Highlights

A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection

Mario Fuentes Reyes,Yuxing Xie,Xiangtian Yuan,Pablo d'Angelo,Franz Kurz,Daniele Cerra,Jiaojiao Tian

- We introduce a high quality synthetic multimodal dataset (SMARS) for urban object classification and 3D change detection, which to the best of our knowledge is the very first data set for the assessment of deep learning based 3D building change detection approaches.
- A novel workflow for synthetic 2D/3D multimodal multi-temporal datasets preparation in four steps: 3D virtual city design, airborne stereo imagery simulation, DSM generation, and orthophoto/reference data preparation.
- A systematic evaluation of the feasibility of the SMARS dataset for building extraction, multi-class semantic classification, and change detection. Besides single domain tests, we evaluate the performance of the datasets on cross-domain tests. To the best of our knowledge, this represents the first attempt at performing building extraction on synthetic and real cross-domain multi-model data.

A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection

Mario Fuentes Reyes[a,**], Yuxing Xie[b,a,**], Xiangtian Yuan[a,**], Pablo d'Angelo[a], Franz Kurz[a], Daniele Cerra[a] and Jiaojiao Tian[a,*]

[a]Remote Sensing Technology Institute, German Aerospace Center (DLR), Muenchener Strasse 20, Wessling, 82234, Germany
[b]Chair of Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany

ARTICLE INFO

Keywords:
3D change detection
building extraction
urban semantic segmentation
synthetic datasets

ABSTRACT

Advances in remote sensing image processing techniques have further increased the demand for annotated datasets. However, preparing annotated multi-temporal 2D/3D multimodal data is especially challenging, both for the increased costs of the annotation step and the lack of multimodal acquisitions available on the same area. We introduce the Simulated Multimodal Aerial Remote Sensing (SMARS) dataset, a synthetic dataset aimed at the tasks of urban semantic segmentation, change detection, and building extraction, along with a description of the pipeline to generate them and the parameters required to set our rendering. Samples in the form of orthorectified photos, digital surface models and ground truth for all the tasks are provided. Unlike existing datasets, orthorectified images and digital surface models are derived from synthetic images using photogrammetry, yielding more realistic simulations of the data. The increased size of SMARS, compared to available datasets of this kind, facilitates both traditional and deep learning algorithms. Reported experiments from state-of-the-art algorithms on SMARS scenes yield satisfactory results, in line with our expectations. Both benefits of the SMARS datasets and constraints imposed by its use are discussed. Specifically, building detection on the SMARS-real Potsdam cross-domain test demonstrates the quality and the advantages of proposed synthetic data generation workflow. SMARS will be published as an ISPRS benchmark dataset by ISPRS Commission I (working groups 1, 3 and 8).

1. Introduction

Recent years have seen dramatic progress in the development of image processing algorithms. Deep neural networks have outperformed traditional image processing approaches on most of the classical image understanding and interpretation problems (Minaee et al., 2021; Xie et al., 2020) .

At the early stages of computer vision, high quality manually labeled data series were published as benchmark datasets for computer vision tasks including classification and recognition, such as PASCAL Visual Object Classes (VOC) 150 (Everingham et al., 2010), KITTI (Geiger et al., 2013), Microsoft Common Objects in Context (MS COCO) (Lin et al., 2014), and Cityscapes (Cordts et al., 2016). These large-scale benchmark datasets have been then used to develop and validate deep learning algorithms. The performance of these networks highly depends on the amount and the quality of the available training data, which are expensive and sometimes difficult to acquire. The vast majority of newly published papers are dealing with the "Training" phase, as the collection of training data represents often the bottleneck for these applications (Zhou, 2018; Pourpanah et al., 2022). The performance of artificial intelligence (AI) algorithms is severely limited whenever insufficient data with low number of samples, unbalanced classes, or inaccurate annotations are available (Li et al., 2020; Xie et al., 2023).

Today, advanced neural network architectures are adopted in many other fields such as medical image analysis and remote sensing. Many excellent approaches originally proposed in the computer vision community have been

*Corresponding author
**These (the first three) authors contributed equally to this work.
Mario.FuentesReyes@dlr.de (M. Fuentes Reyes); Yuxing.Xie@dlr.de (Y. Xie); Xiangtian.Yuan@dlr.de (X. Yuan); Pablo.Angelo@dlr.de (P. d'Angelo); franz.kurz@dlr.de (F. Kurz); Daniele.Cerra@dlr.de (D. Cerra); Jiaojiao.Tian@dlr.de (J. Tian)
ORCID(s): 0000-0002-6593-5152 (M. Fuentes Reyes); 0000-0002-6408-5109 (Y. Xie); 0000-0001-7648-5938 (X. Yuan); 0000-0002-8407-5098 (J. Tian)

successfully applied and further developed for earth observation tasks, including building/road extraction, semantic classification, and change detection (Zhu et al., 2017; Xie et al., 2020). Additional hindrances are added to the ones listed above regarding the availability of training datasets for specific problems in remote sensing.

The nature of remote sensing data is often multimodal, as the different sensors usually provide complementary information on a target on the ground, by measuring the backscattered radiation in different frequency ranges (including visible, infrared, thermal emissions, and microwaves), and estimating ground and canopy height parameters yielding 3D information. Additionally, this information is seldom acquired in a single acquisition from the same observation platform, therefore introducing variations in viewing angle, sensing geometry, acquisition time, atmospheric conditions, and position of the source of radiation. The availability of different sources of information on the same area is often beneficial for remote sensing applications: for example, the fusion of 2D/3D data is advantageous for image classification (Ghamisi et al., 2016), building extraction (Hosseinpour et al., 2022), and change detection (Tian et al., 2013; Qin et al., 2016). In addition to the limited availability of the corresponding 2D/3D training data, sufficient variability in the data must be present in order to train a valid deep learning (DL) neural network. Furthermore, annotating changes in a large scale remote sensing images is time-consuming and error-prone. To our knowledge, there is no 2D/3D multimodal building change detection benchmark dataset available until now, which in part limits the implementation of effective AI techniques for 3D change detection.

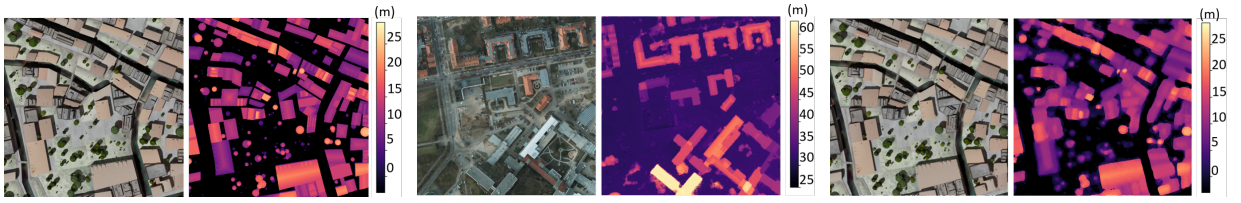
To this end, synthetic data have been proposed in order to fill this gap in a less expensive way. Currently, several available studies highlight the advantages of using synthetic data for solving real-world problems, especially in the fields of medicine and healthcare, for which real physical experiments are often linked to expensive retrieval costs (Chen et al., 2021). Besides avoiding data acquisition problems, using synthetic data has an increased flexibility when coping with data balancing, in particular for the studying of rare diseases (Chen et al., 2021). Several other studies experience similar problems, such as the ones conducted in the field of physics research, where the process of observing real data may be particularly long and expensive (Stoecklein et al., 2017; Li et al., 2021). Existing literature in remote sensing use synthetic data in order to evaluate their algorithms or fuse them with real data, yielding an efficient training for augmentation tasks. However, in addition to evaluating AI models, the synthetic data should be suitable for integration with real data in order to solve application oriented problems (Nikolenko, 2021). Thus, the domain gap between synthetic and real data should be limited.

Directly rendering of digital surface models (DSM) from a 3D environment retrieves highly accurate products as presented in Fig. 1 (a), exhibiting sharp boundaries around the buildings without any occlusions or gaps. Such precise DSM can be hardly achieved using real data with the currently available optical acquisition and stereo matching techniques, as results obtained from photogrammetry pipeline are characterized by blurred boundaries and contain outliers (Fig. 1 (b)). In order to reduce the gaps between rendered and real data, we aim at defining a novel approach generating synthetic DSMs with the same limitations of real ones, as for the DSM reported in Fig. 1 (c), which more closely resembles the level of detail in Fig. 1 (b) with respect to the generation using directly rendered samples.

Considering all the points above, we propose a novel synthetic photogrammetric data generation procedure with special focus on the application of 2D/3D multimodal classification (or segmentation), building detection and 3D change detection. We use this dataset as source and real remote sensing imagery as target for domain adaptation experiments. The main contributions of our paper are the following:

- A workflow to produce synthetic data with higher level of realism.
- A 2D/3D multimodal remote sensing dataset, which we name the Simulated Multimodal Aerial Remote Sensing (SMARS).
- A systematic evaluation of the performance of SMARS on building extraction, multi-class semantic classification and change detection.

This paper is organised as follows. Section 2 presents an overview of the state of the art of synthetic data used in remote sensing and the related studies in virtual city synthetic data generation. In section 3, the proposed synthetic data generation, which include the multi-temporal stereo imagery simulation as well as the data process procedure is introduced in detail. Section 4 illustrates the proposed method in details. In section 4, we further describe the details of the generated SMARS dataset and the tasks to be addressed, including building extraction (section 5), multi-class semantic segmentation (section 6) and building change detection (section 7). Moreover, extensive discussions are presented in Section 8. Section 9 provides the conclusions.



(a) Directly rendered samples (b) Real dataset samples-Potsdam (c) Synthetic samples with proposed approach

Figure 1: Quality differences between synthetic and real data. Elevation scale for the DSM is in meters.

2. State of the art

2.1. Existing real 2D/3D multimodal benchmark datasets

Due to the aforementioned reasons, the number of available 2D/3D multimodal benchmark datasets is limited. The ISPRS Potsdam dataset[1] is at the moment of writing the most popular public benchmark for 2D/3D semantic labeling, and it is also widely used to test and validate building extraction methods (Xie et al., 2023). This dataset provides airborne orthoimages and corresponding DSMs generated via dense image matching. The ground sampling distance of both images and DSMs is 5 cm. The original training data have 24 pairs of tiles, each having a size of 6000×6000 pixels (300 m×300 m). The ISPRS Vaihingen[2] is another airborne benchmark dataset containing both 2D images and DSMs. However, its limitation of having only near-infrared, red, and green bands restricts its applicability in mainstream applications requiring RGB images, as the blue band is not available. DroneDeploy[3] is a 2D/3D multimodal dataset containing aerial scenes captured from drones. Its main limitation is that it provides only original irregular mosaics, furthermore, it lacks a clear separation between training, validation, and test sets. Hence, it is not widely used in the community.

On the subject of change detection, there are a number of several single modal benchmark datasets available (Caye Daudt et al., 2018; Gupta et al., 2019; Caye Daudt et al., 2019; Shao et al., 2021). To the best of our knowledge, 3DCD is currently the only benchmark that provides 2D/3D multimodal data suitable for evaluating deep learning algorithms in remote sensing change detection (Coletta et al., 2022; Marsocci et al., 2023). Nonetheless, in this dataset, DSMs are obtained by LiDAR sensors, whose acquisition dates as well as the Ground Sample Distance (GSD) differ from the corresponding optical images. This may potentially affect their paired use in multimodal algorithms. Apart from the voids in the DSM, the changes are not exclusively defined for buildings but also for other land cover changes. In addition, the dataset only covers the urban centre of the city of Valladolid in Spain, and therefore is not suitable for domain adaptation experiments.

2.2. Synthetic data in remote sensing

Curating real 2D/3D multimodal datasets requires valid data acquisition and processing, which is then compounded by the time consuming and costly step of manual annotation. Therefore, the generation of synthetic data for remote sensing applications is preferred whenever real-world data are not available or difficult to collect. Authors in Börner et al. (2001) propose SENSOR (Software Environment for the Simulation of Optical Remote Sensing Systems) to simulate hyperspectral images. Artificial orbit and attitude data are used in Schwind et al. (2012) to analyse the co-registration errors between visible and near-infrared (VNIR) and short-wavelength infrared (SWIR) imagery for the design of the EnMAP (Environmental Mapping and Analysis Program) satellite. Simulated SAR images are generated in Tao et al. (2013) for change detection. Synthetic data have been explored in vegetation studies. Li and Strahler (1985) proposed a geometric-optical forest canopy model to explain the variance of a pixel in low resolution images of forest stands. The model represents conifers with Lambertian surfaces shaped as cones, which cast shadows on the ground.

Moreover, multi-temporal datasets are more costly to prepare and the annotation is more challenging with respect to single images: the number of public change detection benchmarks is therefore rather limited, furthermore, most of them are single modal data and some are characterized either by small size or a low ground sampling distance (Shi et al., 2020). The described difficulties in curating the described multi-temporal datasets can be mitigated by relying

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx
[2] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx
[3] https://github.com/dronedeploy/dd-ml-segmentation-benchmark

on synthetic data. For instance, Townshend et al. (1992) simulate a set of different mis-registrations degrees to find out their impact on vegetation change detection. Simulated change detection datasets have been used in Almutairi and Warner (2010) to compare state-of-the-art algorithms. The simulated data therein are rather simple with few shape patterns and additional artificial noise. A real LiDAR point cloud is used in de Gélis et al. (2021) to generate one Level of Detail 2 (LoD2) model as a pre-event dataset. By manually adding or removing buildings in the model, the construction or demolition of buildings can thus be simulated. This results in a time-consuming process, and with buildings as the only objects present in the 3D model, the results have a large domain gap with real urban 3D models.

In order to close the domain gap between simulated and real data, Hoeser and Kuenzer (2022) propose an artificial data generation procedure by including expert knowledge in a highly structured manner to control the automatic image and label generation, by employing an ontology in the process. However, with more complex background information, urban change detection is more difficult to simulate and control.

Radiative transfer models have been explored to simulate remote sensing data. Recently, in order to analyse vegetation behaviours, several synthetic data generation tools have been introduced based on the radiative transfer model (Qi et al., 2019; Disney et al., 2006). As one of the most representative software for radiative transfer modeling, the Discrete Anisotropic Radiative Transfer (DART) can accurately simulate 3D radiative budget and chlorophyll fluorescence of vegetation (Gastellu-Etchegorry et al., 2015), as well as passive remote sensing and LiDAR signals of natural and urban scenarios. It is capable of precisely simulating the vegetation reflectance in several wavelengths and also works for dense forests with complex canopy structures (Janoutová et al., 2019). However, rendering more realistic urban scenes using DART is quite challenging for inexperienced users due to its complex parameters requiring expert knowledge in the relevant fields. In contrast, 3D rendering engines such as Blender or Unity are considerably easier to use, offer more sophisticated rendering features, and support several formats of 3D models and materials (Richter et al., 2016; Shah et al., 2018; Fabbri et al., 2021). Moreover, in order to construct a large urban scene, many detailed and realistic 3D models for vegetation and buildings are needed. 3D rendering engines not only have more large-scale 3D city models but can also edit those models while DART does not support editing, which poses difficulty in simulating urban changes. In comparison, the 3D rendering engine is more advantageous in creating multi-temporal urban scenes of large regions.

2.3. Virtual city synthetic data

Generating data from a virtual model is currently becoming more popular in computer vision due to the capabilities of modeling software. However, the application of synthetic data is rather limited if the domain gap with the real data is too large. A virtual model can contain anything from a small object to a city. For example, building models can be used to create indoor based point clouds (Ma et al., 2020) or depth and semantics, as in Hypersim (Roberts et al., 2021). Studies related to autonomous driving have also benefited from the developments of synthetic data creation. A widely known example is the SYNTHIA dataset (Ros et al., 2016) that provides synthetic images of urban scenes labeled for semantic segmentation. Such scenes are rendered from a virtual New York City 3D model with the Unity game engine. The dataset includes segmentation annotations for 13 classes including pedestrians, cyclists, buildings, and roads. Another approach is used by CARLA (Dosovitskiy et al., 2017), an open source simulator that supports the training, prototyping, and validating of autonomous driving models. CARLA facilitates the data acquisition from street view for the generation of segmentation and depth maps. Similarly, the ParallelEye dataset (Li et al., 2019) generates images from the CityEngine software with depth and optical flow as part of the ground truth.

A similar setting can be considered for the simulation of aerial or satellite imagery. The Synthinel-1 dataset (Kong et al., 2020), also based on CityEngine, targets the building/no-building classification from an airplane perspective. The article also addresses the advantages of synthetic imagery by ablation studies. The VALID dataset (Chen et al., 2020), on the other hand, focuses on panoptic segmentation and depth estimation over urban infrastructure. Furthermore, the SyntCities dataset (Fuentes Reyes et al., 2022) provides semantics and disparity maps, making the data suitable for stereo reconstruction. The STPLS3D dataset (Chen et al., 2022) provides point clouds, and semantic and instance maps based on open geospatial data sources. Authors in Xiao et al. (2022) simulate LiDAR acquisition for an urban environment and deliver the dataset as point clouds.

