic labels.

4.3 Vector-Based Change Detection

Conventional post-classification change detection compares pixel-wise predictions across epochs in the raster domain. However, such differencing is less suitable for cartographic applications, where topographic map updates require object-level interpretation. Therefore, we adopt a vector–to–vector workflow in which semantic predictions are converted into polygons and directly compared with the reference topographic map to identify object-level changes.

To integrate multimodal information, semantic predictions from the 2D and 3D networks are first vectorized independently. The resulting polygons are then fused using the Selective Label Fusion (SLF) method proposed by Anjanappa et al. (2025). Since BGT updates are represented in the 2D map domain, 3D-to-2D fusion is performed to include 3D structural information, correcting or supplementing 2D-derived polygons while preserving accurate boundaries. The resulting fused polygons remain topologically consistent with the BGT schema and form the basis for detecting “from–to” changes.

This study focuses on the **Water, Road, Parking Space, Railroad, Building**, and **Bridge** classes. Changes involving the *No Label* category and seasonal transitions between *Vegetation* and *Bare Ground* are excluded. During fusion, missing 2D segments for buildings and water are supplemented using polygons derived from 3D predictions. For bridges and parking spaces, 2D and 3D polygons are intersected to retain overlapping regions, add missing 3D components, and remove isolated 2D false positives. For roads and railroads, 3D information is used only to fill local gaps in the 2D predictions.

Finally, the fused vector map is compared with the reference BGT map using a rule-based procedure. The “from” label is obtained from BGT polygons, while the “to” label is derived from the fused predictions. Class-wise change detection is performed using spatial joins between reference and predicted polygons. Polygons without spatial intersections are marked as *missing* or *new*. To reduce noisy detections, class-specific minimum-area thresholds (Table 2) are applied before reporting changes.

Class	Area Threshold (m^2)
Building	30
Bridge	5
Water	50
Railroad	30
Road	40
Parking Space	20

Table 2. Class-specific minimum area thresholds used to suppress noisy polygons.

Among the detected results, non-overlapping polygons indicate potential changes. Polygons present in the BGT map but absent in the predictions suggest possible demolitions or outdated map objects, whereas polygons appearing only in the predictions indicate potential new constructions or inconsistencies in the reference map. Figure 3 illustrates an example of multimodal change detection using airborne data and the BGT map.

5. Experiments and Results

This section presents the implementation details and evaluation of the proposed framework. The experiments evaluate semantic segmentation performance for both modalities and compare change-detection results obtained using a 2D-only workflow with those from the proposed fused 2D+3D workflow.

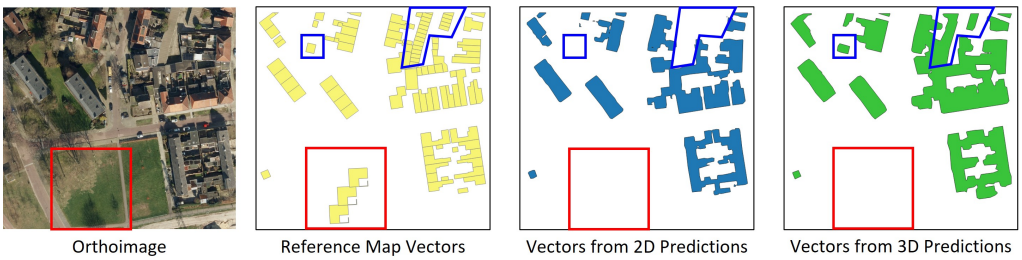


Figure 3. Example of multimodal building change detection. The orthoimage shows the current scene and the BGT map provides previous building footprints. Red boxes indicate demolished buildings absent in both 2D and 3D data, while blue boxes show areas missing in 2D but confirmed by 3D, illustrating how multimodal fusion reduces false detections.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-1-2026 | © Author(s) 2026. CC BY 4.0 License.

5.1 Implementation Details

The 2D segmentation model used a pretrained Swin backbone from MMSegmentation (Contributors, 2020). It was trained for 160K iterations using cross-entropy loss, with an input patch size of 512×512 and a batch size of 8. Data augmentation included random resizing (scale factor [0.9, 1.1]), cropping, flipping, and color adjustments. The model was optimized using AdamW (Loshchilov and Hutter, 2017) with a base learning rate of 6×10^{-5}, weight decay of 0.01, and a two-phase schedule: Linear Warmup (1500 iterations) followed by Polynomial Decay (PolyLR).

