BRIGHTEARTH: PIPELINE FOR ON-THE-FLY 3D RECONSTRUCTION OF URBAN AND RURAL SCENES FROM ONE SATELLITE IMAGE

S. Tripodi[a,*], N. Girard[a], G. Fonteix[a], L. Duan[a], W. Mapurisa[b], M. Leras[a], F. Trastour[a], Y. Tarabalka[a], L. Laurore[a]

[a] LuxCarta Technology, Mouans Sartoux, France [b] LuxCarta South Africa, Cape Town, South Africa

KEY WORDS: Deep learning, optical satellite images, semantic segmentation, 3D reconstruction, digital terrain model.

ABSTRACT:

With the growth of the availability and quality of satellite images, automatic 3D reconstruction from optical satellite images remains a popular research topic. Numerous applications, such as telecommunications and defence, directly benefit from the use of 3D models of both urban and rural scenes. While most of the state-of-the-art methods use stereo pairs for 3D reconstruction, such pairs are not immediately available anywhere in the world. In this paper, we propose an automatic pipeline for very-large-scale 3D reconstruction of urban and rural scenes from one high-resolution satellite image. Convolutional neural networks are trained to extract key semantic information. The extracted information is then converted into GIS vector format, and enriched by both terrain and object height information. The final classification step is applied, yielding a 16-class 3D map. The presented pipeline is operational and available for commercial purposes under the BrightEarth trademark.

1. INTRODUCTION

With the evolution of satellites sensors, both spatial and temporal resolutions of remote sensing optical images have been greatly improved over the last decades. Tremendous set of data for Earth observation are now available allowing global land-cover classification and 3D reconstruction on high-resolution images. Three dimensional models represent a fundamental information for several applications including telecommunications, defence, urban planning, building mapping and monitoring. Traditionally, the acquisition of two images of the same location taken from different angles enables to produce a 3D elevation model. For this reason, stereo imagery is generally used to generate urban models (Pepe et al., 2021). Nevertheless, it remains arduous to obtain instantaneous stereo pairs at any point on the Earth.

To tackle this image availability issue and give the opportunity to create three-dimensional models anywhere on the globe at a reasonable cost, the main goal of this work is to obtain an equivalent result from a single cloud-free image. Nowadays, the availability of full global base maps together with the associated metadata such as sun elevation and azimuth angles makes possible height estimation and thus automatic digital 3D mapping of the Earth's surface.

To produce this on-the-fly 3D reconstruction in an automatized way, we set out to establish a methodology focusing on solving three major issues:

- A high-quality classification map must be extracted from the satellite image and properly converted into a GIS vector format. The accuracy obtained will have a direct impact on the quality of the 3D scene reconstitution.
- Heights of buildings and trees must be accurately estimated from a single image.
- From a mono-image only a height above the ground of each object can be estimated, and not the absolute base elevation. In order to place the extracted objects in a 3D cartographic coordinate system, we propose to reconstruct a digital terrain model (DTM). DTM is commonly produced by image correlation when a stereo pair is available. The goal here is to extract the terrain from an available world digital surface model (DSM).

To address the above issues, we propose an automated pipeline for city and rural modelling from a single orthorectified image as input. The major contributions of this work lie in predicting shadows and facades directly during the deep-learning segmentation in order to be able in the following steps to compute the terrain object heights. In addition, the DTM generation from DSM according to the resolution of DSM allows us to validate heights and thus increase result confidence and completeness. Finally, a refined classification (needed by the industry) yields a 16-class 3D map including :

- **9 building classes** : dense urban, urban, high density residential, medium density residential, sparse/village (low density residential), multi-family residential, industrial and commercial, building blocks and high building.
- **3 vegetation classes** : dense trees, sparse trees and low vegetation.
- **2 water classes** : sea and inland water.
- **2 ground class** : road and bare ground.

2. PROPOSED PIPELINE

Our input is one orthorectified image with corresponding metadata (e.g. azimuth and elevation angles for the sun and the satellite), as well as the global digital surface map.