However, further applications of synthetic data are limited by the large differences in characteristics between real and synthetic data. A remarkable example is the much higher quality of the DSM obtained from the virtual 3D models in comparison with the one generated from photogrammetric matching. Edges are usually sharper in the simulated data, and the occlusions are absent in the generated ground truth. In addition, images from real scenarios show imperfect textures, light reflection, seasonal changes, the presence of temporary objects (cars, pedestrians, street advertisements,

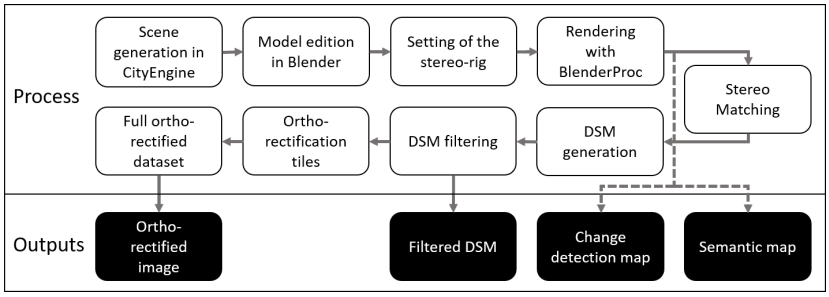


Figure 2: Basic description of the pipeline used to generate the dataset.

etc.), atmospheric effects, and other elements that cannot be easily modeled in software. Hence, the simulation is mostly restricted to the geometry of the scene, textures, and camera properties. Still, the rendered images can visually resemble real cases and help to compensate for the limits of real sensors (such as sparsity) and reduce the costs to generate ground truth.

3. Methodology on synthetic data generation

To close the gap between synthetic data collection and remote sensing applications we combine two techniques, airborne data collection from virtual city and photogrammetric stereo data preparation. In this section, we propose a novel workflow to generate a 2D-3D multimodal dataset. A diagram to summarize it is shown in Fig. 2. It consists of three parts: 3D virtual city design, imagery simulation, and data processing.

3.1. 3D virtual city design

In order to produce a realistic change scenario we used a 3D virtual city as a starting point to simulate the scene growth process, instead of directly generating artificial images. We built the 3D scenes based on the CityEngine software[4], a suite facilitating the modelling of urban environments based on the computer-generated Architecture (CGA) shape grammar language. The software was also used to develop the above-mentioned ParallelEye and Synthinel-1 datasets (Li et al., 2019). CityEngine supports building a city model from land cover maps, such as Open Street Map, or a manually designed base map. However, designing a virtual world with carefully customized features would require relevant expert knowledge and would be time-consuming. Therefore, we selected two predefined city models from ESRI and further refined them accordingly.

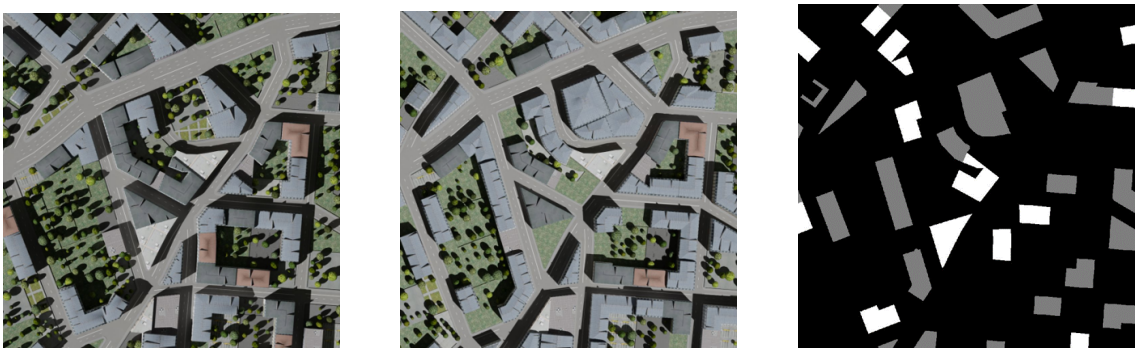
In this paper, we chose two typical European cities: Paris and Venice. Henceforth we refer to them as SParis and SVenice, respectively. The selected city models have a variety of textures and architectures resembling the original cities, as well as a large surface that allows the inclusion of a large number of buildings in the subsequent rendered images. The buildings are defined in terms of roof type, roof angle (if any), height, number of floors, floor height, and size of the parcel. In order to have a lifelike view, we further edited the 3D model of the cities by modifying the streets in order to have a more realistic topography, as the original version has streets with the shape of letters. The trees were replaced with textured ellipsoids instead of the original ones represented with a uniform color. Additionally, some areas were manually corrected in order to ensure that any parcel in the area included urban content.

A large pool of textures has been used in the provided models, namely 219 for buildings (rooftops and facades) and 87 for vegetation. For the latter ones, we edited the default textures of the ellipsoids by creating a dense representation of leaves in order to resemble canopies. While still limited with respect to the full variety of the real world, these refinements helped produce a scene with sufficient variability.

As the dataset is mainly intended for change detection applications in urban areas, each city model was generated with two versions simulating the city's growth:

- A case where approximately 50% of the parcels are covered by buildings. This is considered the model before changes happen and we refer to it as pre-model in the remainder of the paper.

[4]https://www.esri.com/en-us/arcgis/products/esri-cityengine/overview



(a) Pre-model (b) Post-model (c) Change ground truth

Figure 3: Samples from the pre- and post-models after rendering with associated ground truth for change detection. The pre-model has a lower building density and different illumination conditions. Black regions in the ground truth exhibit no change, while gray indicates new buildings and white replaced ones, respectively.

- A case with approximately 70% of the parcels covered by buildings. Some areas defined previously as gardens are replaced by constructions, while some buildings have been removed and substituted with green areas. This model contains the changes to be detected, and is therefore named post-model.

In Fig. 3, we show samples for both the pre- and post-model, respectively 3(a) and 3(b). The central image exhibits a higher number of buildings and less vegetation cover. Also, some of the original buildings have been replaced with lawns or vegetation.

According to the requirements described above, we adapted a total of four city models (two cities, two epochs) and exported all cases in the Wavefront (with extension .obj) format for further editing. The manipulation of the scenes in CityEngine demands about 17GB of RAM memory.

Subsequently, we loaded the Wavefront files in Blender, an open source tool for modeling, simulation, and rendering. We applied the BlenderProc pipeline (Denninger et al., 2020) to render the images. Our approach for the rendering is based on the one described in SyntCities (Fuentes Reyes et al., 2022) and we generated for this case the colored images (we refer henceforth to them as "optical") and the semantic maps.

Within Blender we split the geometry of the scenes according to their textures, separating all the surfaces into the required semantic labels. The available categories include: vegetation, streets, rooftops (mansard, gambrel, gable, hip and flat styles), facades, grass, landmarks, cars, and background. We combined them into five typical land cover classes used for urban mapping, including buildings (all rooftops, facades and landmarks), streets, high vegetation (trees), grass (lawns) and others (cars, water, bare soil or background).

We simulate different illumination conditions by setting an artificial Sun in two specific positions for the pre-/after-event models, reproducing two different times for data acquisition. The selected angles were 70° for elevation, and 217° (pre-model) and 160° (post-model) for azimuth. The same conditions were applied to both cities. Finally, we added a homogeneous plane under the ground level of each scene to avoid undefined regions (no value pixels) in the rendering process, which is assigned to the "other" category. Without it, distance would be considered to be infinite if there is an empty region in the objects. This plane guarantees a color and depth value for each rendered pixel.

3.2. Airborne stereo imagery simulation

SMARS is designed to resemble aerial imagery and the simulated camera is constrained by a stereo rig, which helps to later generate a digital surface model (DSM). In this part, we provide more details on the simulated data acquisition and camera parameters.

Firstly, the simulated camera is located 2km above the origin of the scenes. Since we used synthetic models that are not georeferenced, the origin of the coordinate system assigned by City Engine is used by default. An arbitrary point located at the center of the model and on the terrain level is taken as a reference for the rendering process.

In Fig. 4(a), we show the configuration of the stereo rig. In order to simulate the stereo imagery acquisition procedure, two cameras are located at the same distance from the rig center with a baseline of 200m in all cases. Both cameras follow the pinhole model and have the same focal length. As image resolution plays an essential role in

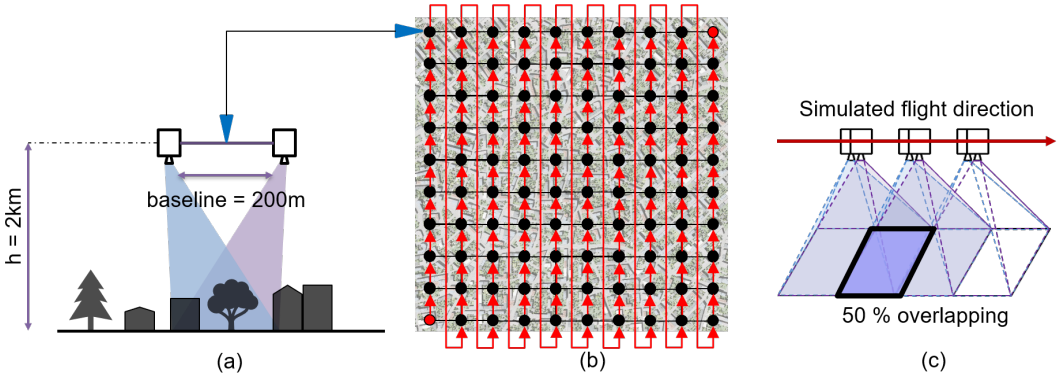


Figure 4: Simulated stereo configuration. (a) Stereo rig, where the converge distance and baseline of the cameras have been adapted to cover the same area on the ground. (b) The trajectory of the simulated camera above the scene. (c) Overlapping between adjacent samples is 50% for both horizontal and vertical directions.

transfer learning, we aim to provide this image dataset in two GSDs, namely 30cm and 50cm. Following Eq. 1, we set the focal length of the cameras to 234.37mm and 140.62mm, respectively.

$$f = \frac{height \cdot sensor\_width}{covered\_area} \tag{1}$$

where f is the focal length, $height = 2000$m as described above, $sensor_width = 36$mm for the simulated camera and $covered\_area = 1024 * \text{GSD}$, being 1024 the size in pixels of the output image. The converge distance is set to 2km (same as the simulated height) with an off-axis camera, which allows us to cover the same area on the ground from two different points of view. This configuration is also modeled with the offset of the principal point in the intrinsic matrix of the camera.

In Fig. 4(b) we illustrate the trajectory of the simulated camera above the scene. We rendered images at 100 positions within a regular square grid, with strides set as 153.6m and 256m for 30cm and 50cm GSD, respectively. The center of the grid is set to be close to one of the scenes, so most of the content is included. In order to simulate a real-world airborne data acquisition campaign, the pair of stereo-cameras are moved from the lower-left to the upper-right corner with a constant stride. The points belonging to the grid represent the location of the center of the stereo rig (see the arrow with blue extremes). This means that the cameras are located symmetrically to the left and right side of each point.

Overlapping between adjacent samples is set to 50% in both the horizontal and vertical directions of the grid. A visual representation of the overlapping is given in Fig. 4(c), where the camera pairs along the simulated flight direction are also included. The images are rendered with a size of 1024×1024 pixels.

After rendering, a semantic segmentation map to be used as ground truth (GT) is delivered with the categories described previously (buildings, streets, vegetation, lawns and others). For the building extraction GT map, we combine all categories except building to no-building, enabling binary semantic segmentation. With the pre-/post-event building extraction GT maps, we calculate the building change detection map by taking only the building class for comparison. Three change classes are included:

- No change: buildings or no-buildings have the same semantic label pre/post-event images.
- Construction: pixels labelled as building in the post event images are no-building in the pre-event images.
- Demolition: pixels labelled as building in the pre-event images are replaced by the no-building label.

The change detection ground truth is directly rendered from the 3D model with an orthographic view. Labels for the semantic categories are also directly rendered from Blender, as BlenderProc generates a category for each object in the scene. We assigned all geometric elements to the desired categories. The building masks are a simplified version of the category maps considering a binary building/non-building case. For the change detection mask, building masks are compared and labelled according to their difference. In this case, all generated ground truth is generated in the rendering step, and therefore perfectly matches the original images. Due to the orthorectification process described in subsection 3.4, the alignment will not be perfect as this simulates the quality obtained from a photogrammetric pipeline. Results show that the alignment differences do not have a significant impact on the three evaluated tasks.

3.3. Stereo matching and DSM generation

Although very precise 3D point clouds and DSMs can be directly delivered with the rendering software, the quality of these data for all cases will be higher than the real-world 3D point clouds generated by stereo matching techniques, where many mismatching errors and occlusions occur. Thus, in this work we only take the synthetic stereo image pairs and generate the orthophotos and 3D point clouds with a traditional approach. First, we assign a fake UTM projection to all synthetic airborne stereo images, in order to enable the photogrammetric processing. Concretely, we assign the tiles to the UTM zone 31N coordinate system (EPSG:32631), even though the simulated model does not match any region on a real map, this area would match the city of Paris. Additionally, for the photogrammetric pipeline we enter the camera extrinsic and intrinsic matrices, including focal length, principal points, and camera rotation and translation parameters. The extrinsic and intrinsic parameters of the synthetic data are precise and there was no artificial noise added. We assume that the deviation of the positional accuracy is negligible, as the relative accuracy of real-world aerial images used for stereo matching is better than 0.2 pixels.

A DSM is generated from all tiles by using the CATENA pipeline (Krauß, 2014), which is used for multiple tasks related to the processing of satellite imagery. The disparity estimation, which is the first step, is computed via Semi-Global Matching (SGM) (Hirschmuller, 2008), an algorithm widely used for stereo matching due to its good balance between accuracy and computational costs. SGM takes a rectified stereo image pair as input and estimates a disparity map. We apply the implementation of SGM described in (d'Angelo and Reinartz, 2011), which takes satellite data as input, and set the penalty parameters $P_1 = 400$, $P_2 = 800$ and the window size for the Census transform (Zabih and Woodfill, 1994) to 7×9.

After the matching and the use of the camera parameters to determine the 3D location of each pixel, we retrieve a georeferenced DSM for each stereo pair. We subsequently merge all the stereo pair DSMs by using the median of all values belonging to the same location, resulting in one final DSM for each virtual city.

As a real DSM generation procedure, gaps are present due to matching failures or occlusions. We apply an inverse distance weighted interpolation in order to fill the remaining holes (Bartier and Keller, 1996).

3.4. Orthophoto and reference data

The orthorectification process for the rendered optical tiles is implemented in a GPU as described in (Kurz et al., 2012), considering as input the generated DSM, and the intrinsic and extrinsic parameters of the optical images. The outputs are take into account occlusions by buildings and vegetation. Bilinear interpolation is used to resample the orthorectified images to a given ground sampling distance.

We merge all the tiles into a single large image with the warp utility from the GDAL library (GDAL/OGR contributors, 2022), having as a result a complete orthorectified optical image, corresponding to the DSMs at pixel level.

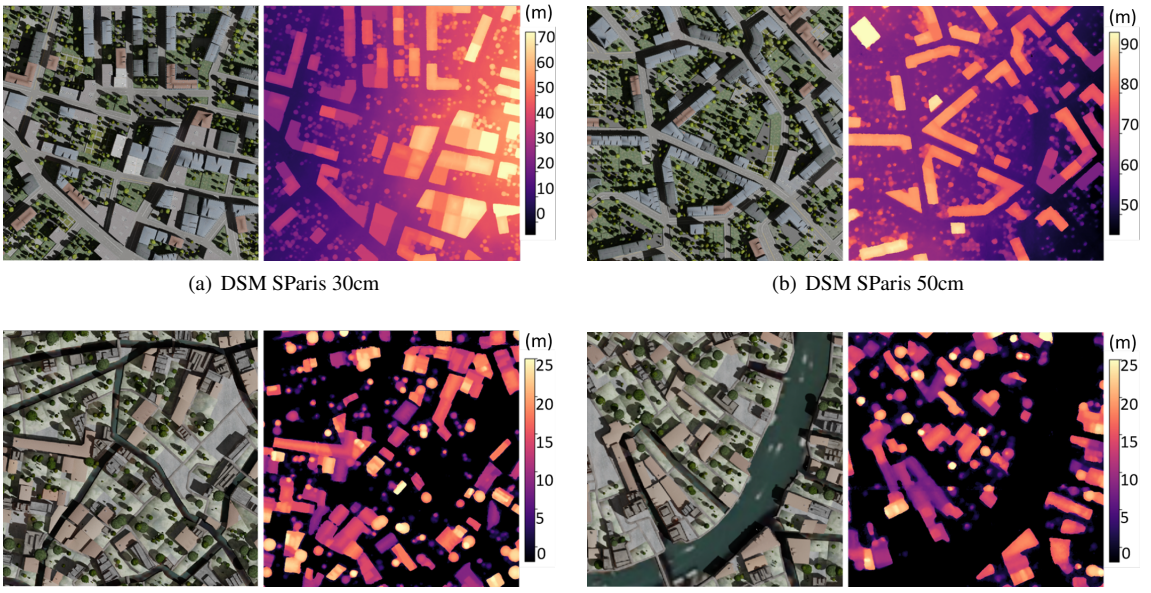
4. Experimental design

In this section we describe some additional details of the generated SMARS dataset and the delimitation of the regions used for training and testing in the deep learning algorithms for both cities. Additionally, we explain the tasks to be addressed with our generated data to show the advantages and constraints of SMARS.