For 3D segmentation, a predefined KPFCNN architecture with simple convolution blocks was used with geometry, normalized reflectance, and HAG input features. The input radius was set to 10m with a subsampling grid size of 0.4m. The network was trained for 500 epochs using an SGD optimizer with epoch-based decay and an initial learning rate of 0.01. A class-balanced sampling strategy was applied during training, while potential-based sampling ensured spatial uniformity during inference. Data augmentation included anisotropic scaling, random rotations, and noise addition.

The other workflows using geoprocessing operations were implemented using readily available Python libraries, including GeoPandas, Shapely, Alphashape, and Rasterio. All spatial operations were performed in EPSG:28992 (Amersfoort / RD New) coordinate reference system to maintain geometric consistency with the BGT data.

5.2 Evaluation Metrics

Quantitative evaluation is conducted for semantic segmentation, while change detection is assessed qualitatively due to the absence of verified change samples. The evaluation aims to assess the framework’s ability to identify meaningful topographic changes and inconsistencies in map data.

Semantic Segmentation: Mean Accuracy (mAcc) and Intersection over Union (IoU) are used to evaluate segmentation performance and are defined as:

$$\mathrm{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c} \tag{1}$$

$$\mathrm{mAcc} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{2}$$

where TP_c, FP_c, and FN_c denote the true positive, false positive, and false negative pixel (or point) counts for class c, and C is the total number of evaluated classes.

Change Detection: To assess the effect of multimodal fusion, change-detection results from the fused workflow are compared with a baseline using only 2D predictions. For each class, we report the number of tiles in which candidate changes are detected, together with the counts and total areas of objects that are either missing from the predictions or newly appearing relative to the BGT map. These quantities represent candidate changes rather than verified changes and therefore do not constitute quantitative accuracy measures. Instead, they indicate the magnitude and spatial distribution of differences identified by the workflow. Since no authoritative ground truth for map changes is available, visual inspection is used to qualitatively assess the plausibility of the detected changes.

5.3 Semantic Segmentation Performance

Table 3 and Figure 4a present the semantic segmentation results for the selected topographic classes. Both networks achieve comparable mean IoU values of approximately **75%**. The 2D model performs well for visually distinctive surface classes such as **water** (78.5%), **vegetation** (88.2%), and **roads** (76.9%), benefiting from high-resolution color information and contextual cues. In contrast, the 3D network achieves higher accuracy for structurally distinctive classes such as **trees** (98.5%) and **buildings** (98.2%), where elevation and geometric features provide strong discriminative signals.

Although smaller object-based classes such as **parking spaces** and **bridges** show lower IoU values, both modalities maintain reliable performance for large surface categories and man-made structures. Overall, the results highlight the complementary strengths of the two modalities: 2D data captures appearance-based cues, while 3D data provide structural information, supporting their combined use for change detection.

5.4 Vector-Based Change Detection Results

Table 4 and Figure 4c summarize the detected changes obtained from the 2D-only predictions and the fused 2D+3D workflow. Overall, candidate changes were identified in **154 tiles** using the 2D-only predictions and in **148 tiles** after multimodal fusion. This slight reduction in the number of tiles containing changes suggests that incorporating 3D information helps suppress isolated or fragmented detections. Compared with the 2D-only approach, the multimodal workflow reduces the number of missing objects (from 417 to 356) while identifying more new objects (from 657 to 760).

For both experiments, a large number of candidate changes occur for the **Water**, **Road**, **Parking Space**, and **Building** classes. These classes also account for the largest affected areas due to their larger spatial extent. Notably, the **Water** class shows a substantial increase of 23K m^2 in the newly detected area in the multimodal workflow compared with the 2D-only approach, while the missing water area decreases from 33K m^2 to 11K m^2. Similarly, the **Building** class shows an increase of approximately 6K m^2 in the newly detected area. These observations indicate that multimodal fusion primarily influences classes characterized by larger spatial structures or stronger geometric features.