The proposed pipeline (see Fig. 1) consists of the following steps:

* Corresponding author stripodi@luxcarta.com

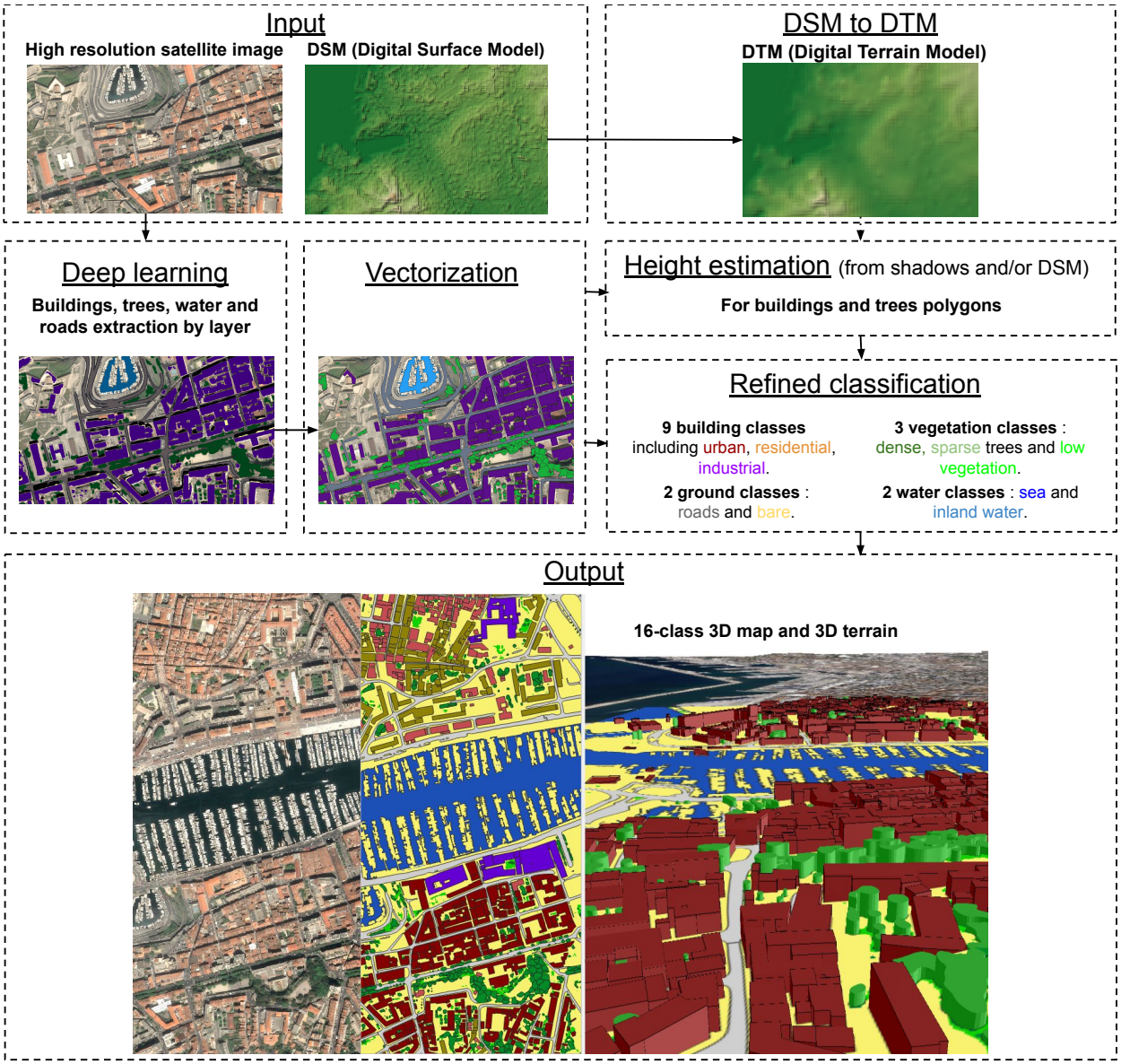

Figure 1. The proposed workflow for 3D reconstruction of urban and rural scene from one satellite image. The DSM can be used in option for height estimation, in addition to height-from-shadows algorithm (which only requires the input orthoimage); it can also be used for the estimation of the terrain height and thus the absolute height position of the objects.

1. Deep learning-assisted extraction of the following classes:
   - Buildings: with sub-classes rooftop, contour/wall (including inner-walls to separate adjacent buildings), facade and shadow.
   - Trees: with sub-classes tree and shadow.
   - Water.
   - Roads: with sub-classes road, road contour.
2. Vectorization of extracted objects.
3. DSM to DTM extraction.
4. Height estimation for buildings and trees.
5. Building and tree type classification.