4.1. SParis and SVenice multimodal data structure

The pre- and post-event DSMs and orthophotos are generated using the workflow described in Section 3. All the datasets are projected to the UTM zone 31N coordinate system and cropped in order to cover the same regions. Fig. 5 reports examples of the generated DSMs. Buildings appear well delimited and easy to identify in most cases, while other elements such as streets or vegetation appear incomplete or blurred. There is a clear difference between the models obtained using 30cm and 50cm GSD respectively, as the former exhibits sharper edges with individual trees easy to identify, while the latter exhibits some blobs merging different objects. Despite some artifacts or the presence of outliers, the DSMs still have a high quality in all cases due to their synthetic nature.

The final dataset splittings are summarized in the diagram below. We list all possible subsets but report the names for only three of them for each city in order to simplify the diagram, with the remaining cases following the same nomenclature. For each subset, we have available optical images, DSMs, semantic maps, and building masks for both pre- and post-event scenarios. Additionally, we have building change detection masks for the difference between pre- and post-images. All these cases are shown in Fig. 6.



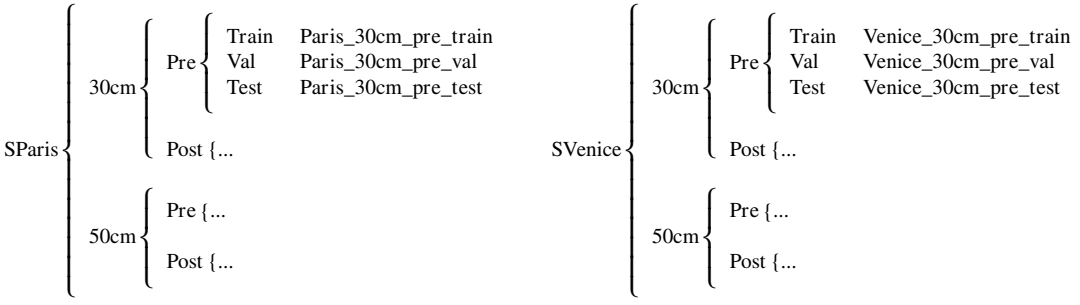
(a) DSM SParis 30cm

(b) DSM SParis 50cm

(c) DSM SVenice 30cm

(d) DSM SVenice 50cm

Figure 5: Example regions of the DSMs generated for SMARS besides the paired orthorectified images. All samples are taken from the pre-event models. Elevation scale for the DSM is in meters.



Figs. 7 and 8 illustrate the pre-event training, validation and test areas for SParis and SVenice, respectively. For post-event data, the splitting in training, validation and test data follows the same process. The size of both SParis and SVenice rasters with 30cm GSD is 5600 × 5600 pixels. For SParis (Fig. 7) 30% of the coverage is used for training (marked in yellow as P1), 30% for validation (P2), and 40% for testing (P3). Training, validation and testing data are marked in yellow as P1, P2, and P3, respectively. For SVenice (Fig. 8), 50% of the coverage is set as training as it contains a large area of water, belonging to the class "others" (V1, marked in blue), while 15% is used for validation (V2) and 35% (V3) for testing. The footprints of the images with a GSD of 50cm are larger with respect to the ones of 30cm, namely 4500 × 3560 pixels (SParis) and 5600 × 5600 pixels (SVenice). The splitting boundaries of the 50cm datasets are the same as the ones in the 30cm datasets. In Fig. 7 and 8, P4/V4, P5/V5 and P6/V6 represent respectively training, validation and testing areas for the 50cm datasets.

The released version of SMARS includes the above-mentioned rasters all in GeoTIFF format. Optical images are stored in three Band (RGB) uint8 format, DSMs with float precision and ground truth maps/masks with discrete values. The released version includes 9.0 GB of GeoTIFF data, covering the original rasters and split training, validation, and test tiles. According to our splitting approach, each city_GSD data group consists of 27 tiles (6 pre-/post-event optical image tiles, 6 pre-/post-event DSMs, 6 pre-/post-event building masks, 6 pre-/post-event semantic maps, and 3 pre-/post-event building change detection masks). With four city_GSD combinations, there are 108 tiles in total. To make it easier for users to start with this data, a Python tool for patch cropping is included in the release version. The default training patches in our work have a size of 512 × 512 pixels with 50% overlapping, but users can customize

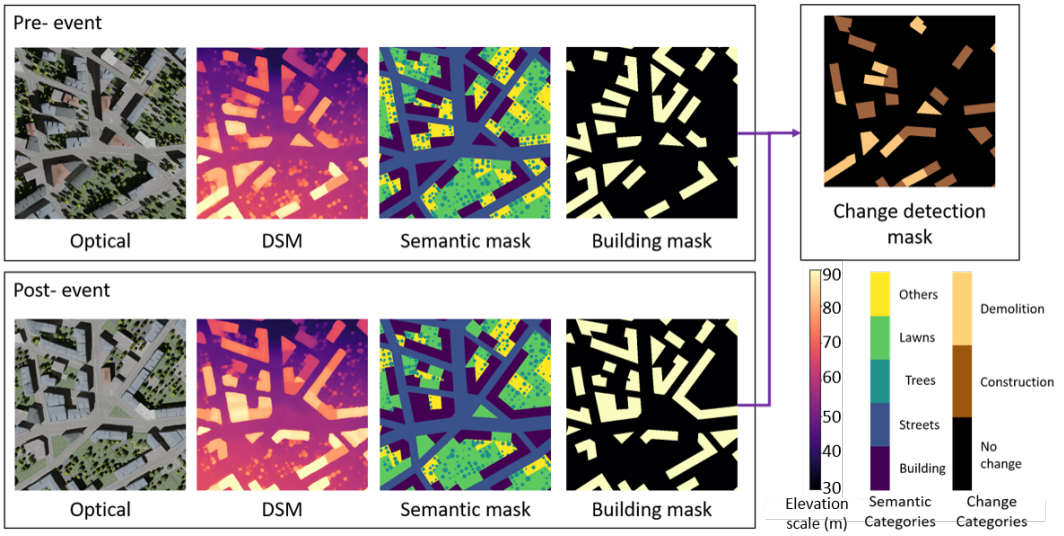


Figure 6: Available information for each tile in pre and post-events scenarios. For each case, an optical image, a DSM and semantic and building masks are included. For the change detection, the difference between the two events is used for the ground truth mask. Scales are given as a reference for displayed information. The elevation scale for the DSM is in meters.

training and validation patches as required. In addition, the DSM rasters can be converted to point cloud formats with another released Python tool, so users can use SMARS data with point cloud building extraction/semantic segmentation networks directly.

We employ the pre-event version with a GSD of 30 cm in the building extraction and 5-class semantic segmentation test design. In order to better visualize the testing results in this paper, the test region of each dataset is split into two regions, I and II (Fig. 9).

4.2. Data quality evaluation design

The proposed SMARS dataset focuses on building extraction, semantic segmentation, and 3D change detection. The building types and distributions of SParis and Venice are distinct and resemble those of the corresponding real cities. In addition, the building blocks of Venice are often separated by water channels instead of roads. The distinct features between the SParis and SVenice data result in large domain gaps for learning tasks, making SMARS a feasible data source for domain adaptation tests.

We experiment with state-of-the-art deep learning neural networks on the SMARS dataset for three tasks: 1) building extraction, 2) multi-class semantic segmentation, and 3) building change detection.

As buildings are dense in the scenes and resemble the architecture of real cities, the SMARS samples are an adequate input for building extraction and multi-class semantic segmentation tasks. Several effective deep learning approaches are available for these tasks. For the first two tasks, we work on two situations. The first is the single domain test with the provided train/Val/Test data from each synthetic city separately. In addition, we perform synthetic data cross-domain experiments by using SParis and SVenice separately for training, and test on the other model. Finally, we evaluate the predictions of samples from real sensors in the building segmentation task, which represents the most interesting experiment. In this case, we take samples from the Potsdam dataset for testing. We use as input either the images or the point clouds, which are addressed by 2D and 3D approaches respectively. Aspects to be studied include the correct detection and completeness of the buildings, as well as the transferability to previously unseen data.

Considering that the data are rendered with different semantic classes (*buildings*, *streets*, *trees*, *lawns* and *others*), we assess the performance of different neural networks using both 2D and 3D data, relying respectively on the images and their associated DSMs. Samples from both models have been generated with the same classes, enabling both single and cross-domain strategies to be tested. As the scenes are based on two different architectures, we expect some difficulties in the cross-domain case. Unfortunately, these experiments cannot be applied to real data due to the incompatibility of the available classes. Apart from the usual metrics such as Jaccard score (intersection over union (IoU)) and accuracy, we investigate the effect of the different nature of the data in relation to the obtained results.

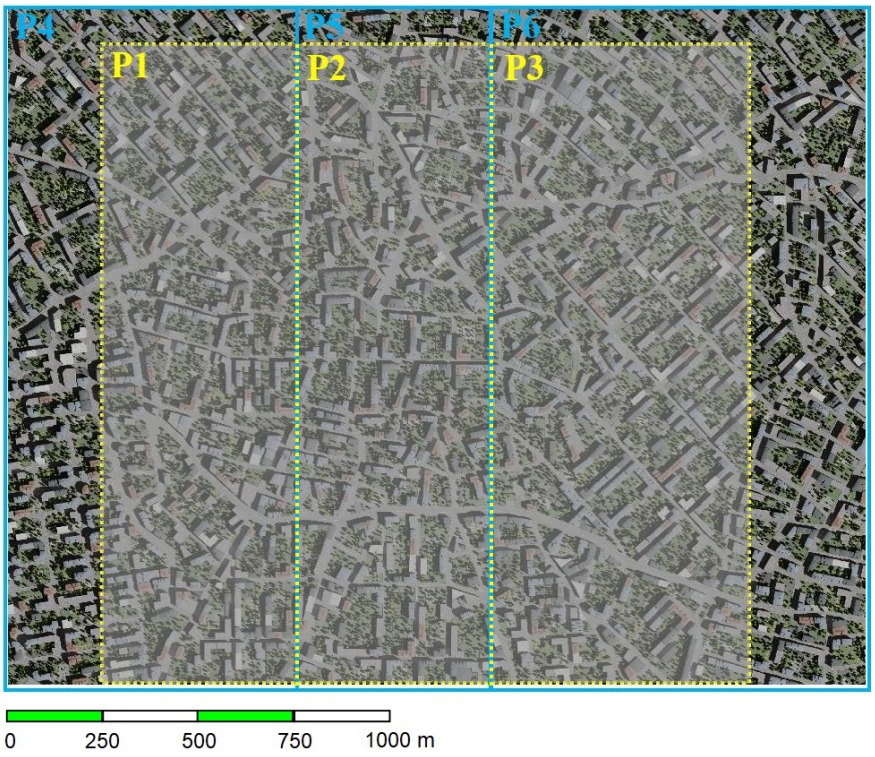


Figure 7: Layout of the SParis images. Yellow dotted lines represent the splitting of the 30cm resolution dataset (1.68 km by 1.68 km). Blue solid lines represent the splitting of the 50cm resolution images (2.25 km by 1.78 km).

The third task, change detection, is a key aspect to evaluate as the virtual scenes are constructed in order to simulate changes caused by city growth. The objective of this task is to localise the regions where the landscape has a significant change, whether because of new constructions or demolitions of buildings. The quality of the processed DSM plays a relevant role in the performance of this task, therefore we expect a difference in performance for the two cities, where the heights and space between buildings are significantly different. A comprehensive analysis based on the results highlights which approaches performed better on SMARS, and the approaches yielding a superior performance. As there are no unanimously accepted deep-learning based 3D change detection approaches available, we apply machine learning based approaches and are not able in this case to evaluate the transferability as in the previous tasks.

The following sections describe how the applied algorithms have been adapted for our experiments, the metrics to assess the performance on the different tasks, and a discussion of the capabilities and constraints of our dataset.

5. Building extraction

To examine the similarities between the SParis and SVenice datasets, and the domain gaps between the subsets of SMARS data and the real multimodal data in a deep learning context, we conduct building extraction experiments using different combinations of training and testing data, as detailed below:

- SMARS-to-SMARS single domain test
  - SParis→SParis: SParis used for training, SParis for testing
  - SVenice→SVenice: SVenice used for training, SVenice for testing
- SMARS-to-SMARS cross domain test
  - SParis→SVenice: SParis used for training, SVenice for testing
  - SVenice→SParis: SVenice used for training, SParis for testing
- SMARS-to-real cross domain test
  - SParis→Potsdam: SParis used for training, ISPRS Potsdam for testing

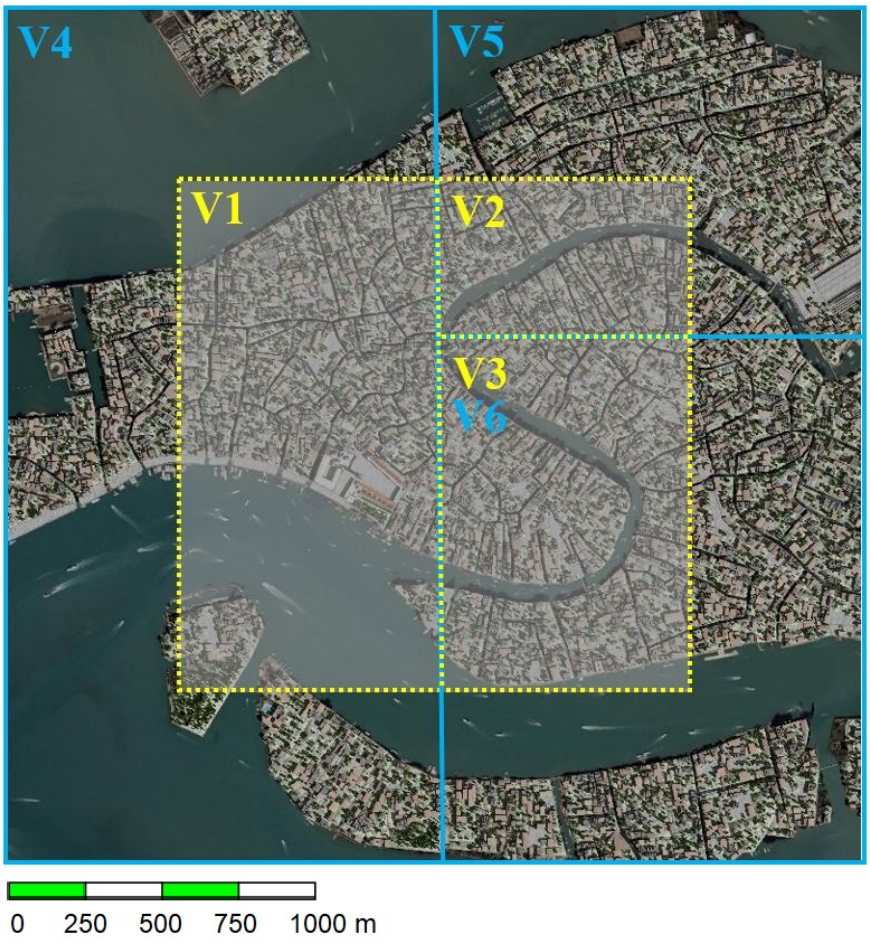


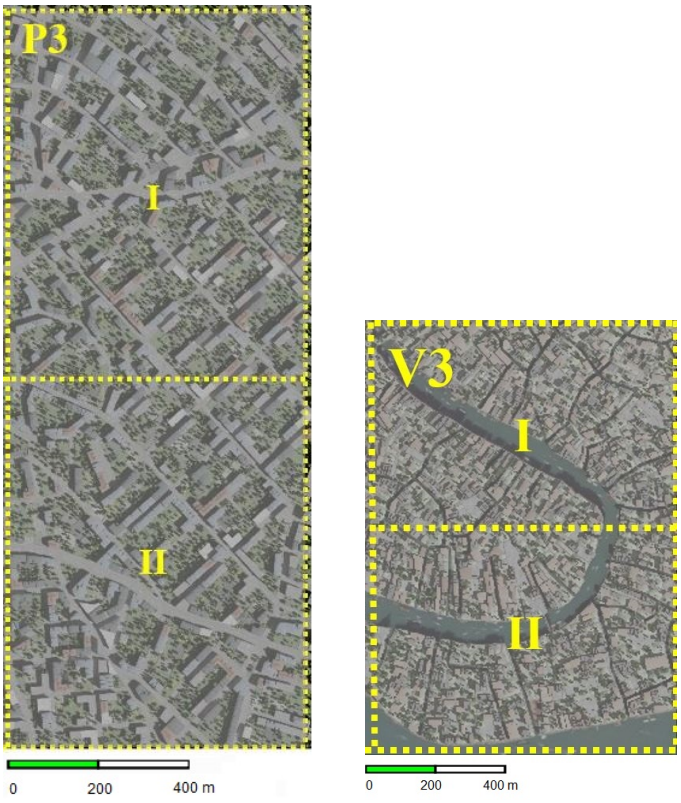
Figure 8: Layout of the SVenice images. Yellow dotted lines represent the splitting of the 30cm resolution dataset (1.68 km by 1.68 km). Blue solid lines represent the splitting of the 50cm resolution images (2.8 km by 2.8 km).

- Svenice→Potsdam: SVenice used for training, ISPRS Potsdam for testing
- Potsdam→Potsdam: ISPRS Potsdam used for training, ISPRS Potsdam for testing (reference)

In order to assess the building extraction task from optical images, we report results obtained by applying the state-of-the-art Swin Transformer(Liu et al., 2021). We also employ the widely-used point cloud network SparseConvNet (Graham et al., 2018) to investigate the domain gaps between DSMs. Point cloud networks are proven to have a reasonable performance in urban scenes (Xie et al., 2020), even in the semantic segmentation task of photogrammetric point clouds (Bachhofner et al., 2020) or DSMs (Xie et al., 2023). We downsample the resolution of both optical imagery and DSM-derived point clouds from the Potsdam dataset from 30 cm from 5 cm in order to reduce the impact of differences in spatial resolution on the results.