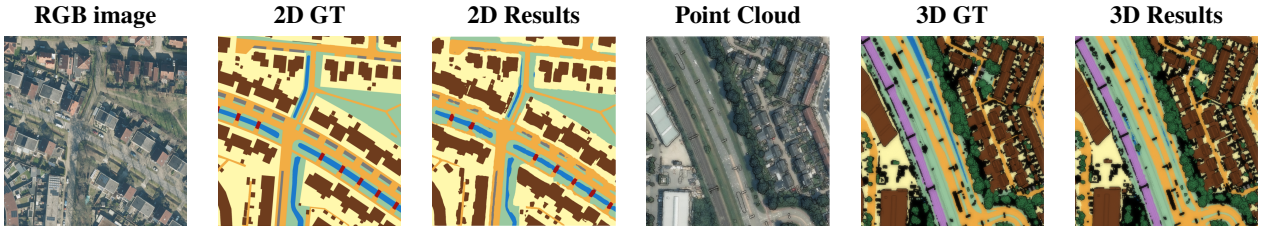
Data	Map-derived Classes (IoU %)									mIoU	mAcc
	Water	Vegetation	Tree	Bare Ground	Road	Parking	Railroad	Building	Bridge	(%)	(%)
3D	63	86.5	98.5	62.9	71.5	39.2	87.8	98.2	65	74.8	91.2
2D	78.5	88.2	-	68.4	76.9	47.1	91.4	86.7	59.6	74.6	83.8

Table 3. Semantic segmentation results for 2D and 3D domains with class-wise IoU, mean IoU, and mean accuracy (mAcc).

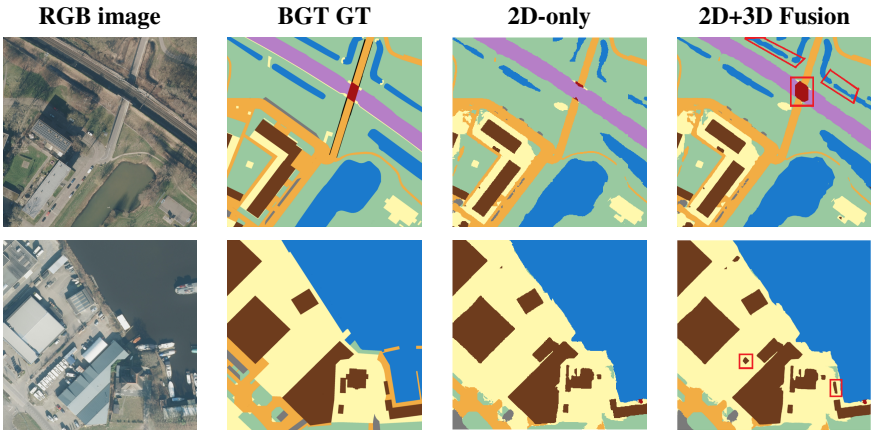
This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-1-2026 | © Author(s) 2026. CC BY 4.0 License.

Class	Changes - 2D Only					Changes - 2D + 3D				
	Tiles	Missing Objects		New Objects		Tiles	Missing Objects		New Objects	
		Count	Area (m^2)	Count	Area (m^2)		Count	Area (m^2)	Count	Area (m^2)
Water	90	96	33K	109	18K	90	22	11K	185	41K
Road	110	53	8K	191	28K	108	54	8K	186	29K
Parking	82	184	16K	218	22K	82	222	17K	203	21K
Railroad	0	0	0	0	0	0	0	0	0	0
Building	75	65	6K	109	15K	88	39	5K	168	21K
Bridge	29	19	397	30	1K	23	19	397	18	1K
Total	154	417	63K	657	84K	148	356	41K	760	113K