Each step of our pipeline is described in the following sections and is validated in Section 8.

3. DEEP LEARNING-ASSISTED EXTRACTION OF BUILDINGS, TREES, WATER, AND ROAD

In our pipeline, we extract semantics such as buildings, trees, water, and roads separately instead of predicting all classes in one neural network as in (Tasar et al., 2019, Zhang et al., 2018). Our goal is to reconstruct urban scenes with all objects positioned in 3D through 2D imagery. Buildings and roads are vulnerable to occlusions (e.g. trees) in terms of preserving their geometrical regularities, predicting buildings/roads by separate models gives more flexibility to enforce the completeness and regularity of man-made shapes in our pipeline. We explain the extraction of each semantic in the following subsections.

3.1 Buildings

For building extraction, we train a deep learning model based on U-Net (Ronneberger et al., 2015) with a ResNet-101 (He et al., 2015) encoder to output a pixel-wise classification of the

input image. Predicted classes are: rooftop, contour/wall (used to separate adjacent buildings), facade, and shadow of buildings on the ground.

3.1.1 Data preparation Our dataset spans orthoimages across the world. For each image we have manually-curated ground truth polygons for individual building rooftops. Each polygon has a ground truth height attribute. The image metadata include azimuth and elevation angles for the sun and the satellite. From the rooftop polygons we rasterize rooftop and contour/wall ground truth as binary masks. Rooftop polygons are projected on the footprint using the satellite angles and the height attribute. It uses a simple projection that assumes a flat ground in the proximity of the building. From this shoebox projection we build ground truth facade polygons which are then rasterized to the corresponding ground truth as binary mask. In much the same manner we compute shadow polygons by projecting the footprint polygons, this time using the sun angles. To obtain shadow polygons corresponding to ground shadows, a difference operation is performed between shadow polygons and facade polygons. After rasterisation we obtain ground shadow ground truth binary mask. See Fig. 2 for an example of all the ground truth masks.

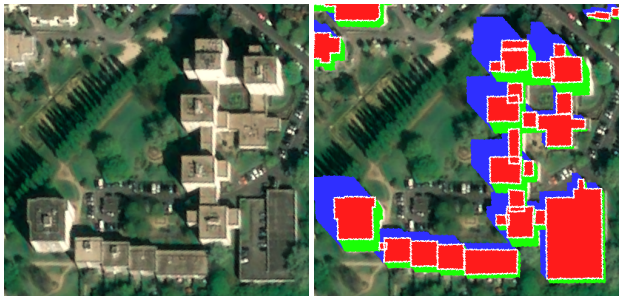

Figure 2. Example ground truth masks. Red: rooftops, green: facades, blue: shadows, white: contours/walls.

3.1.2 Model training Our fully-convolutional neural network is trained with a combined multi-class loss, adding cross-entropy and IoU (Intersection over Union) loss together. The contour/walls class is given a higher weight for balancing.

3.2 Trees

For trees, we train the same model as for buildings but with a different output head to predict two classes: tree and ground shadow of trees. Our dataset is a collection of various satellite images across the globe for which we have vegetation ground truth polygons. For the ground truth ground shadow of trees however, we had to resort to manual annotation. As this is resource-intensive, we only labeled a fraction of our dataset. Training is adapted to apply the loss on the shadow class only when the ground truth for it is available. Images for which the shadow ground truth is available are sampled more often during training. We are thus able to train for tree segmentation on the whole dataset (to increase generalisation), while also being able to train for tree shadow on the small shadow-labeled portion of the dataset, all with the same model.

3.3 Water

Concerning water extraction, a slightly different approach was chosen. The main idea consists in combining deep learning-based classification results with an additional external data source. The semantic segmentation of water bodies on high-resolution images is a complex task for several reasons. Indeed, the water areas are not textured, the colors depend on many factors such as sun exposure and can be confused with shadows, shapes are irregular and other objects can hide part of the water areas.

To get a convincing result, the proposed pipeline consists of two main steps. The first consists in classify every pixel of the image based on deep learning. The second marries this deep learning-based classification with an external data source for automatic correction.