5.1. Single domain test: 2D data

In order to verify whether deep learning methods can be applied to the SMARS data for remote sensing tasks such as building extraction from earth observation data, we train the Swin Transformer with the optical images of SParis and SVenice separately, using the data split described in chapter 4.2, and test the models with the corresponding test sets. Results are listed in Table 1, rows 1 and 3. The segmentation results are reported in Figs. 10 and 11 (a) and (b). Within the same dataset, the building extraction IoUs of SParis and SVenice are above 95% and 92% respectively, indicating very satisfactory results. We can conclude that the synthetic data can be used for remote sensing tasks with deep learning approaches, yielding results similar to the ones obtained using real data.



(a) Test regions, Pre-SParis-30cm, 0.67 km by 1.68 km

(b) Test regions, Pre-SVenice-30cm, 0.84 km by 1.18 km

Figure 9: The test regions of the 30 cm datasets.

5.2. Cross domain test: SMARS-SMARS 2D data

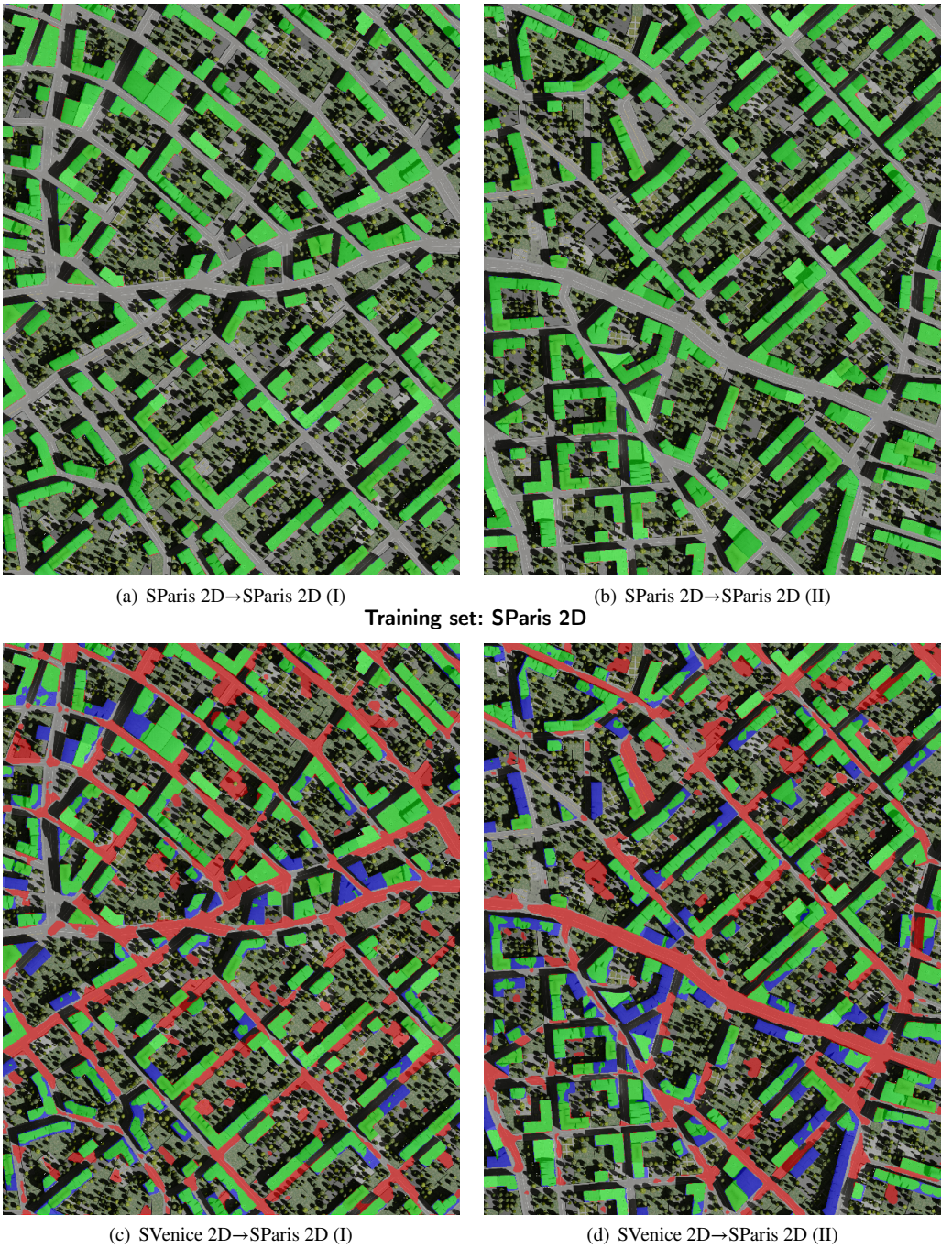
In order to investigate domain shifts between the two synthetic datasets, and how these affect in turn the downstream task of building extraction, we test the Swin Transformer trained with one of the two sets on the other one, according to the data split described in chapter 4.2. Results are presented in Table 1, rows 2 and 4. The segmentation results are reported in Figs. 10 and 11 (c) and (d). With respect to the results presented in Section 5.1, the building IoU scores are significantly degraded, from 95% and 92% to 57.37% and 44.59%, respectively. The decrease in performance can be attributed to large domain shifts, as evidenced by the distinct architectural styles, street appearance and ground features of the two scenes. The decrease in performance when training and testing data have distinct distributions can also be observed in real remote sensing data.

Table 1
SMARS optical imagery building extraction results

Train	Test	Precision [%]	Recall [%]	F1 Score [%]	IoU [%]
SParis	SParis	97.27	98.38	97.82	95.73
SParis	SVenice	65.62	81.89	72.86	57.30
SVenice	SVenice	95.92	95.84	95.88	92.09
SVenice	SParis	47.28	88.69	61.68	44.59

5.3. Cross domain test: SMARS-real 2D data

In order verify the suitability of employing a synthetic dataset to assess algorithms to be applied to real data, we test our network trained with SMARS with the ISPRS Potsdam data (for brevity named Potsdam thereafter). In



(a) SParis 2D→SParis 2D (I) (b) SParis 2D→SParis 2D (II)

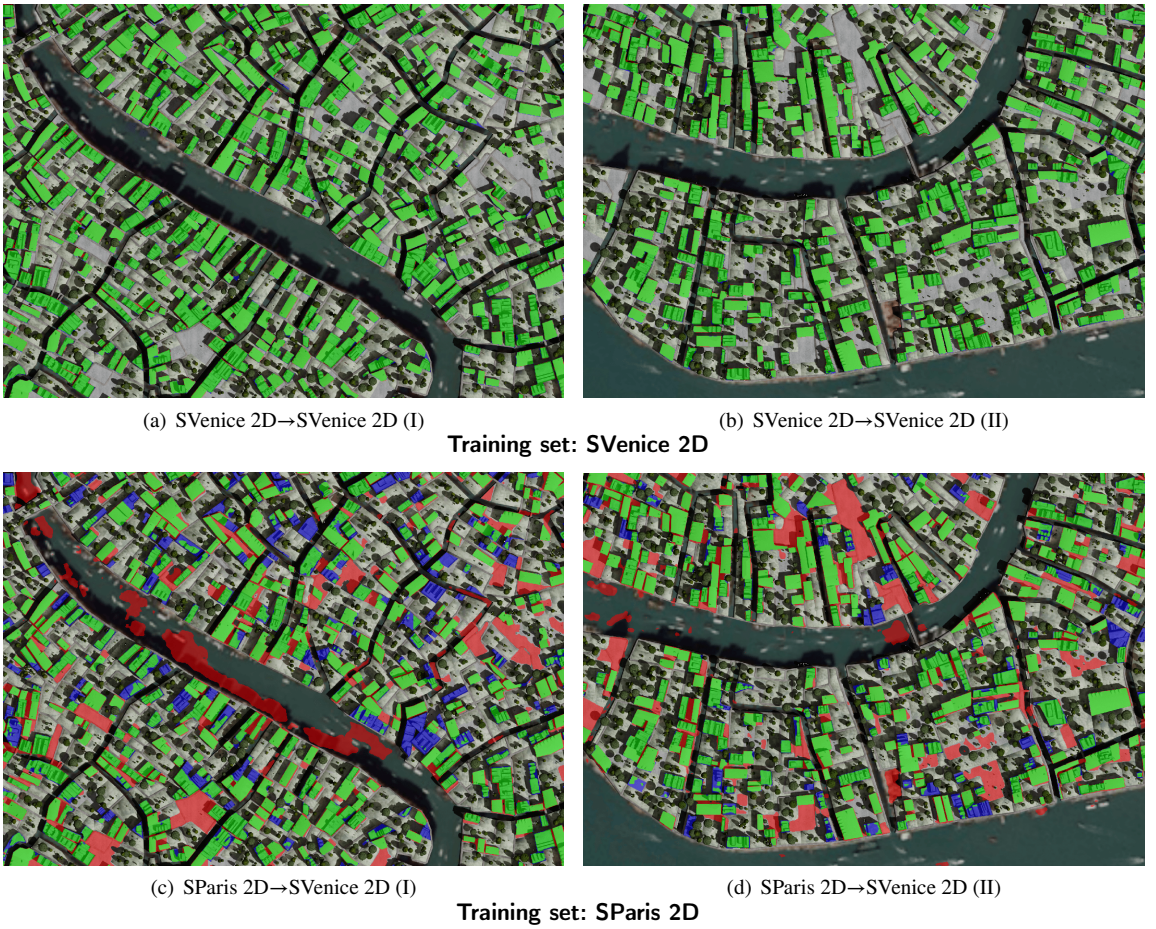
Training set: SParis 2D

(c) SVenice 2D→SParis 2D (I) (d) SVenice 2D→SParis 2D (II)

Training set: SVenice 2D

Figure 10: The image building extraction results of SVenice: (a) and (b) Swin Transformer trained on SParis; (c) and (d) Swin Transformer trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

addition, we apply the CIELAB color space transformation (He et al., 2021) to the SMARS 2D data in order to reduce the domain gaps between the synthetic and real datasets. Adopting similar workflow and settings as in our previous work (Li et al., 2022), we select 10 images from the Potsdam data to be used as reference, and transform the SParis and SVenice data to the Potsdam data in the LAB color space (LAB), and then convert SParis and SVenice back to



(a) SVenice 2D→SVenice 2D (I) (b) SVenice 2D→SVenice 2D (II)

Training set: SVenice 2D

(c) SParis 2D→SVenice 2D (I) (d) SParis 2D→SVenice 2D (II)

Training set: SParis 2D

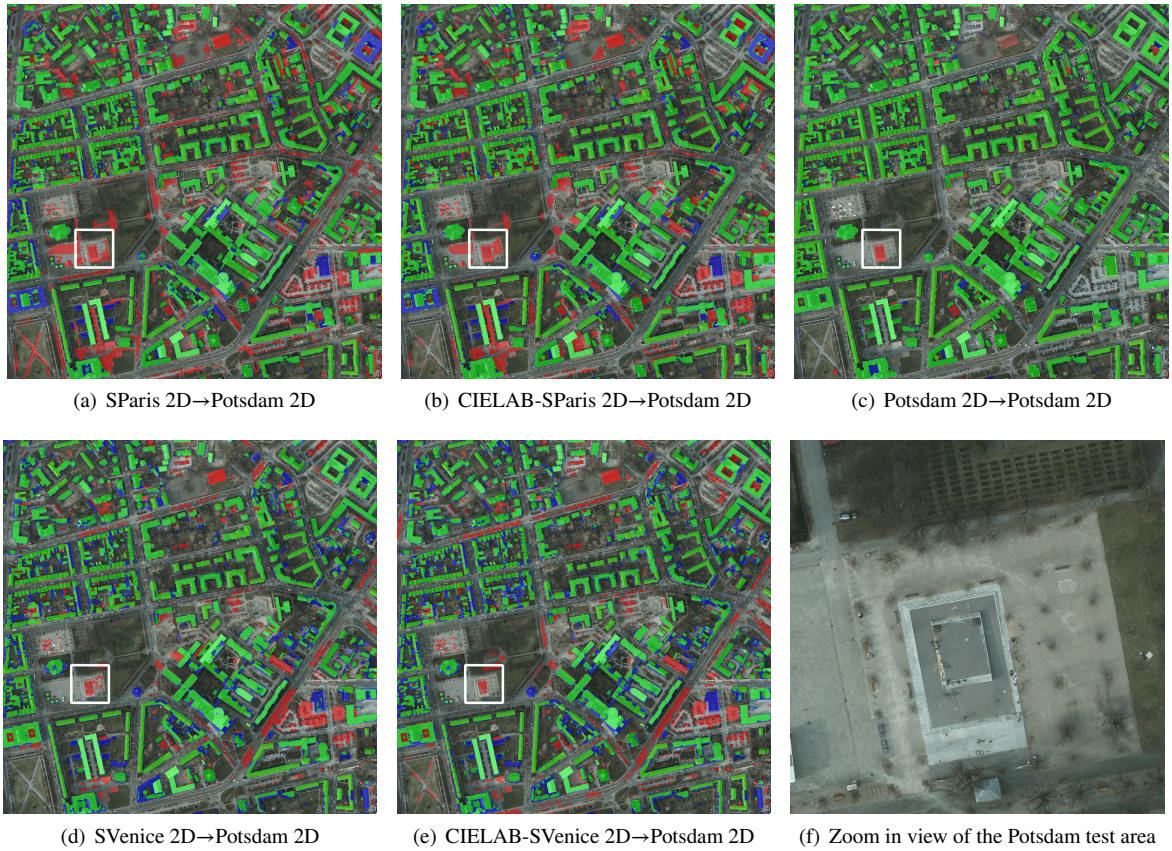
Figure 11: The image building extraction results of SVenice: (a) and (b) Swin Transformer trained on SVenice; (c) and (d) Swin Transformer trained on SParis. Legend: True Positive False Positive False Negative. True Negative is not displayed.

RGB colorspace. Quantitative results are listed in Table 2. The result of Potsdam→Potsdam is listed in the last row for reference. Surprisingly, the test results on Potsdam data yield better performance than the SParis/SVenice cross domain experiments. This can be explained by the fact that the buildings in Potsdam are more similar to SParis than SVenice in terms of their structure and appearance. The CIELAB transformation does not lead to consistent performance changes. For the SParis trained model, the IoU score of building extraction in Potsdam dataset increases 4%, while for SVenice trained model decreases over 3%. Another performance discrepancy is observed in the relationship between precision and recall. For model trained on SParis, precision is significantly lower than recall, while the opposite is observed for model trained on SVenice. Results of building extraction in Potsdam is shown in Fig. 12. In spite of being far from perfect for the Potsdam dataset, the majority of buildings is correctly extracted, suggesting that simulated optical data can be suited to train a neural network for building extraction and other tasks employing real earth observation data. To further validate the suitability of the SMARS as training data for building extraction, we include the cross-domain test results of Potsdam reported by Peng et al. (2022), which address the difficulties in cross-domain building extraction. The results are shown in Table 2. Two real datasets, namely WHU (Ji et al., 2019) and MASS (Mnih, 2013) are used as source domain data for training. Without any domain adaptation strategy (denoted w/o DA), the models trained with SMARS significantly outperform those trained with WHU and MASS datasets notwithstanding that they are real data. In the same work by Peng et al. (2022), a unsupervised domain adaptation method named FDANet was proposed, which consist Wallis filter, adversarial learning and consistency regularization to tackle domain shift. Nevertheless, our model trained with CIELAB-transformed-SParis data outperforms FDANet trained with MASS data, further demonstrating

the potential of SMARS as training data. In addition, the result of intra-domain experiment that used Potsdam as training data is listed in the last row of Table 2.

Table 2
2D cross-domain study, row 1-4: SMARS and ISPRS Potsdam as training and testing sets, respectively; row 5-8: WHU and MASS as training data, and Potsdam as testing data from Peng et al. (2022), where 'w/o DA' denotes without domain adaptation and FDANet is described in chapter 5.3; the last row: Potsdam as training and testing.

Train	**Test**	Precision [%]	Recall [%]	F1 Score [%]	IoU [%]
SParis	Potsdam	69.47	84.57	76.28	61.65
SParis (CIELAB)	Potsdam	73.18	86.70	79.37	65.79
SVenice	Potsdam	81.68	73.37	77.30	63.00
SVenice (CIELAB)	Potsdam	78.28	71.44	74.68	59.59
WHU (w/o DA)	Potsdam	-	-	68.83	52.47
WHU (FDANet)	Potsdam	-	-	88.87	79.96
MASS (w/o DA)	Potsdam	-	-	39.05	24.26
MASS (FDANet)	Potsdam	-	-	78.63	64.78
Potsdam	Potsdam	94.45	95.30	94.88	90.25



(a) SParis 2D→Potsdam 2D (b) CIELAB-SParis 2D→Potsdam 2D (c) Potsdam 2D→Potsdam 2D

(d) SVenice 2D→Potsdam 2D (e) CIELAB-SVenice 2D→Potsdam 2D (f) Zoom in view of the Potsdam test area

Figure 12: SMARS 2D→Potsdam results, trained respectively with SParis(a), SParis-CIELAB(b), SVenice(d), and SVenice-CIELAB(e).Legend: True Positive False Positive False Negative. True Negative is not displayed. A detailed view of the area misclassifed as buildings by all models is shown in (f). The area is highlighted in (a)-(e) with a white rectangle.