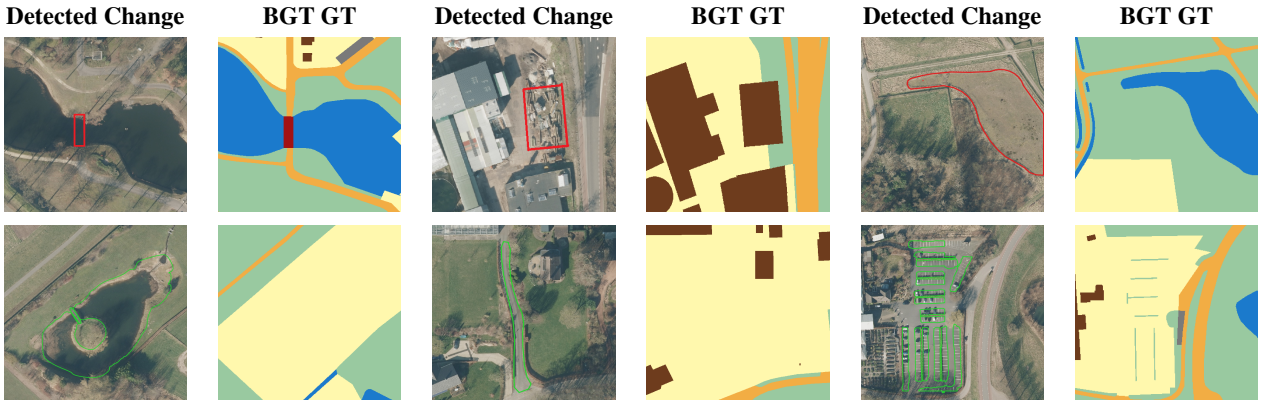
Table 4. Class-wise summary of detected topographic changes, indicating the number of tiles affected, the counts of missing and newly added objects, and their corresponding affected areas in square meters (m^2).



(a) Semantic segmentation results for the 2D and 3D networks.



(b) Example comparison of 2D-only and fused 2D+3D predictions used for change detection step. Highlighted regions in red boxes indicate cases where including 3D information improves object delineation.



(c) Detected changes (left) and BGT reference (right). Red polygons outline missing objects, and green polygons delineate new objects from the proposed workflow. Examples include water, buildings, roads, and parking areas.

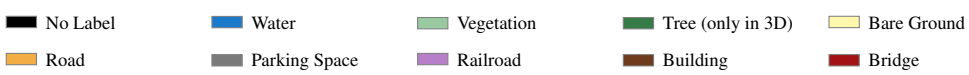


Figure 4. Qualitative results demonstrating multimodal semantic segmentation (a) and vector-based change detection (b).

https://doi.org/10.5194/isprs-annals-XI-3-2026-1-2026

The results also show variations between object counts and affected areas. For example, **Parking Space** generates a large number of candidate objects but contributes comparatively smaller areas, whereas **Water** bodies account for fewer objects but substantially larger change areas. In contrast, the smaller object-based **Bridge** class contributes relatively small change areas due to its limited spatial footprint. Although the 2D-only workflow detects more new bridges, many correspond to small noisy segments that are not preserved after multimodal fusion. Furthermore, the **Railroad** class shows no detected changes.

Manual inspection of selected tiles indicates qualitative differences between the two approaches. In several cases, the 2D-only predictions produce fragmented or patchy detections caused by local segmentation errors, particularly for the **Water** and **Bridge** classes. Incorporating 3D structural information helps reduce such artifacts and produces more spatially coherent objects, as illustrated in Figure 4b. For example, the 2D-only workflow detects more new bridges (30) than the multimodal workflow (18), but many of these detections correspond to small fragmented segments incorrectly labeled as bridges.

The values reported in Table 4 represent **candidate changes** derived from differences between predicted objects and the reference BGT map. Not all detected changes correspond to real-world modifications; some arise from segmentation errors, incomplete delineations, or differences in representation between the predicted objects and the BGT itself.

6. Discussion

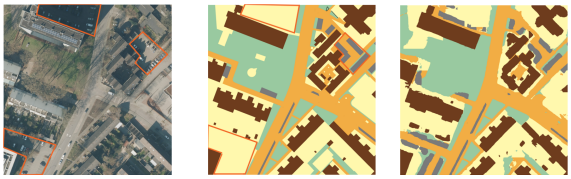
The results in Section 5.4 demonstrate that the proposed vector-to-vector workflow can highlight object-level changes between airborne data and the reference map. While the overall number of detected changes remains comparable between the 2D-only and fused 2D+3D workflows, the multimodal workflow helps reduce fragmented detections and produce more spatially coherent polygons. However, polygon-level change detection requires accurate boundaries and topological consistency. As a result, the reliability of detected changes strongly depends on the quality of semantic segmentation and the robustness of the vectorization, fusion, and comparison stages.