3.3.1 Deep learning-based classification As for buildings and trees, we train a deep learning model based on U-Net to predict in this case only one class. This resulting model is applied to satellite images to detect all kinds of water bodies : inland waters (rivers, canals, lakes, and ponds) and maritime waters (seas and oceans). Three-band-images (near-infrared, red and green) are used as input for the model. Indeed, according to the most popular water index, the NDWI (Normalized Difference Water Index) (McFeeters, 1996), it is recommended to use near-infrared and green bands for extracting water bodies. These bands allow better water discrimination when compared to natural colors. Moreover, the amount of accurately annotated ground-truth data resulting from archiving products delivered by LuxCarta[1] over the years has yielded an efficient model that succeed in generalizing all over the world.

3.3.2 Automatic correction Once the deep learning-inferred water segmentation is obtained, the main idea is to use an additional external data source to enhance the result. To solve this problem the method described in (Fonteix et al., 2021) seems promising where deep-learning based segmentation is combined with an another available data source (e.g., outdated prediction, OpenStreetMap, etc.) to perform an automatic correction of natural environments and fit the high-temporal resolution satellites images of Sentinel-2A. We have applied this approach on high-resolution satellite images, and have found that this method is well adapted for our use case as shown in Fig. 3.

3.4 Roads

Following the method for building prediction, we train a U-Net (Ronneberger et al., 2015) model to predict 3 classes: road, road contour, other. Training set is composed of the satellite images with 50 cm/pixel spatial resolution, acquired by different sensors (Pleiades, WorldView-3 etc.). Ground truth data are polygons drawn by experts in GIS data labeling, which provide a high quality and good completeness of road coverage of cities worldwide. Contrary to building rooftops, roads form a contiguous and connected network while trees and cars often appear on roads, sometimes even completely covering road areas. In our ground truth creation policy, a road underneath an obstacle such as a tree or a car is labeled as road. From our experiments, our model is capable of handling a certain degree of object occlusion and extract complete road bodies. Fig. 4 shows an example of our road prediction quality.

4. VECTORIZATION OF EXTRACTED OBJECTS

Building polygonization. A computational geometry approach is applied to vectorize the raster prediction of building rooftops

[1] https://www.luxcarta.com/

Figure 3. Example of a water segmentation before and after the automatic correction. Pléaides image over Marseille, France. Water prediction, water prediction after combining with an external source.

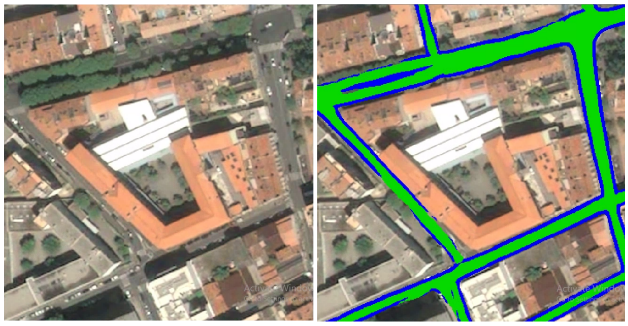

Figure 4. Left: input orthoimage, right: our road prediction with road, contour, other (transparent).

to polygons. By polishing the contour/rooftop labels and border positions, we polygonize each buildings into compact and regularized polygons. A joint optimization is designed to optimize simultaneously the regularity of each individual polygon shape and the snapping with adjacent/neighboring polygons in the polygonization framework. An example of building polygonization result is show in Fig. 5, in red.

Tree polygonization. We apply a superpixel segmentation over the tree areas in our prediction to obtain relatively homogeneous tree canopy clusters and polygonize them into preliminary polygons. Several postprocessing operations including shape optimization, snapping are applied to improve the visualization of our final tree polygons. As shown in Fig. 5, in green, our tree polygons fit the tree contours in the input image with homogeneous size, without any geometry conflicts such as overlapping, tiny cracks, self-intersection, etc.

Water body polygonization. Our prediction of water body is represented as a binary raster mask: water or other. We polygonize valid water areas into simple polygons and polish boundaries using the polynomial approximation method (Bodansky et al., 2002) to obtain more natural water body contours, as shown in Fig. 5, in blue.