5.4. Single domain test: 3D data

As mentioned above, we also employ the point cloud network SparseConvNet (Graham et al., 2018) as a reference in order to examine the quality of the DSMs. The quantitative results of single-domain building extraction from DSM-based point clouds are listed in Table 3, rows 1 and 3. The classification results are presented in Fig. 13 (a) and (b), and Fig. 14 (a) and (b). The IoU scores of SParis→SParis and SVenice→SVenice are 95.16% and 91.03%, respectively, which are slightly inferior to the results obtained by the Swin Transformer with the simulated optical imagery but still satisfactory. Based on the evaluation metrics and visual quality, the synthetic data can be considered a valid substitute or integration for the training of deep networks for the considered tasks, whenever sufficient annotated real earth observation data are not available.

5.5. Cross domain test: SMARS-SMARS 3D data

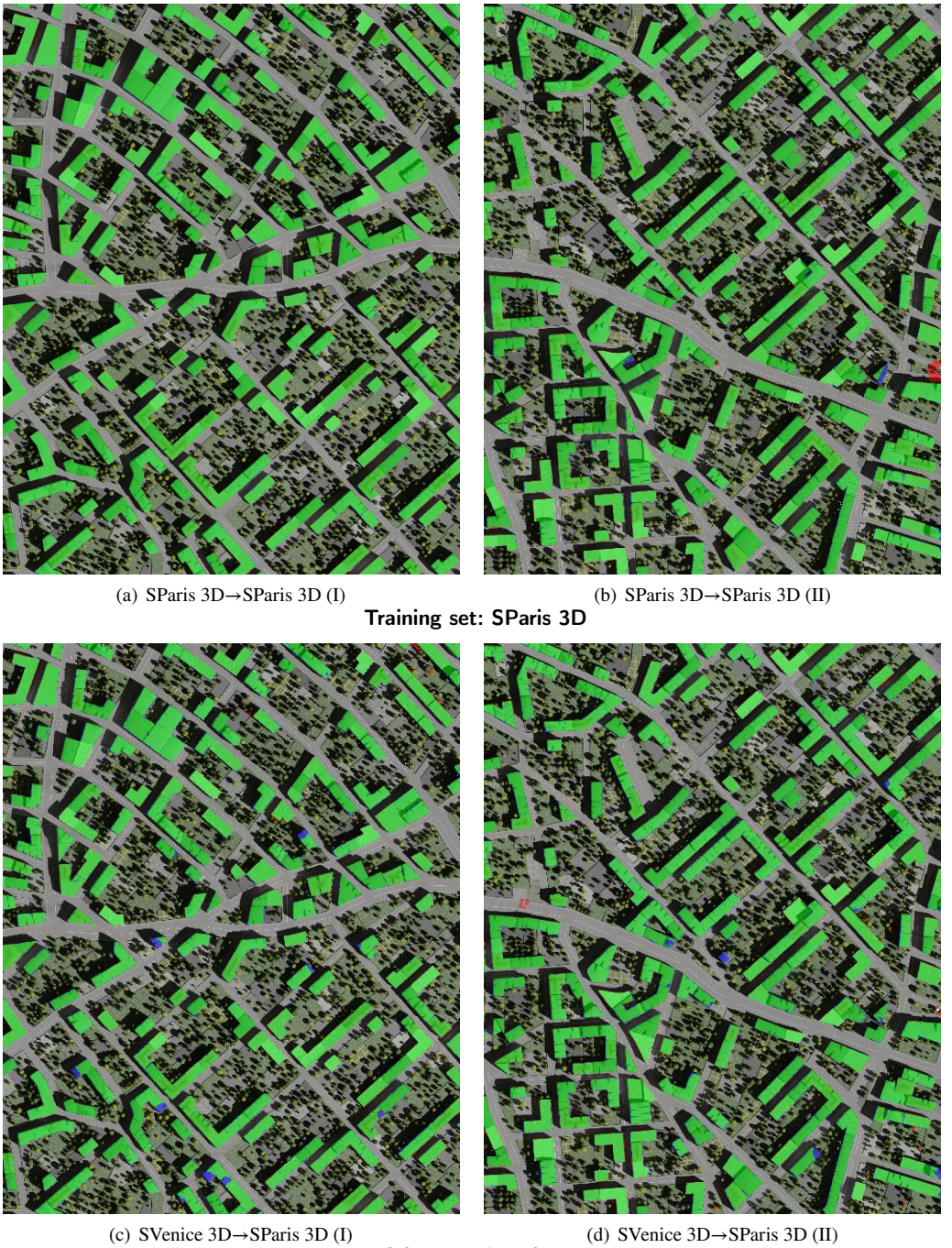
Using a similar workflow as described in Section 5.2, we carry out experiments SParis→SVenice and SVenice→SParis by integrating the DSM-based point clouds, in order to investigate domain shifts between the two synthetic DSMs. Rows 2 and 4 of Table 3 list the quantitative results. Compared to the single-domain case, the score of each metric decreases. In the experiment of SVenice→SParis, precision, F1, and IoU decrease 2.55%, 1.1%, and 2.07% compared with the results of SParis→SParis, respectively. Such decreases in performance appear to be acceptable. According to the qualitative results shown in Fig. 13 (c) and (d), the SparseConvNet model trained on the SVenice data correctly covers all building objects. Compared with the building masks generated by the model of SParis→SParis, it contains more false negative pixels on several building instances. For the SParis→SVenice case, precision, recall, F1, and IoU decrease 11.61%, 12.28%, 11.95%, and 20.38% compared with the results of SVenice→SVenice, respectively. The predicted building masks exhibit non-negligible noise (Fig. 14 (c) and (d)). Several pixels belonging to other classes which are adjacent to buildings are here misclassified as buildings, with the same happening for some pixels belonging to the water semantic class. This phenomenon can be explained by two factors. Firstly, the majority of buildings in the SVenice dataset are smaller with respect to the ones contained in SParis. Consequently, the SparseConvNet model trained on SParis fails at recognizing them. Secondly, the water class is not present in SParis. As a result, several flat water areas in SVenice's DSMs are more easily misidentified as rooftops by the point cloud building extraction network trained on SParis data.

Table 3
SMARS building extraction using DSM-derived point clouds as the input.

Train	Test	Precision [%]	Recall [%]	F1 Score [%]	IoU [%]
SParis	SParis	97.74	97.29	97.52	95.16
SParis	SVenice	86.13	85.01	85.57	74.78
SVenice	SVenice	95.91	94.70	95.30	91.03
SVenice	SParis	95.19	97.68	96.42	93.09

5.6. Cross domain test: SMARS-real 3D data

As illustrated by the qualitative results in Fig. 15, models trained with synthetic data achieve reasonable performance on the ISPRS Potsdam dataset when inspected visually. Nevertheless, partial building structures, which are seldom found in synthetic data, are often not detected, such as four quadrilateral building clusters in Fig. 15 (b). In Fig. 15 (c), such errors appear considerably reduced. Table 4 shows that the IoU and F1 scores for SVenice→Potsdam are 9.98% and 7.11% higher than those for SParis→Potsdam, respectively. This indicates that the SparseConvNet trained on SVenice has better generalization capabilities on real data with respect to the model trained on SParis. However, when compared to the reference results of Potsdam→Potsdam, both models trained on synthetic data still yield a decreased performance. Results further show a good capability for transfer learning, especially for SVenice. This could also be used as a step for pre-training neural networks, and supplementing it with a few additional samples for fine-tuning might alleviate the domain gap. Additionally, including a larger variety of building models in the training data might help to correctly identify some missing shapes.



(a) SParis 3D→SParis 3D (I) (b) SParis 3D→SParis 3D (II)

Training set: SParis 3D

(c) SVenice 3D→SParis 3D (I) (d) SVenice 3D→SParis 3D (II)

Training set: SVenice 3D

Figure 13: Building extraction results of SParis test data using DSM-derived point clouds as the input: (a) and (b) SparseConvNet trained on SParis; (c) and (d) SparseConvNet trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

6. Multi-class semantic segmentation

In order to assess the performance in semantic classes different from buildings, we carry out multi-class semantic segmentation on the 2D optical and point cloud data, with both single- and cross-domain tests.

(a) SVenice 3D→SVenice 3D (I) (b) SVenice 3D→SVenice 3D(II)

Training set: SVenice 3D

(c) SParis 3D→SVenice 3D (I) (d) SParis 3D→SVenice 3D (II)

Training set: SParis 3D

Figure 14: Building extraction results of SVenice test data using point clouds as the input: (a) and (b) SparseConvNet trained with SVenice data. (c) and (d) SparseConvNet trained with SParis data. Legend: True Positive False Positive False Negative. True Negative is not displayed.

Table 4
3D cross-domain study, with SMARS and ISPRS Potsdam datasets as training and testing set, respectively.

Train	**Test**	Precision [%]	Recall [%]	F1 Score [%]	IoU [%]
SParis	Potsdam	67.60	89.50	77.02	62.63
SVenice	Potsdam	80.58	88.00	84.13	72.61
Potsdam	Potsdam	93.75	92.54	93.14	87.17

6.1. 2D multi-class semantic segmentation

The SwinTransformer is here trained on SParis and SVenice using all 5 semantic classes. Quantitative results are listed in Table 5, while segmentation maps are reported in Fig. 16. For the model trained on SParis, all classes except *trees* achieve IoU over 90% on the SParis test set; however, when tested with the SVenice test set, the performance significantly decreases, with the exception of the *trees* class. This indicates a large domain gap between the two datasets, especially regarding *buildings*, *streets*, *lawns*, and *others*. On the contrary, *trees* in both datasets have relatively uniform appearance, thing which can explain the comparatively smaller performance degradation in the cross-domain setting. In the SVenice→SVenice results, the lowest accuracy and IoU scores are associated to the class *streets*: probably, this can be due to the small number of instances of the class*streets* in SVenice, as well as to their different structure. Interestingly, the *lawns* class appears to be the least affected in the SVenice→SParis experiment, while the exact opposite happens

(a) Potsdam 3D→Potsdam 3D (b) SParis 3D→Potsdam 3D (c) SVenice 3D→Potsdam 3D

Figure 15: Building extraction results of ISPRS Potsdam data using DSM-derived point clouds as the input. (a) SparseConvNet trained on ISPRS Potsdam. (b) SparseConvNet trained on SParis. (c) SparseConvNet trained on SVenice. Legend: True Positive False Positive False Negative. True Negative is not displayed.

Table 5
Transferability study of SMARS optical images, 5 classes

Train	Test		Building	Street	Tree	Lawns	Other	Mean
SParis	SParis	IoU[%]	96.07	96.80	85.49	92.97	92.70	92.81
		Acc[%]	97.90	98.54	93.06	96.11	95.88	96.30
SParis	SVenice	IoU[%]	45.37	0.37	83.45	0.01	13.58	28.56
		Acc[%]	72.80	0.66	92.62	0.02	29.84	39.19
SVenice	SVenice	IoU[%]	86.55	56.68	78.95	87.19	87.85	79.45
		Acc[%]	94.87	64.92	86.97	94.84	93.43	87.01
SVenice	SParis	IoU[%]	15.67	10.16	54.06	66.73	17.41	32.81
		Acc[%]	20.02	10.23	66.22	87.92	54.51	47.78

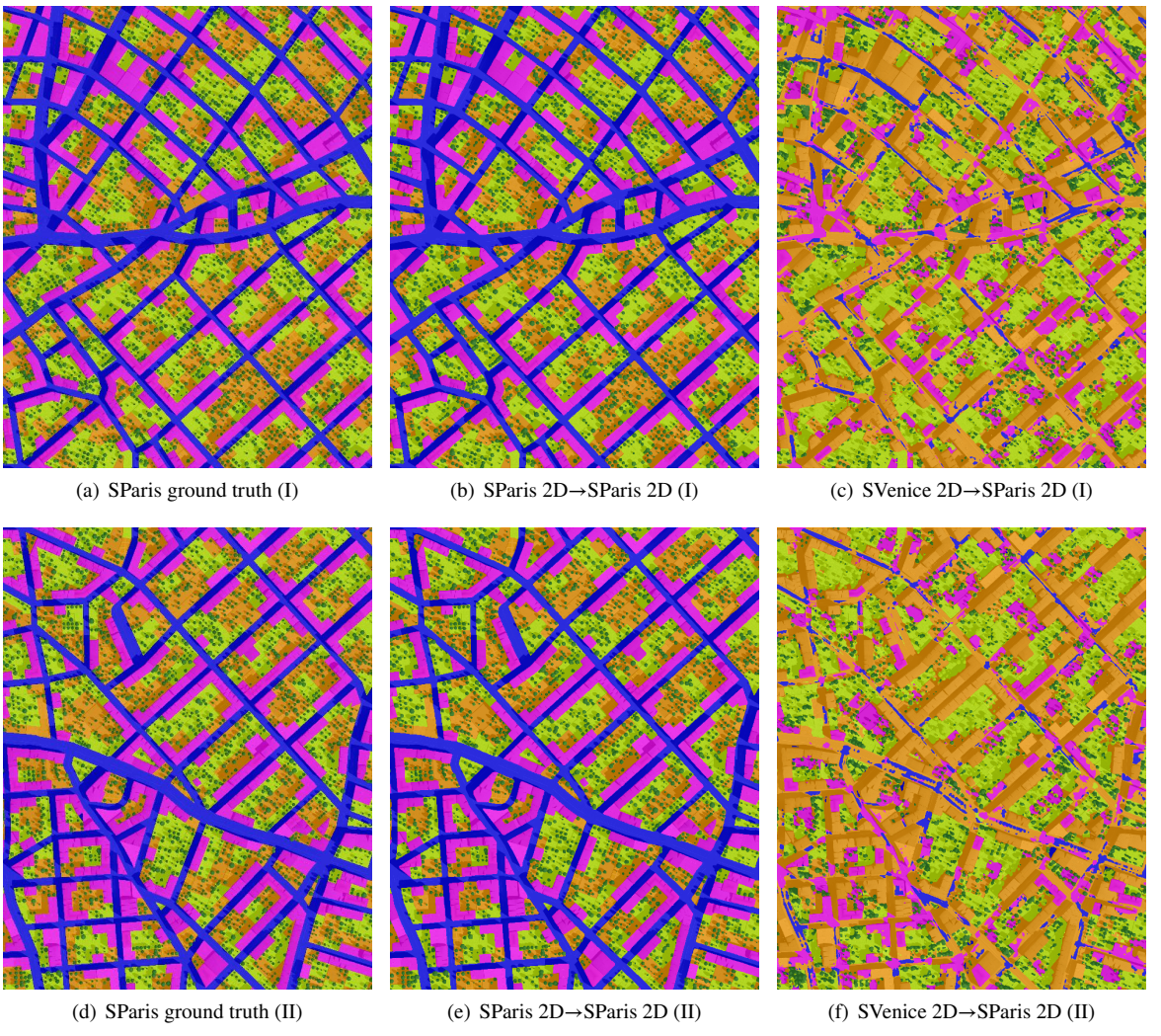
Table 6
Transferability study of SMARS DSM-derived point clouds, 5 class

Train	Test		Building	Street	Tree	Lawns	Other	Mean
SParis	SParis	IoU[%]	94.85	90.00	78.66	46.39	47.48	71.48
		Acc[%]	96.99	96.98	83.26	58.59	69.66	81.10
SParis	SVenice	IoU[%]	72.81	9.26	37.55	33.78	2.33	31.15
		Acc[%]	83.85	48.23	54.55	44.23	2.77	46.73
SVenice	SVenice	IoU[%]	90.04	25.71	80.20	62.89	69.09	65.59
		Acc[%]	97.54	38.64	87.45	83.26	76.15	76.61
SVenice	SParis	IoU[%]	93.19	4.54	75.30	39.45	4.07	43.31
		Acc[%]	96.48	4.75	86.06	84.54	8.43	56.05

for the SVenice→SParis results. Fig. 17 (e) and (f) show the river (belonging to the *others* class) as being mostly misclassified as *street*, due to the absence of water in the SParis dataset; meanwhile, the majority of *streets* and *lawns* are misclassified as *others*, as their structure is different in the SParis dataset.