Segmentation performance strongly influences the accuracy of change detection. Classes with high IoU, such as **Buildings** and **Roads**, produce cleaner polygons and lead to more stable change candidates. In contrast, lower-accuracy classes such as **Parking Space** tend to produce noisy predictions, leading to unreliable detections. The comparison between the 2D-only and multimodal workflows suggests that incorporating 3D structural information can reduce such noise in several cases, particularly for **Water** and **Bridge** classes. However, these improvements are not uniform. Erroneous 3D predictions can override correct 2D outputs during the fusion stage, as illustrated in Figure 5. This highlights an inherent trade-off: while 3D information provides valuable geometric context, it may also propagate errors when the 3D segmentation is inaccurate, potentially introducing false positives or missed detections.

(a) Image **(b)** 2D Preds **(c)** 3D Preds

Figure 5. Comparison of 2D and 3D predictions, where incorrect 3D predictions can override correct 2D results when fused, potentially indicating false changes.

The use of BGT maps in both the segmentation and change detection processes introduces additional limitations. Although the BGT provides detailed annotations, its mapping conventions may lead to segmentation errors and irrelevant candidate changes across several classes. For example, private areas are not explicitly mapped in the BGT and are often represented as bare ground, as illustrated in Figure 6a. When the workflow segments objects within these regions, as shown in Figure 6c, the comparison with the BGT map identifies them as candidate changes. Although several detected changes correspond to visually identifiable structures in the airborne data, distinguishing them from changes caused by BGT mapping conventions remains challenging.



(a) RGB Image **(b)** BGT GT **(c)** 2D Predictions

Figure 6. Example of label inconsistencies due to BGT mapping conventions (orange box). Private spaces are often represented as bare ground, although they visually contain a mix of elements, such as parking areas, internal roads, and other features, which are identified correctly in semantic predictions.

Another limitation of this study is the absence of a verified ground truth for map changes. Consequently, the detected differences represent candidate change regions rather than validated updates. While manual inspection indicates that many highlighted changes correspond to plausible real-world changes or inconsistencies in the reference map, a systematic quantitative evaluation requires temporally verified change annotations. Developing benchmark datasets with validated map updates would enable more rigorous quantitative evaluation of map-based change detection methods.

Despite these limitations, the results indicate that combining multimodal semantic segmentation with vector-based comparison provides an interpretable framework for supporting topographic map updates. Since the workflow operates directly in the vector domain, it aligns with how topographic maps are maintained in practice. The framework highlights localized changes between airborne data and the reference map, allowing operators to focus inspection on a smaller set of candidate regions rather than the entire dataset. In this way, the proposed approach can assist mapping agencies by reducing the area requiring manual inspection and improving the efficiency of map maintenance.

7. Conclusion and Future Work

This study presents a multimodal framework for detecting object-level changes between airborne data and existing topographic maps using vector-based comparison. Semantic predictions generated from 2D orthoimages and 3D point clouds are converted

into polygon representations and directly compared with reference map vectors. The comparison with a 2D-only baseline indicates that integrating 3D geometric information can reduce fragmented detections and produce more spatially coherent candidate changes.

In addition to highlighting potential real-world developments, the proposed workflow also identifies inconsistencies and outdated objects in the reference map. These results demonstrate the potential of using existing maps both as supervision for semantic segmentation and as reference data for identifying changes. However, the reliability of detected changes is affected by segmentation errors, class ambiguity, and the mapping conventions used to create the reference maps. Furthermore, the lack of verified ground truth limits the scope of quantitative evaluation.

Future work will focus on improving the robustness and scalability of the proposed framework. Incorporating confidence-aware fusion or multimodal uncertainty estimation may help reduce error propagation during the fusion stage. In addition, future research could explore hybrid models that leverage existing maps alongside multimodal data to further improve change detection performance. Finally, integrating additional multi-temporal datasets and developing benchmark datasets with verified map updates would enable evaluation of automated map-updating workflows at larger spatial scales. Overall, the proposed approach demonstrates that multimodal semantic segmentation combined with vector-based comparison can serve as a practical tool for supporting topographic map maintenance.

8. Acknowledgments

This publication is part of the project titled *Learning from Old Maps to Create New Ones*. It is supported by the Open Technology Programme of the Dutch Research Council (NWO).

References

Agarwal, A., Kumar, S., Singh, D., 2019. Development of Neural Network Based Adaptive Change Detection Technique for Land Terrain Monitoring with Satellite and Drone Images. *Defence Science Journal*, 69(5), 474-480. DOI:10.14429/dsj.69.14954.