Road polygonization. Since there exist uncountable noises on roads, such as trees, cars, occlusions from buildings and so on, we apply a preprocessing operation to polish lightly our road prediction by ignoring tiny holes inside road body areas, completing short missing/open contours. Then we polygonize the road prediction with contour+road as foreground. Road polygon shapes are refined by the method of (Bodansky et al., 2002), and long road polygons are sliced at junctions to avoid self-intersection issues when massive road bodies connect together. An example is shown in Fig. 5, in violet.

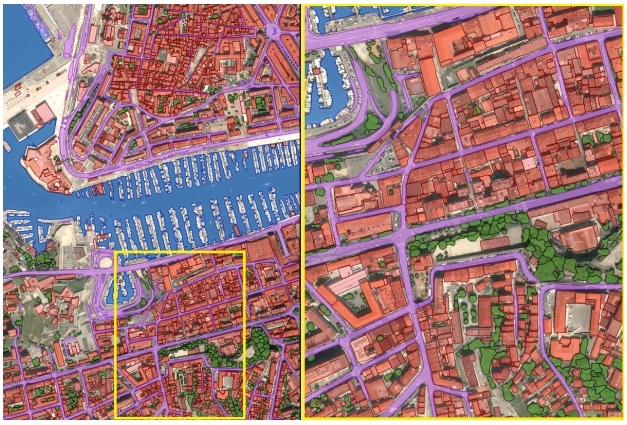

Figure 5. Vector layers of building, tree, water, and road polygons on top of the orthoimage (left) with a crop (right). Red: rooftops, green: trees, blue: water, violet: roads.

5. DSM TO DTM EXTRACTION

5.1 Overview of problem

For a complete 3D scene reconstruction, a DTM (Digital terrain model) is necessary to put 3D objects such as buildings and trees on the ground. The first constraint of the proposed pipeline is that we have only one image as input, so the DTM cannot be extracted by image correlation. The second constraint is our reconstruction has to work anywhere in the world. Today, world DEMs (Digital Elevation Model) are available for free at 30 meter resolution which represents the surface of the earth DSM (Digital surface model), including objects on the ground, such as AW3D30 (Japan Aerospace Exploration Agency, ALOS World 3D 30m, 2021), SRTM (NASA Shuttle Radar Topography Mission Global, 2013). However, a DSM does not represent only the terrain (see Fig. 6), it has to be properly cleaned to be used in order to place our reconstructed 3D buildings/trees. We propose in this section an algorithm to extract the DTM from DSM. This section illustrates our approach with the DSM AW3D30 because it is the best compromise between coverage (the whole world), resolution (30m) and accuracy. However our pipeline is not restrained to this data, a LIDAR DSM could also be injected in our pipeline for added accuracy.

The DTM extraction from DSM is a well-known problem (Mousa et al., 2017) but existing solutions are not accurate enough. Several methods identify objects on the earth as trees and buildings, then reconstruct the terrain under the objects. One of the most promising approaches in the literature (Duan et al., 2019) performs object detection by analysing the morphological profile of the DSM, and then applies computer graphics state-of-the-art interpolation based on SAKE (Budninskiy et al., 2017). In our case, as the input is a single orthoimage, object

detection is done using deep learning (see section 3). As for interpolation, techniques from previous works fail to interpolate when applied on large areas, for example in very dense urban areas or large forests.

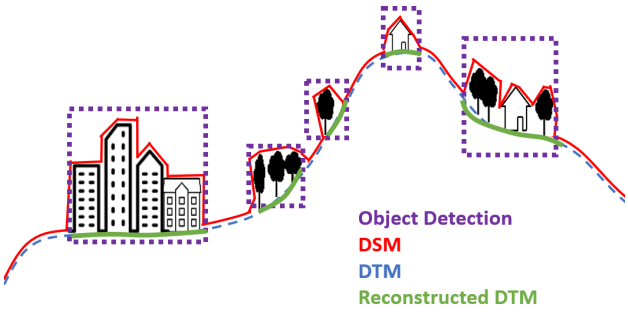

Figure 6. DSM vs DTM.

5.2 Our Approach

We propose an original and efficient approach based on a physical simulation under constraints and a GPU implementation to solve this problem.