6.2. 3D multi-class semantic segmentation

In these experiments we train the model including SParis and SVenice DSMs with the described 5 semantic classes using SparseConvNet. Table 6 reports a quantitative assessment of the results for SParis→SParis, SParis→SVenice, SVenice→SVenice, and SVenice→SParis models. In the two experiments having the same source for training and test data, namely SParis→SParis and SVenice→SVenice, SparseConvNet achieves a satisfactory performance for the classes *buildings* and *trees*. SParis→SParis exhibits clearly superior results with respect to SVenice→SVenice for the



(a) SParis ground truth (I) (b) SParis 2D→SParis 2D (I) (c) SVenice 2D→SParis 2D (I)

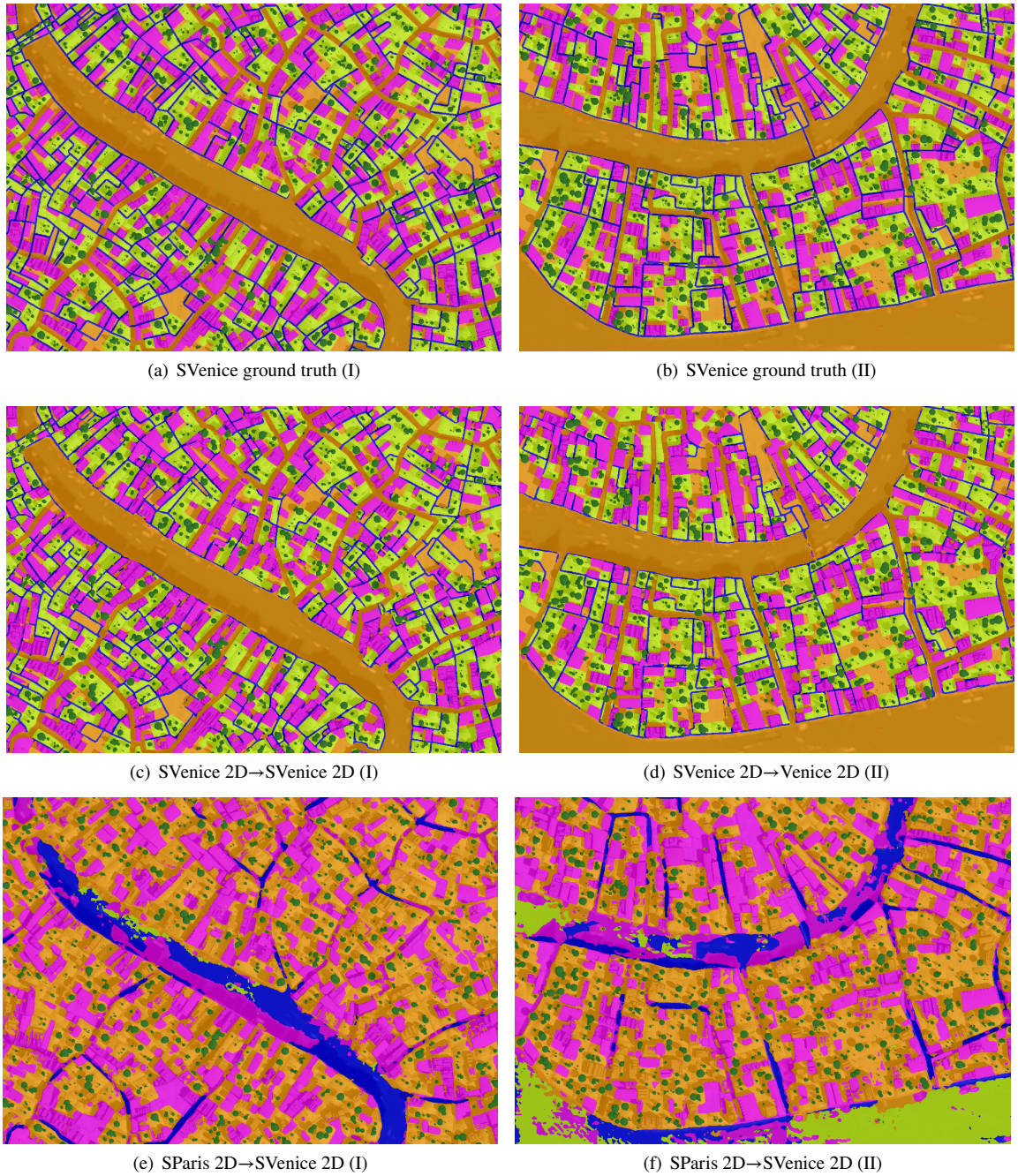
(d) SParis ground truth (II) (e) SParis 2D→SParis 2D (II) (f) SVenice 2D→SParis 2D (II)

Figure 16: Results of image semantic segmentation for SParis (5 classes). Legend: Buildings Street Trees Lawns Other

class *streets*. As mentioned, this is due to the limited number of samples for this class available for training in SVenice. In the cross-domain experiment SParis→SVenice, the performance of each class decreases severely. Among the results of SVenice→SParis, the IoU and accuracy scores for the class *buildings* are excellent, and comparable to the scores achieved in SParis→SParis. This is in line with the results presented for SVenice→SParis building extraction in section 5.5, as SVenice features a wide variety of building sizes covering most of their variability for the respective class in SParis data. The generalization capability of recognizing buildings is preserved in the 5-class semantic segmentation point cloud model. The visual assessment of the results presented in Fig. 18 suggests that the model trained with SVenice is not optimal to recognize streets in the DSMs of SParis. In Fig. 19, most of the areas covered by water are classified as streets in SParis→SVenice. This is because SParis lacks training data for this class, as discussed in section 5.5.

7. Building change detection

In order to assess the feasibility of SMARS for 3D change detection applications, in this section change indicators from both 2D and 3D data are extracted and evaluated. In addition, we present several state of the art change detection approaches for comparison (Tian and Dezert, 2019).



(a) SVenice ground truth (I)

(b) SVenice ground truth (II)

(c) SVenice 2D→SVenice 2D (I)

(d) SVenice 2D→Venice 2D (II)

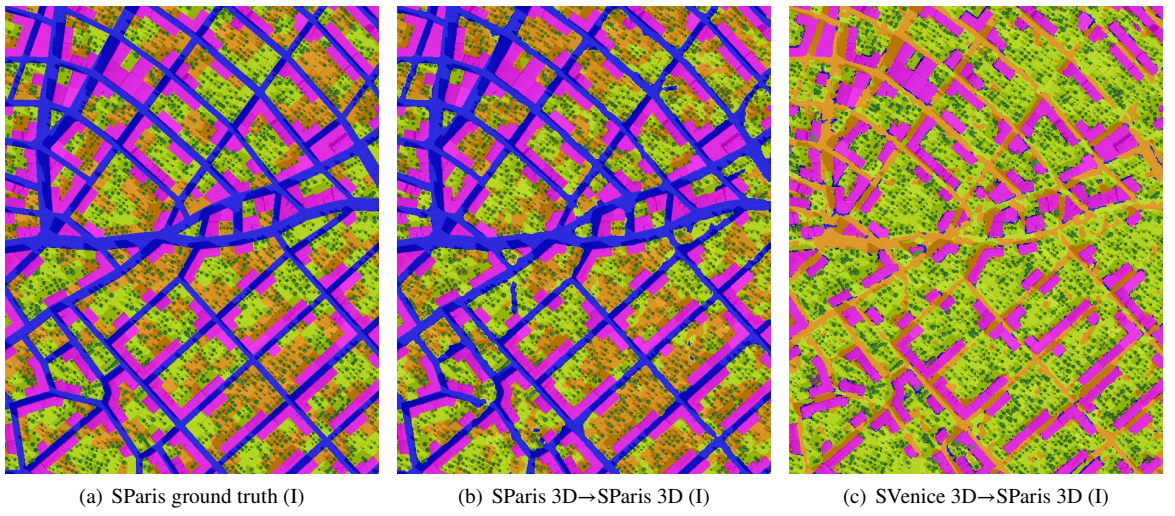
(e) SParis 2D→SVenice 2D (I)

(f) SParis 2D→SVenice 2D (II)

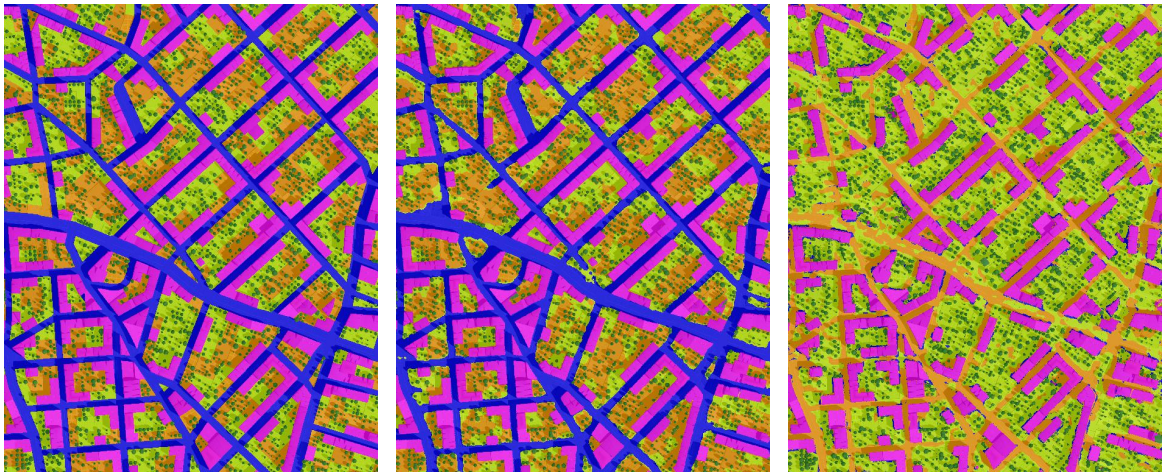
Figure 17: Results of image semantic segmentation for SVenice (5 classes). Legend: Buildings Street Trees Lawns Other

7.1. Robust height differences

As detailed in our previous work (Tian et al., 2013), the quality of the pre- and post-event DSMs can exhibit relevant differences according to GSD, sensors characteristics, illumination conditions, stereo viewing angles and other parameters of the multi-view images from which the DSMs are generated. Hence, methods based on pixel-based subtraction do not in all cases deliver ideal results (Tian et al., 2013; Qin et al., 2016). Thus, robust distance measurements yielding a refined height change indicator have been proposed. The main motivation of the experiments reported in this section is assessing the differences between DSMs generated from synthetic and real data, along with



(a) SParis ground truth (I) (b) SParis 3D→SParis 3D (I) (c) SVenice 3D→SParis 3D (I)



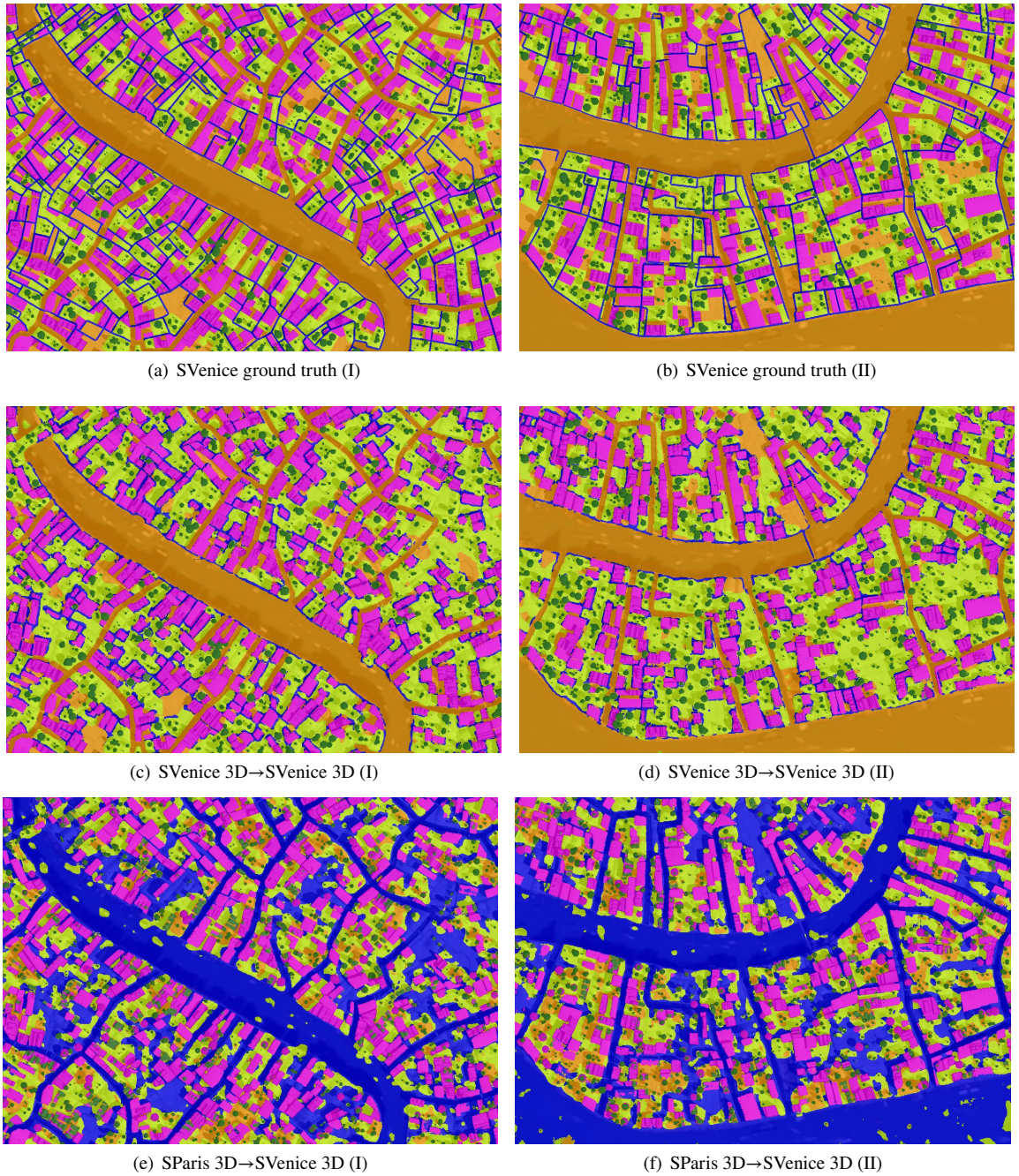
(d) SParis ground truth (II) (e) SParis 3D→SParis 3D (II) (f) SVenice 3D→SParis 3D (II)

Figure 18: 5-class semantic segmentation results of SParis test data using DSM-derived point clouds as the input. (a) and (d) Ground truth. (b) and (c) SparseConvNet trained with SParis data. (e) and (f) SparseConvNet trained with SVenice data. Legend: Buildings Street Trees Lawns Other

their impact on practical applications. We compare the robust height differences proposed in (Tian et al., 2013) (window size set to $w = 5$) to the use of direct height difference (considering only positive height changes). In addition, the pre- and post-event images are "acquired" with similar settings by the virtual camera, such as GSD and different illumination conditions, lowering the impact of the sources of errors when using methods based on direct subtraction of the DSMs. Nevertheless, in Fig. 20, results obtained by applying robust height differences appear superior, as they exhibit reduced noise in the building boundary regions.

7.2. Building change mask generation

In order to further assess the quality of the proposed data for 2D and 3D change detection applications, extended experiments with different change detection approaches are summarized in this section. In this paper, we test direct height differences with threshold values manually and automatically selected to generate positive building change masks for the test regions. In addition, 2D change detection results are extracted and evaluated using the state of art Interactively Reweighted Multivariate Alteration Detection (IR-MAD) (Nielsen, 2007). For the case of fusion-based change detection approaches, we follow the method proposed in (Tian and Dezert, 2019), which employs the decision fusion model to combine the 2D and 3D change indicators. Three decision criteria are considered, including



(a) SVenice ground truth (I)

(b) SVenice ground truth (II)

(c) SVenice 3D→SVenice 3D (I)

(d) SVenice 3D→SVenice 3D (II)

(e) SParis 3D→SVenice 3D (I)

(f) SParis 3D→SVenice 3D (II)

Figure 19: 5-class semantic segmentation results of SParis test data using DSM-derived point clouds as the input. (a) and (d) Ground truth. (b) and (e) SparseConvNet trained with SVenice data. (c) and (f) SparseConvNet trained with SParis data. Legend: Buildings Street Trees Lawns Other

Maximum of Belief (MaxBel), Maximum of Plausibility (MaxPl) and Maximum of Betting Probability (MaxBetP). In order to calculate the Basis Belief Assignments (BBAs) of the concordance and discordance indices, two thresholds are required. In our previous work (Tian and Dezert, 2019), we adopt an extension of Otsu thresholding to project the change indicators to a sigmoid distribution. As here the training data are provided by SMARS, we use them to automatically calculate the two thresholds for each change indicator, namely the mean value of the change indicators

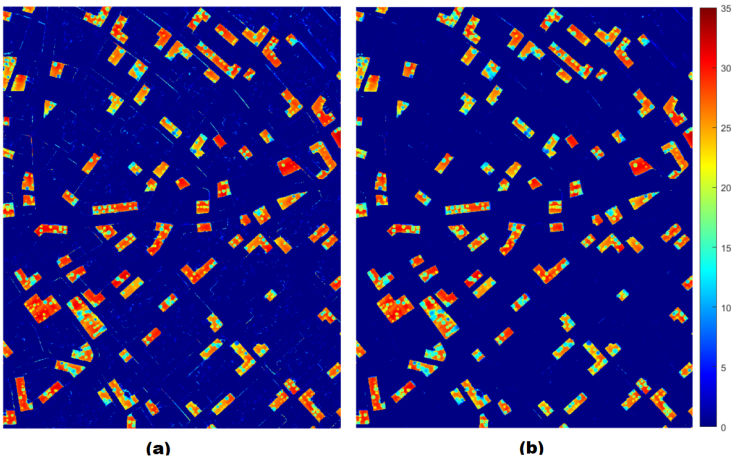
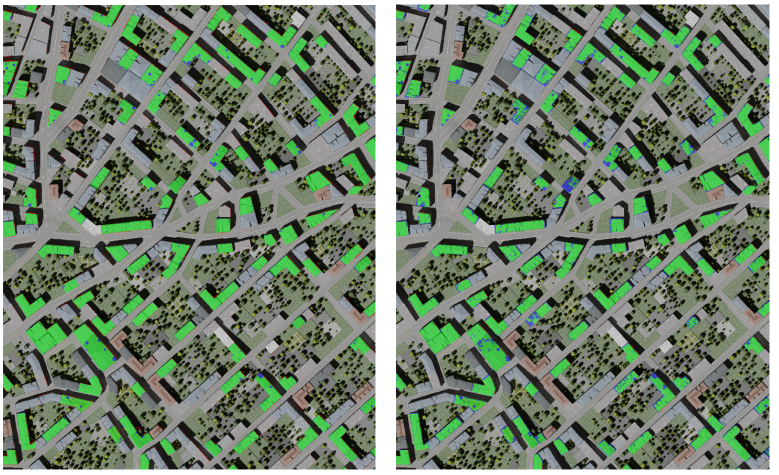


Figure 20: Positive height differences: (a) direct subtraction; (b) Robust height differences with ($w = 5$)

.

for each class (change (T_0), no-change(T_1)). We refer to this approach as automatic threshold values selection (AUTO), and set $T = mean(T_0, T_1)$ for the height differences and IR-MAD, separately.