Albrecht, C. M., Zhang, R., Cui, X., Freitag, M., Hamann, H. F., Klein, L. J., Finkler, U., Marianno, F., Schmude, J., Bobroff, N., Zhang, W., Siebenschuh, C., Lu, S., 2020. Change Detection from Remote Sensing to Guide OpenStreetMap Labeling. *ISPRS International Journal of Geo-Information*, 9(7). DOI:10.3390/ijgi9070427.

Anjanappa, G., Elberink, S. O., Maiti, A., Lin, Y., Vosselman, G., 2026. Map2ImLas: Large-scale 2D-3D airborne dataset with map-based annotations. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 19, 100112. DOI:10.1016/j.ophoto.2025.100112.

Anjanappa, G., Oude Elberink, S., Vosselman, G., 2025. Learning From Detailed Maps: Joint 2D-3D Semantic Segmentation for Airborne Data with Selective Label Fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-G-2025, 101–108. DOI:10.5194/isprs-annals-X-G-2025-101-2025.

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

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, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., Li, H., 2021. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 1194-1206. DOI:10.1109/JSTARS.2020.3037893.

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.

Contributors, M., 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation.

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.

Daudt, R. C., Saux, B. L., Boulch, A., Gousseau, Y., 2018. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. arXiv:1810.08468.

de Gélis, I., Lefèvre, S., Corpetti, T., 2023. Siamese KPConv: 3D multiple change detection from raw point clouds using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 274-291. DOI:10.1016/j.isprsjprs.2023.02.001.

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.

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.

Esri, I., 2022. Esri Nederland. https://www.esri.nl/ [Accessed: 2026-03-04].

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

Girard, N., Charpiat, G., Tarabalka, Y., 2019. Aligning and updating cadaster maps with aerial images by multi-task, multi-resolution deep learning. C. Jawahar, H. Li, G. Mori, K. Schindler (eds), *Computer Vision – ACCV 2018*, Springer International Publishing, Cham, 675–690. DOI:10.1007/978-3-030-20873-8_43.

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.

Ji, S., Shen, Y., Lu, M., Zhang, Y., 2019. Building Instance Change Detection from Large-Scale Aerial Images using Convolutional Neural Networks and Simulated Samples. *Remote Sensing*, 11, 1343. DOI:10.3390/rs11111343.

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

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.

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.

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

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.

Niroshan, L., Carswell, J. D., 2022. Osm-gan: Using generative adversarial networks for detecting change in high-resolution spatial images. S. Bourennane, P. Kubicek (eds), *Geoinformatics and Data Analysis*, Springer International Publishing, Cham, 95–105.

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.

Rahman, F., Vasu, B., Van Cor, J., Kerekes, J., Savakis, A., 2018. Siamese network with multi-level features for patch-based change detection in satellite imagery. *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 958–962.

Ramachandram, D., Taylor, G. W., 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34(6), 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.

Rudner, T. G. J., Rußwurm, M., Fil, J., Pelich, R., Bischke, B., Kopačková, V., Biliński, P., 2019. Multi3Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 702-709. DOI:10.1609/aaai.v33i01.3301702.

Shafique, A., Cao, G., Khan, Z., Asad, M., Aslam, M., 2022. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sensing*, 14, 871. DOI:10.3390/rs14040871.

Singh, A., 1989. Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10, 989-1003. DOI:10.1080/01431168908903939.

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.

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.

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.

Yamazaki, K., Hanyu, T., Tran, M., de Luis, A., McCann, R., Liao, H., Rainwater, C., Adkins, M., Cothren, J., Le, N., 2023. AerialFormer: Multi-resolution Transformer for Aerial Image Segmentation. https://arxiv.org/abs/2306.06842, arXiv:2306.06842.

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.

Zhan, Y., Fu, K., Yan, M., Sun, X., Wang, H., Qiu, X., 2017. Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1845-1849. DOI:10.1109/LGRS.2017.2738149.

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, Q., Guo, X., Li, Z., Li, D., 2022. A review of multi-class change detection for satellite remote sensing imagery. *Geo-spatial Information Science*, 1-15. DOI:10.1080/10095020.2022.2128902.

Zorzi, S., Bittner, K., Fraundorfer, F., 2020. Map-repair: Deep cadastre maps alignment and temporal inconsistencies fix in satellite images. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 1829–1832. DOI:10.1109/IGARSS39084.2020.9323370.