5.2.1 DSM Analysis The physical simulation can model the deformation of the terrain in a nonlinear way and respect some constraints. These constrains can be easily integrated in the simulation being an iterative process. Beginning with the elevation for each point given by the DSM, at each iteration a new value of elevation is simulated until a stable state is reached corresponding to the value of the DTM. Constraints are evaluated at each step, $e.g.$ to force the new value to be included in an interval or remain still if the DSM=DTM (attach point), see Fig. 7. These constraints allow for example to keep the relief under large forests and does not remove much of the terrain, compared to the traditional approach of interpolation. To get these constraints, a deep analysis of the DSM is done based on slope, accumulation flow, valley, and hill detection to:

- remove noise inherent to DSM and get the attach points,
- estimate the height of objects to compute an interval containing the value of the DTM. This estimation can be approximate because it serves just to guide the physical simulation which is already robust to deform the terrain in a realistic and fast way.

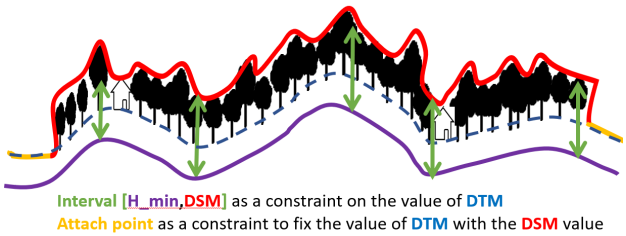

Figure 7. Constraints on the DTM values to inject in the physical simulation.

5.2.2 Physical simulation The physical simulation is based on a soft body simulation and particle system. One particle corresponds to one value of a DSM. To optimise computation time, a DSM is meshed so that the number of particles (number of vertices to move) is reduced, as is done in recent work (Zheng et al., 2016). The idea is to reduce the number of particles under flat areas and keep details on the relief. Our simulation includes two kinds of force fields: one for the tensile stiffness of terrain and one for the bending of terrain. Our simulation is implemented on GPU for enabling an efficient computation. Fig. 8 illustrates results of a DTM estimation, using a DSM AW3D30 at the input. The proposed technique can also be used on the LIDAR DSM. Finally, according to the resolution and the vintage of the input DSM we can extract the height of objects from DHM (DSM-DTM) with a certain accuracy to validate, fill or help the shadow height estimation (see section 6), in particular for large forests or very dense urban areas.

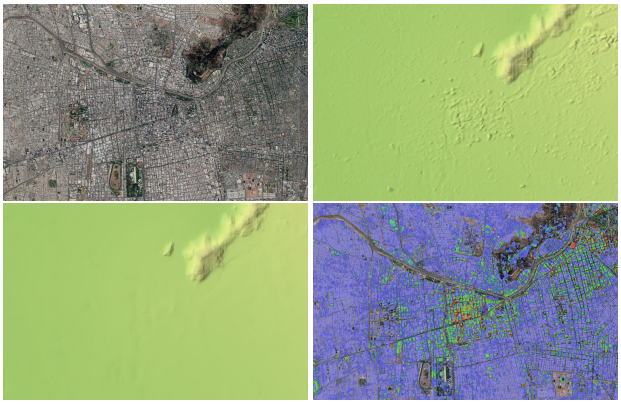

Figure 8. Example of the DTM estimation: AW3D30 DSM (top right), DTM (bottom left) and DHM (bottom right).

6. HEIGHT ESTIMATION FROM SHADOW

Height estimation of a building or a tree using their shadow uses the geometrical link between shadow length and height given a known sun elevation angle as described in eq. 1 (assuming a flat ground under the shadow).

$$h = l * \tan(\alpha)\,, \tag{1}$$

where h = height,
l = shadow length,
α = sun elevation.

The difficult part consists in computing the correct shadow length, given the predicted shadow mask, extracted building polygons, shadow occlusions (e.g. by its own building), etc.

The basic principle is to shoot rays from the contour of the extracted polygon in the direction of the shadow (which is known from the sun azimuth angle) and stop the ray when it goes out of the predicted shadow mask, see Fig. 9 for an illustration. Each ray has a length in pixels, which is converted to meters with the GSD (Ground Sample Distance) of the image. Ray lengths are aggregated by computing the median, which is less perturbed by outliers (see for example ray 0 in Fig. 9, which would affect the average greatly). Assuming the image is from nadir, the building height is computed with eq. 1. When the image is not from nadir, the measured shadow lengths with the rays are under-estimated when the building occludes parts of its own shadow (as is the case in Fig. 9). However, the missing shadow length corresponds to the displacement of the footprint relative to the rooftop, which can be computed given the height and the satellite angles.