The performance of the difference change detection approaches is evaluated based on overall accuracy (OA), kappa accuracy (KA) and IoU (Table. 7). Each synthetic image has two test regions, which are marked as AOI (I) and AOI (II) in Table. 7, respectively. SParis appears to be an easier test region, featuring mainly high-rise and well-separated buildings. In addition, the buildings are considerably higher than most of the trees, introducing a relevant increase in height in the transitions from trees to buildings. Therefore, the direct height differences with automatic thresholding approach (Hdiff (AUTO)) achieve the best accuracy according to the figures of merit listed in Table 7. However, a visual assessment of Fig. 21 reveals that the decision fusion results present a reduced amount of false positives, especially around building boundary regions. Further details are reported in Fig. 22. The best results are achieved by directly comparing the two building masks derived from Section 5.5: we refer to this case as "Post-classification" in Table. 7.



(a) Height difference (b) DS fusion

Figure 21: 3D change detection results of SParis (I) generated by direct height difference (a) and decision fusion (b). Legend: True Positive False Positive False Negative. True Negative is not displayed.

SVenice is a challenging test region for 3D change detection compared to SParis, as it features small-sized buildings and narrow streets. In addition, the trees are sometimes taller than nearby residential buildings, resulting in negative

Test Regions	Methods	AOI(I)			AOI(II)		
		OA[%]	KA[%]	IoU[%]	OA[%]	KA[%]	IoU[%]
SParis	HDiff>5m	97.83	90.85	85.36	97.91	90.12	84.00
	Robust HDiff(AUTO)	98.41	92.84	88.24	98.49	92.34	87.25
	Hdiff(AUTO)	98.45	93.22	88.87	98.49	92.55	87.63
	IR-MAD (AUTO)	85.87	43.84	35.10	86.42	40.05	31.29
	Decision Fusion-MaxBel	98.08	91.07	85.47	98.13	90.07	83.67
	Decision Fusion-MaxPl	98.04	90.89	85.18	98.10	89.94	83.48
	Decision Fusion-MaxBetP	98.09	91.10	85.51	98.14	90.16	83.82
	Region- DS-MaxBel	91.46	67.15	56.29	91.54	62.50	50.66
	Post-classification	98.86	94.99	91.64	98.78	93.95	89.83
SVenice	HDiff>5m	93.30	77.26	68.54	94.30	76.47	66.37
	Robust HDiff(AUTO)	93.54	77.07	68.02	94.40	75.68	65.15
	Hdiff(AUTO)	93.19	77.02	68.13	94.24	76.39	66.31
	IR-MAD (AUTO)	85.90	36.26	27.27	87.36	32.74	24.19
	Decision Fusion-MaxBel	93.38	75.84	66.36	94.39	75.01	64.21
	Decision Fusion-MaxPl	93.39	75.84	66.36	94.36	74.81	63.98
	Decision Fusion-MaxBetP	93.37	75.80	66.32	94.37	74.89	64.07
	Region- DS-MaxBel	90.70	69.01	59.60	91.45	66.12	55.17
	Post-classification	96.94	89.53	84.14	97.57	90.03	84.24

Table 7
Results of different change detection approaches on SParis and SVenice.

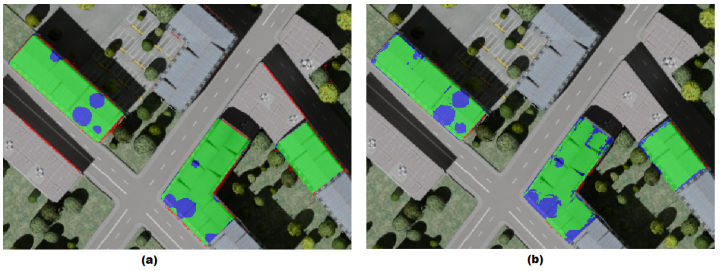
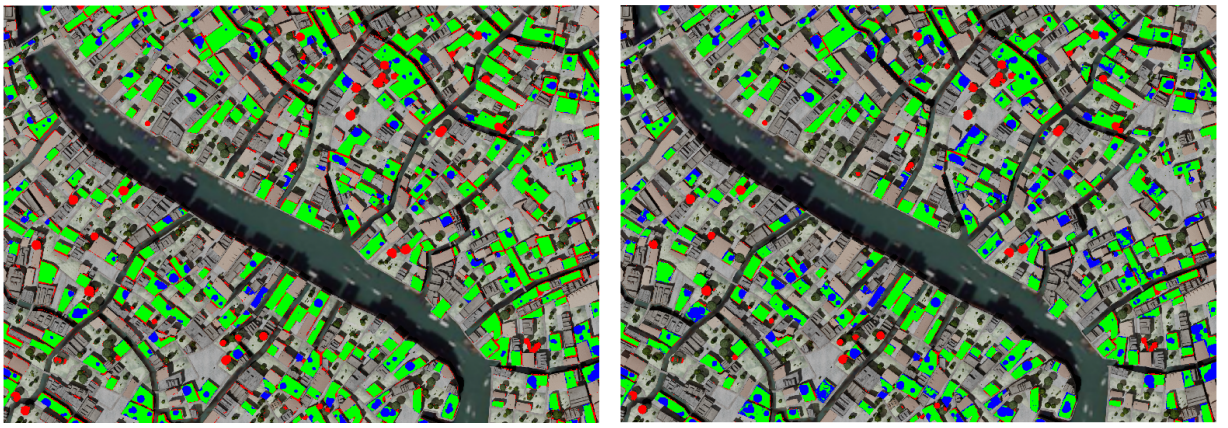


Figure 22: Comparison of results obtained for single buildings: (a) direct height differences (b) DS fusion. Legend: True Positive False Positive False Negative. True Negative is not displayed.

height changes for newly constructed buildings. This rarely occurs in the SParis dataset (see in Fig. 22). Moreover, the *water* class occupying around 5 % of each test region is not defined for this dataset. Differences between *water* and other semantic classes are particularly evident in the synthesized optical images, which were simulated relying on low-resolution satellite data. The 2D change detection results have an associated IoU of 27.27 % and 24.17% in the two test regions, respectively, confirming the impact of differences in illumination conditions between the pre- and post-event images on the final results. When applied on the SVenice data, robust height differences achieve slightly higher accuracy with respect to fusion-based approaches. Nevertheless, Fig. 23 shows that both height differences and DS fusion detect regions as false negatives, if a tall tree is replaced by a building in the post-image. Additionally, a relevant number of new trees is detected as newly constructed buildings (highlighted in red in Fig. 23), as these match both conditions of having an increased height and exhibiting changes in the spectrum of the optical data. In a similar way to the experiments carried out on SParis, the differences in performance between the three decision approaches are not obvious. Relying on the accurate 2D/3D multimodal building detection result of section 5.5, post-classification clearly outperforms other approaches, achieving an IoU equal to 84.14 % and 84.24 % in the two test regions, respectively. The second test region of SVenice is presented in Fig. 24 (b), in which most of the newly constructed buildings are correctly identified.



(a) Height difference (b) DS fusion

Figure 23: 3D change detection result for SVenice (I) generated by direct height difference (a) and decision fusion. Legend: True Positive False Positive False Negative. True Negative is not displayed.

In order to reduce false negatives for newly constructed buildings, we test region-based 3D change detection by fusing the post-event building mask with the fusion-based change detection results. As all three DS fusion methods yield similar results, we only report results obtained with Decision Fusion-MaxBel for the following region-based change detection experiment. Buildings belonging to the post-event building mask are considered as newly constructed if more than 30 % of their pixels belongs to the "building change" category in the pixel-based change detection results. The performance of the region-based change detection approach is rather poor for both SParis and SVenice, as shown in Table. 7. This can be explained by examining Fig. 24, where a relevant number of newly constructed buildings are connected to the unchanged buildings in the virtually simulated environment. Therefore, a relevant number of both false positives and negatives are introduced when averaging the change decisions in these regions.

(a) Region-based DS-MaxBel (b) Post-classification

Figure 24: 3D change detection result of SVenice-AOI2 relying on region-based approaches (a) and post-classification results (b). Legend: True Positive, False Positive and False Negative. True Negative is not displayed.

8. Discussion

This paper proposes a novel workflow for synthetic data generation filling the gaps in the available 2D/3D multimodal data for building extraction, multi-class semantic segmentation and 3D change detection. Our data analysis goes in two directions: 1) the feasibility of using SMARS to evaluate the efficiency of existing approaches for building

extraction, multi-class semantic segmentation and building change detection and 2) the effects of the domain gap when the models trained on our synthetic data are tested on real data.

8.1. Quality of the synthetic dataset

This subsection discusses the main advantages and disadvantages of the rendered images described in section 3. The proposed SMARS dataset meets our expectations in most of the reported experiments. Nevertheless, it also presents some limitations. Both will be discussed below for each of the available semantic categories in SMARS.

8.1.1. Buildings

The buildings generated by CityEngine exhibit good quality in terms of geometry, architectural appearance, and textures. They can be favorably compared to models with LoD2 and LoD3, as some rooftops have additional features such as chimneys. Moreover, the buildings resemble the expected distribution of a city in terms of size and arrangement and contribute to creating realistic scenarios. Taking into account the options to manipulate the building properties, it is easy to simulate the city growth as required for the change detection task. Furthermore, as *buildings* achieve a very good reconstruction in the DSMs, they can be easily detected by the algorithms considered in this article.

Nonetheless, the pool of textures to generate the *buildings* is limited and might lead to overfitting in the learning process. Besides, no construction sites are part of the dataset, as would be the case for real images; these regions represent a challenge for change detection depending on the progress of the constructions. Another constraint is given by the generation of mostly residential buildings, as facilities such as commercial buildings, parks, sports centers, or transport stations are not included in our dataset.

In the experiments, we notice that the discrepancy in height between the two city models leads to errors for prediction in the learning models, as the DSMs values have different ranges. With traditional approaches, the similarity in height between *trees* and *buildings* can also increase the challenges of classification, especially when they are close to each other. In the SMARS dataset, the building roofs are generally well visible and do not suffer from occlusion problems as in real data, making the task of building extraction easier.

8.1.2. Street

A major difference between the two models is the street category. In SParis the streets match the common design with sidewalks, concrete material, and broken and solid lines. Besides, streets in this model are wide and have a height profile different from all other elements, with the exception of lawns.

SVenice is more difficult in this category. In the same way as the real city, the streets are designed for pedestrians, and are therefore narrow, causing sidewalks to be absent and are not marked either by broken or solid lines. Additionally, the width of the streets is comparable to the one of the multiple canals crossing the city. This problem is aggravated by the similarities in terms of height between the "others" (where canals and sea are included) and street categories. Because of that, we can notice in the semantic segmentation task that cross-domain experiments drop significantly in performance for this category. For learning models trained with SParis, the canals of SVenice are considered streets and the lawns are predicted as "others". Likewise, for learning models trained with SVenice, the streets of SParis are many times wrongly labeled as "others" and only a few streets are actually detected.

As width and height are within the expected ranges for streets, a suitable solution would be to enhance the available categories in order to incorporate canals, squares, roundabouts, alleys, and other elements that could be confused with roads.

8.1.3. Vegetation and lawns

Representation of shapes and structures of trees and bushes in 3D is a critical issue. A detailed representation requires a complicated geometric definition leading to high computational costs. A common simplified case with only two intersected vertical planes greatly reduces the memory requirements but exhibits poor visual quality in the models. Due to the trade-off between memory and appearance, we used textured ellipsoids. This allows the inclusion of a large number of trees and bushes in the virtual scenes. We include many textures, but these are limited to a specific number of plant species.

Yet, the *vegetation* regions largely suffer from the domain gaps between synthetic and real data. Real scenes have no simplified geometry (with the exception of man-trimmed trees) and cannot be easily modeled. Using only ellipsoids makes the learning biased towards this shape, and cannot adequately lead to correct predictions of other types of vegetation. Also, seasonal effects (such as leave colors, snow covering, or fallen leaves) are not considered.

On top of that, the *lawns* category has been simplified too. Actual grass has a non-negligible height (even if this is relatively small in comparison to the other objects), no uniform texture, and can include small vegetation such as low bushes. For the simulated cities, the *lawns* are simplified by a flat area with grass-like texture, which appears realistic enough in the orthophotos. Without the texture, the *lawns* would be similar to the *roads* or *bare soil* category, as the height information of *lawns* is set close to 0.

In DART, *trees* are defined by tree species, various attributes of trunk and crown, and are simulated using turbid voxels or isosceles triangles (Gastellu-Etchegorry et al., 2015). Tree crown shapes can be chosen from ellipsoidal, ellipsoid-composed, truncated cone, trapezoid, and cylinder with truncated cone. In addition, branches and twigs can be added. However, the tree modeling requires many manual input and is still not realistic as desired. Nevertheless, there is still potential to improve the quality of the *trees* class by using existing detailed 3D tree models. For example, the RAdiation transfer Model Intercomparison (RAMI) experiments derived detailed and realistic 3D models of various tree species by in situ measurements. The 3D models have been exported to DART, and can be edited in Blender as well. But those tree models do not include enough typical urban tree species to represent the urban tree scenario. For the reasons described above, we did not adopt these accurate 3D tree models.

8.1.4. Water

Water is not an annotated category in our SMARS dataset. However, it is an important land cover type in the SVenice scene. In the provided Venice city model of CityEngine, the water bodies are actually covered by a real low-resolution satellite image, exhibiting shadows that might not correspond to the simulated sun conditions. In addition, elements present in the water (such as boats and bridges) do not have an above ground height, so the captured multi-view images do not present a meaningful disparity in the epipolar image pairs. Therefore, in the generated DSMs the surface of water bodies is rather flat and smooth. In reality, the elements present in the water would have a height value larger than zero.

On the other hand, the SParis model has no *water*, so these are absent in the ground truth for either city, an aspect which can lead to errors in the semantic segmentation task, especially for cross domain experiments. It is particularly complex for the algorithms to separate water from streets in the SVenice model, where the canals have similar contextual features as the *streets* in SParis. The collection of a larger number of samples with labeled *water* coverage might help solve this issue.

Finally, since we use an aerial photo as the source for the water areas, these do not change between the pre- and post-models and remain also constant within the simulated flight campaigns. In reality, the waves and tides produce an irregular surface, causing the matching algorithms to yield poor results. Usually, the DSM pipeline would fail to reconstruct such regions, while our DSM has a constant value. As discussed above, a physical simulation of water would lead to enhanced realism in the scenarios. Since our work focuses mainly on buildings, this is currently left out of our studies.

8.2. Single domain test

In single-domain building extraction experiments, both optical image and point cloud methods produce satisfactory results. As training and testing data share common features, very precise results are therefore derived for the task of building segmentation. The optical images have slightly better performance concerning 3D point clouds, as the images exhibit denser features in comparison while the point clouds have sparse representations.

Buildings in SParis and SVenice exhibit large variability in roof texture and their details, or the size and shape of the buildings. As a result, the evaluation metrics show that building extraction is less complex for SParis with respect to SVenice. The situation is more complex for the multi-class segmentation experiments. In SParis→SParis, the use of optical images achieves a mean IoU above 90%, while information from the 3D point cloud underperforms, with a mean IoU of 71%. The most problematic classes appear to be *lawns* and *background* classes. This suggests that point cloud features alone are not sufficient to represent some of the classes. For the case of SVenice→SVenice, both 2D and 3D methods exhibit relatively poor performance for the class *street*, as these are predominantly narrow pedestrian walks, which can be easily confused visually with the stone-paved square, belonging to the class *background* (Fig. 9(b)). Similarly, point cloud features of *streets* are not discriminative enough to allow separating this semantic class from the others. Different results are obtained for the class *buildings*: here, the 3D method can achieve satisfactory results not only for building extraction but also for multi-class semantic segmentation, indicating a good ability of point clouds to characterize features relating to man-made regular objects. However, it is worth noting that optical image analysis still outperforms the 3D method, achieving slightly higher IoU scores in all single-domain test scenarios, except

for SVenice→SVenice multi-class semantic segmentation. These differences are observed in the building extraction experiments of SParis→SParis and SVenice→SVenice, as well as the multi-class semantic segmentation experiment of SParis→SParis, with differences of 0.57%, 1.06%, and 1.22%, respectively. This is due to the reason that building objects in DSMs are easily confused with other classes having similar heights by geometric features. In Fig.13 (b), an evident false positives area is noticeable at the right border of the image, where several trees are incorrectly recognized as buildings. In Fig.10 (b), no such error is present. In addition, due to limitations of the matching algorithms, some building boundaries in DSMs are incomplete and missing a few pixels (Tian et al., 2013; d'Angelo and Reinartz, 2011), leading to more false negatives in a 3D single-domain test when compared with optical image analysis.

The difference in performance between the binary and the multi-class segmentation lies partly in the optimization. It is intrinsically more difficult to optimize a multi-class problem with respect to a binary one, which results in a longer convergence time and less definite decision boundary. In addition, as the optimizers take into account the loss values of all classes, the gradient for weight update is different from the binary building extraction experiment.

In conclusion, from the single domain experiments performed we do not observe particular differences from the use of real multimodal data. Therefore, we can conclude that the SMARS dataset could be suitable as a training dataset for multimodal remote sensing tasks. Compared with SParis data, SVenice dataset is more challenging.

8.3. Cross domain test

In the remote sensing field, domain gap or shift is a common challenge for deep learning models. Preparing labeled datasets is normally costly and time-consuming, therefore many weakly and semi-supervised learning approaches are proposed by utilizing existing benchmark datasets (Li et al., 2022). However, target and source domain datasets may be different in terms of city styles, ground object types, seasonal changes, or characteristics of the acquiring sensors, leading to widespread attention of domain adaptation in recent years (Tuia et al., 2016). The lack of benchmark datasets hinders in-depth research in this field, especially for domain adaptation of the joint use of 2D/3D multimodal datasets. The experiments show that the two synthetic data generated using the proposed approach, namely SParis and SVenice, have clear domain gaps, and the results of 5-class semantic segmentation still have significant room for improvements in both 2D and 3D experiments. For example, for the SParis→SVenice and SVenice→SParis scenarios, it is common for streets to be confused with other classes.

For building extraction tasks, the 2D version is suitable for testing domain adaptation methods, while the 3D version of the SParis→SVenice case can be further refined based on baseline methods. The synthetic→real workflow is a challenge presenting wide opportunities for its exploration. Training with synthetic data and testing on real data can significantly reduce the cost of annotating training samples. Likewise, training with real data and testing on synthetic data for evaluating models can greatly reduce the cost of annotating testing samples, which typically require higher accuracy. Furthermore, the reference data associated to the generated synthetic data is ensured to be free from annotation errors. Therefore, this benchmark provides a starting point for the remote sensing community to investigate such topics.

When using different baseline methods, 3D data are more robust to domain shifts for buildings with respect to optical 2D images, while the opposite happens for single-domain tests. Point cloud networks, which are based on geometric features, have better generalization abilities in building extraction tasks for unseen domains, as they are not influenced by possible confusion between spectral features. For instance, in Fig. 16, the image network wrongly recognizes several roads as buildings, as their colors and 2D geometry are similar, while such errors do not occur in the results derived from the point cloud network. The point cloud network SparseConvNet outperforms the image network Swin Transformer for the building class in the synthetic→real building extraction and 5-class semantic segmentation cases. As illustrated in Fig. 12(f), a non-building object is misclassified as a building due to the lack of geometric information, while the prediction from the point cloud network is correct.

Cross-domain results are similar to what would be expected to achieve using real data, demonstrating the feasibility of the SMARS datasets to be integrated into practical applications employing real images. In this paper, no new domain adaptation approach is proposed: we encourage other researchers to test their approaches on this dataset, or prepare their own synthetic data with the proposed approach for their test regions of interest.

8.4. 3D building change detection

Recent years witnessed an increase in demand for accessible and high quality 3D dataset (Tian et al., 2013; Tian and Dezert, 2019; Xie et al., 2020) . Their multi-temporal availability represents a desired feature enabling applications to 3D change detection, where the accuracy of the results is increased by the provided information on targets height,

complementary to the spectral information conveyed by optical earth observation data (Qin et al., 2016). Nevertheless, the lack of available benchmark datasets of this kind makes the development of 3D change detection approaches difficult, especially the ones relying on deep learning, as demonstrated by their scarcity in literature. The production of data for 3D change detection presents several problems. On the one hand, large cities in developed countries have limited changes, not sufficient to train a deep network (Tian et al., 2013). On the other hand, in developing countries building changes are often confused with different categories of changes, such as construction of highways and train stations, hindering their correct annotation. In addition, 2D/3D multimodal multi-temporal datasets are generally expensive to acquire, and several research institutes collect new data in the frame of specific projects: therefore, they can not easily disclose them as publicly available benchmarks.

This paper presents a novel workflow to generate synthetic data suitable for training classification algorithms for 3D change detection. The illumination conditions of the simulated optical images present relevant differences, making this task non-trivial for algorithms relying solely on spectral changes. Pre- and post- event data are almost perfectly co-registered, allowing the user to remove this source of error propagation in their change detection workflow, which must be dealt with when using real data.

The introduced SMARS dataset presents aspects which may be improved in the future. Regarding the intrinsic quality and rendering of the data, results show that DSMs exhibit sharp boundaries and a reduced number of occluded areas with respect to typical real digital elevation models. Regarding the content of the scenes, in SParis most of the building blocks have been extended or partially removed in the transition from the simulated pre- to the post- images, and the changes are evenly distributed throughout the entire virtual city. This usually does not correspond to the pace and distribution of urban pattern changes in the real world.

The reported experiments suggest that traditional machine learning approaches are not optimal at detecting building changes relying on optical images only, as no elevation data are available. The use of high quality DSMs increases the accuracy of the results: however, when using only the generated synthetic DSMs, changes in buildings are often confused with changes in trees, keeping this task highly challenging. In this paper, the best change detection results are obtained by employing both simulated optical data and their associated DSM, by directly comparing the pre- and post-event building masks generated by multimodal co-learning approaches.

9. Conclusion

In this paper we introduce SMARS, a synthetic large and accurately annotated 2D/3D multi-temporal earth observation dataset, as an effort to meet the demand for multimodal benchmark data suitable for change detection applications in urban areas. In addition to 3D change detection, we provide orthorectified images, DSMs and ground truth for semantic segmentation, along with a pipeline to generate similar synthetic images resembling the characteristics of real aerial acquisitions, including their limitations. By modifying the scenes within the pipeline, it is easy to set and adjust the changes between two simulated acquisition times, which is a difficult task when using real data. As a result, the pipeline has the potential to create larger samples with high variability. As the main goal of this paper is the generation of synthetic 2D/3D multimodal data as similar as possible to real data, deep-learning based 3D change detection approaches are not discussed here.

The ground truth associated to the dataset is free from wrongly annotated labels or confusion between classes, being generated during the rendering process. This aspect propagates its advantages to the change detection applications, where a large number of modifications can be handled and are ensured to be correct in the change mask to be used as reference. The quality of the presented synthetic data has been investigated in several experiments, which yielded results similar to what would be expected using real data. The quality of SMARS data is high in terms of coregistration, orthorectification and ground truth quality.

In addition to testing segmentation and change detection approaches, the presented synthetic data can be adapted to train a valid building extraction or semantic segmentation model that can be applied to real datasets. For instance, building extraction shows a good performance on the ISPRS Potsdam dataset, even without a fine-tuning step. Considering the 3D case, most of the buildings are properly classified with sharp boundaries. However, land cover classes not present in the synthetic data were not properly handled by the networks and lead to wrong classification. In terms of multi-class semantic segmentation, we observed a good performance within the same domain, but this decreased when using cross-domain datasets. Besides, it is not a trivial task to evaluate the transferability since the semantic classes present are different in the considered datasets. Further reducing the domain gaps between real and synthetic data, as well as increasing the available number of classes could help to overcome these difficulties. On the

other hand, for the building semantic segmentation experiments, we observe good results as most of the classes have been properly predicted, with the exception of building edges and vegetation for some cases. In general, the synthetic data represent a feasible option for training neural networks for building detection, semantic segmentation, and change detection tasks, in spite of the described constraints due to domain gaps.

Acknowledgement

Mario Fuentes Reyes is currently funded by a DLR-DAAD Research Fellowship (No. 57478193) to pursue his PhD studies. Yuxing Xie was supported by a DLR-DAAD Research Fellowship (No. 57424731).

References

Almutairi, A., Warner, T.A., 2010. Change detection accuracy and image properties: a study using simulated data. Remote Sensing 2, 1508–1529.

Bachhofner, S., Loghin, A.M., Otepka, J., Pfeifer, N., Hornacek, M., Siposova, A., Schmidinger, N., Hornik, K., Schiller, N., Kähler, O., et al., 2020. Generalized sparse convolutional neural networks for semantic segmentation of point clouds derived from tri-stereo satellite imagery. Remote Sensing 12, 1289.

Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Computers & Geosciences 22, 795–799.

Börner, A., Wiest, L., Keller, P., Reulke, R., Richter, R., Schaepman, M., Schläpfer, D., 2001. SENSOR: a tool for the simulation of hyperspectral remote sensing systems. ISPRS Journal of Photogrammetry and Remote Sensing 55, 299–312.

Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Urban change detection for multispectral earth observation using convolutional neural networks, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2115–2118.

Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y., 2019. Multitask learning for large-scale semantic change detection. Computer Vision and Image Understanding 187, 102783.

Chen, L., Liu, F., Zhao, Y., Wang, W., Yuan, X., Zhu, J., 2020. VALID: A comprehensive virtual aerial image dataset, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 2009–2016.

Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L., 2022. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset, in: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022, BMVA Press.

Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering 5, 493–497.

Coletta, V., Marsocci, V., Ravanelli, R., 2022. 3dcd: a new dataset for 2d and 3d change detection using deep learning techniques. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B3-2022, 1349–1354.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223.

d'Angelo, P., Reinartz, P., 2011. Semiglobal matching results on the ISPRS stereo matching benchmark. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVIII-4/W19, 79–84.

Denninger, M., Sundermeyer, M., Winkelbauer, D., Olefir, D., Hodan, T., Zidan, Y., Elbadrawy, M., Knauer, M., Katam, H., Lodhi, A., 2020. Blenderproc: Reducing the reality gap with photorealistic rendering, in: International Conference on Robotics: Sciene and Systems, RSS 2020.

Disney, M., Lewis, P., Saich, P., 2006. 3D modelling of forest canopy structure for remote sensing simulations in the optical and microwave domains. Remote Sensing of Environment 100, 114–132.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An open urban driving simulator, in: Proceedings of the 1st Annual Conference on Robot Learning, PMLR. pp. 1–16.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88, 303–338.

Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., Cucchiara, R., 2021. Motsynth: How can synthetic data help pedestrian detection and tracking?, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10849–10859.

Fuentes Reyes, M., D'Angelo, P., Fraundorfer, F., 2022. Syntcities: A large synthetic remote sensing dataset for disparity estimation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 10087–10098.

Gastellu-Etchegorry, J.P., Yin, T., Lauret, N., Cajgfinger, T., Gregoire, T., Grau, E., Feret, J.B., Lopes, M., Guilleux, J., Dedieu, G., et al., 2015. Discrete anisotropic radiative transfer (dart 5) for modeling airborne and satellite spectroradiometer and lidar acquisitions of natural and urban landscapes. Remote Sensing 7, 1667–1701.

GDAL/OGR contributors, 2022. GDAL/OGR Geospatial Data Abstraction software Library. Open Source Geospatial Foundation. URL: https://gdal.org, doi:10.5281/zenodo.5884351.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research 32, 1231–1237.

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.

Ghamisi, P., Höfle, B., Zhu, X.X., 2016. Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 3011–3024.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xbd: A dataset for assessing building damage from satellite imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops, pp. 10–17.

He, J., Jia, X., Chen, S., Liu, J., 2021. Multi-source domain adaptation with collaborative learning for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11008–11017.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 328–341.

Hoeser, T., Kuenzer, C., 2022. SyntEO: Synthetic dataset generation for earth observation and deep learning–demonstrated for offshore wind farm detection. ISPRS Journal of Photogrammetry and Remote Sensing 189, 163–184.

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 184, 96–115.

Janoutová, R., Homolová, L., Malenovskỳ, Z., Hanuš, J., Lauret, N., Gastellu-Etchegorry, J.P., 2019. Influence of 3d spruce tree representation on accuracy of airborne and satellite forest reflectance simulated in dart. Forests 10, 292.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing 57, 574–586.

Kong, F., Huang, B., Bradbury, K., Malof, J., 2020. The Synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation, in: 2020 Winter Conference on Applications of Computer Vision, pp. 1814–1823.

Krauß, T., 2014. Six years operational processing of satellite data using CATENA at DLR: Experiences and recommendations. KN-Journal of Cartography and Geographic Information 64, 74–80.

Kurz, F., Türmer, S., Meynberg, O., Rosenbaum, D., Runge, H., Reinartz, P., Leitloff, J., 2012. Low-cost optical camera systems for real-time mapping applications. Photogrammetrie-Fernerkundung-Geoinformation , 159–176.

Li, H., Tian, J., Xie, Y., Li, C., Reinartz, P., 2022. Performance evaluation of fusion techniques for cross-domain building rooftop segmentation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences , 501–508.

Li, H., Wang, Z., Hong, T., 2021. A synthetic building operation dataset. Scientific data 8, 1–13.

Li, R., Jiao, Q., Cao, W., Wong, H.S., Wu, S., 2020. Model adaptation: Unsupervised domain adaptation without source data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9641–9650.

Li, X., Strahler, A.H., 1985. Geometric-optical modeling of a conifer forest canopy. IEEE Transactions on Geoscience and Remote Sensing GE-23, 705–721.

Li, X., Wang, K., Tian, Y., Yan, L., Deng, F., Wang, F.Y., 2019. The ParallelEye dataset: A large collection of virtual images for traffic vision research. IEEE Transactions on Intelligent Transportation Systems 20, 2072–2084.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision, Springer. pp. 740–755.

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, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Ma, J.W., Czerniawski, T., Leite, F., 2020. Semantic segmentation of point clouds of building interiors with deep learning: Augmenting training datasets with synthetic BIM-based point clouds. Automation in Construction 113, 103144.

Marsocci, V., Coletta, V., Ravanelli, R., Scardapane, S., Crespi, M., 2023. Inferring 3d change detection from bitemporal optical images. ISPRS Journal of Photogrammetry and Remote Sensing 196, 325–339.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3523–3542.

Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto (Canada).

Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. IEEE Transactions on Image Processing 16, 463–478.

Nikolenko, S.I., 2021. Synthetic data for deep learning. volume 174. Springer.

Peng, D., Guan, H., Zang, Y., Bruzzone, L., 2022. Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–17.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., Wu, Q.J., 2022. A review of generalized zero-shot learning methods. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Qi, J., Xie, D., Yin, T., Yan, G., Gastellu-Etchegorry, J.P., Li, L., Zhang, W., Mu, X., Norford, L.K., 2019. LESS: LargE-Scale remote sensing data and image simulation framework over heterogeneous 3D scenes. Remote Sensing of Environment 221, 695–706.

Qin, R., Tian, J., Reinartz, P., 2016. 3D change detection–approaches and applications. ISPRS Journal of Photogrammetry and Remote Sensing 122, 41–56.

Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016. Playing for data: Ground truth from computer games, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer. pp. 102–118.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M., 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10912–10922.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M., 2016. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3234–3243.

Schwind, P., Müller, R., Palubinskas, G., Storch, T., 2012. An in-depth simulation of EnMAP acquisition geometry. ISPRS Journal of Photogrammetry and Remote Sensing 70, 99–106.

Shah, S., Dey, D., Lovett, C., Kapoor, A., 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles, in: Field and Service Robotics: Results of the 11th International Conference, Springer. pp. 621–635.

Shao, R., Du, C., Chen, H., Li, J., 2021. SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network. Remote Sensing 13, 3750.

Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z., 2020. Change detection based on artificial intelligence: State-of-the-art and challenges. Remote Sensing 12, 1688.

Stoecklein, D., Lore, K.G., Davies, M., Sarkar, S., Ganapathysubramanian, B., 2017. Deep learning for flow sculpting: Insights into efficient learning using scientific simulation data. Scientific reports 7, 1–11.

Tao, J., Auer, S., Palubinskas, G., Reinartz, P., Bamler, R., 2013. Automatic SAR simulation technique for object identification in complex urban scenarios. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 994–1003.

Tian, J., Cui, S., Reinartz, P., 2013. Building change detection based on satellite stereo imagery and digital surface models. IEEE Transactions on Geoscience and Remote Sensing 52, 406–417.

Tian, J., Dezert, J., 2019. Fusion of multispectral imagery and DSMs for building change detection using belief functions and reliabilities. International Journal of Image and Data Fusion 10, 1–27.

Townshend, J.R., Justice, C.O., Gurney, C., McManus, J., 1992. The impact of misregistration on change detection. IEEE Transactions on Geoscience and Remote Sensing 30, 1054–1060.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE geoscience and remote sensing magazine 4, 41–57.

Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S., 2022. Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2795–2803.

Xie, Y., Tian, J., Zhu, X.X., 2020. Linking points with labels in 3D: A review of point cloud semantic segmentation. IEEE Geoscience and Remote Sensing Magazine 8, 38–59.

Xie, Y., Tian, J., Zhu, X.X., 2023. A co-learning method to utilize optical images and photogrammetric point clouds for building extraction. International Journal of Applied Earth Observation and Geoinformation 116, 103165.

Zabih, R., Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence, in: European Conference on Computer Vision, Springer. pp. 151–158.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. National science review 5, 44–53.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5, 8–36.