Figure 9. Left: input image, right: predicted masks for rooftop, facades, shadow, and shadow rays to compute shadow length.

7. REFINED CLASSIFICATION

Many real-life applications, such as land-use and radio network planning, require classification of buildings and trees into subcategories. For buildings, this is usually an identification of classes such as urban areas, residential, isolated and industrial buildings. The algorithm developed in this paper uses polygon density and building heights to classify building polygons into 9 building classes.

The first step is to assign for each polygon a "density" class. Initially the polygons are grouped into structures, where each structure represents a building or a block of adjacent buildings with the common edges. A dense triangulation is created, where each vertex represents a building structure. The triangulation is then converted into a graph with each vertex representing a building and edges representing the distance to neighbouring structures. Since the graph can be dense with millions of polygons, we reduce the graph to a minimum spanning tree which simplifies subsequent analysis. Finally, using the minimum spanning tree clusters are identified with graph segmentation based on proximity and analysing spanning tree branches. Polygons are clustered based on initial identified structures using proximity into three classes namely high, medium and low density areas.

After the identification of density classes, we couple the information about:

- building height,
- density class,
- building size and shape

to classify buildings into 9 classes: dense urban, urban, high density residential, medium density residential, sparse/village (low density residential), multi-family residential, industrial and commercial, building blocks and high buildings. Fig.10 shows a sample classification over Marseille.

The same density classification as for buildings can also be applied to any polygon types including trees where a density measure is required. Like this trees are separated into dense and sparse trees. Moreover, a low vegetation class computed with a vegetation index is added to the final map as well as inland water and sea classes distinguished from the water extraction thanks to a global seas and oceans mask.

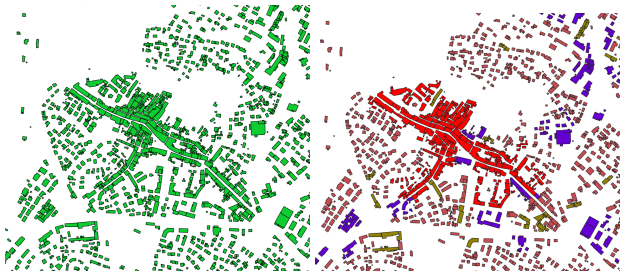

Figure 10. Left: building polygons before classification, right: classified Buildings. Red represents urban area, brown represents high density residential and purple represents industrial and commercial buildings

8. EXPERIMENTS AND CONCLUSIONS

8.1 Experiments and validation

Throughout the paper, the automatic pipeline has been illustrated over Marseille, France (see the Fig. 1). Our pipeline is also operational and available for commercial purposes under the BrightEarth trademark as an online service and an execution on-the-fly allowing to experiment any use case on the world by just drawing an area of interest see the Fig. 11.

The following subsections give some details of validation of our technology for the 2D segmentation, height estimation and DTM extraction.

8.1.1 2D segmentation We have tested our pipeline on more than twenty cities and rural areas on which metrics have been carefully calculated as well as comparison with other state-of-the-art automatic methods.

Regarding the evaluation of the semantic segmentation, the straightforward common metric Intersection over Union (IoU) also known as Jaccard index (Rahman and Wang, 2016), was calculated for each class. It measures the number of pixels in common divided by the total number of pixels present in both predictions.

A very precise, manually labeled, ground truth dataset over twelve areas on four continents was used to evaluate results. We have reported an IoU of 0.83 on the final segmentation for trees, 0.89 for water and 0.85 for buildings. Moreover, it can be pointed out that the percentage of missing polygons for buildings is 1.79%. To get this percentage we consider as non missing polygon a building detected within a radius of 3 meters from the ground truth.

Our building extraction model improves upon (Tripodi et al., 2019), in part because it was trained on around one hundred big satellite images from diverse locations, but mostly because of the addition of the facade class which allows to remove confusion between facade and rooftop (see Fig. 12), which methods classifying only rooftops are prone to. Additionally, detecting shadows with deep learning compared to more traditional methods allows to compute clean shadows without confusion between shadows of different classes (e.g. tree shadows are correctly classified as background by the building model and vice versa, an example of this in Fig. 9). The shadow↔water confusion is also solved with the semantic approach of deep learning compared to more traditional approaches.

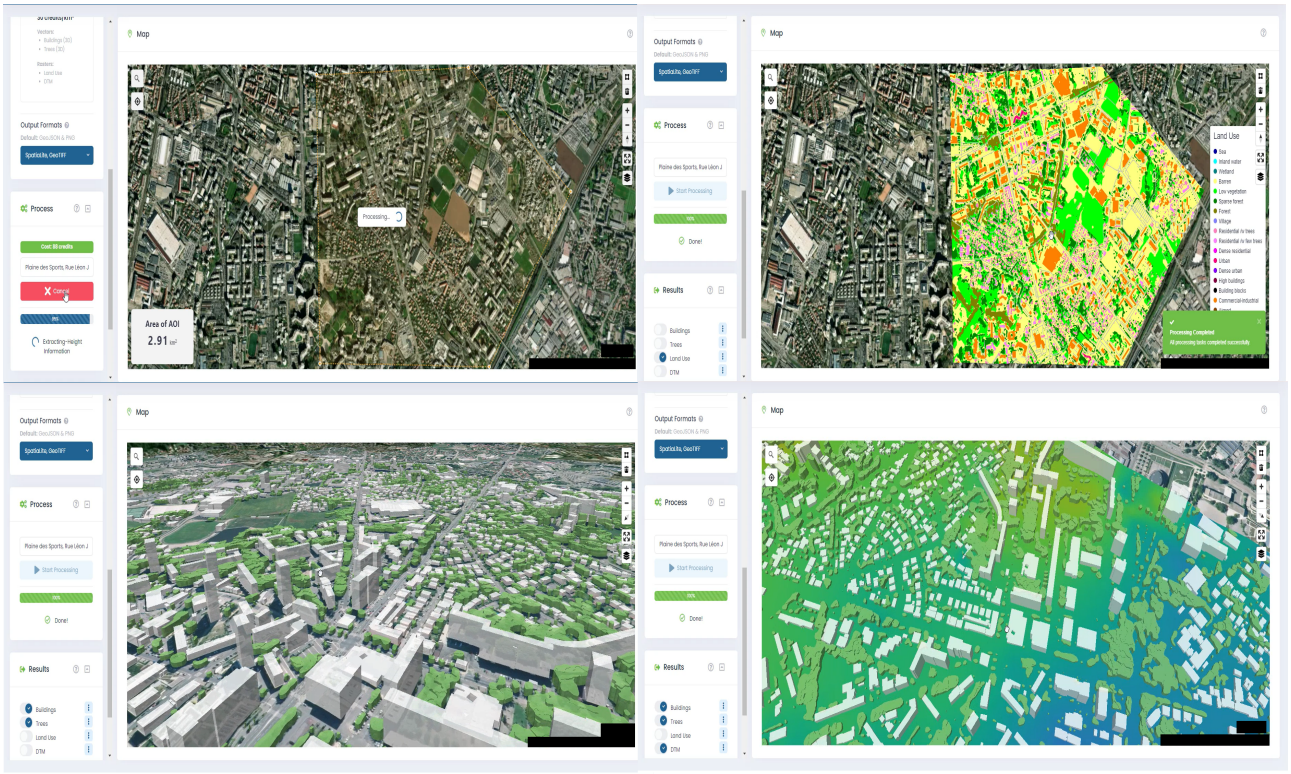

Figure 11. BrightEarth: online and on-the-fly service of our operational pipeline to reconstruct 2D land use (top right), urban scene in 3D (left bottom) with DSM (left right) anywhere in the world from an area of interest (bottom right), Grenoble France Area

Figure 12. Left: building extraction from (Tripodi et al., 2019), right: our building extraction. Both use our polygonisation technique.

8.1.2 Height estimation

We tested our height estimation algorithm on a few satellite images unseen by our deep learning model. See Table 1 for the results. As large images should be processed in a reasonable amount of time, we optimized the algorithm to process at least 700 polygons per second (see timings per image in Table 1). In the majority of cases, we achieve an average height error around the meter. Notice that the accuracy is proportional to the sun elevation angle α: the lower the angle, the more precise height computed from shadow length is. This is a direct result of the $\tan(\alpha)$ term in eq. 1.

8.1.3 DTM extraction

Our method of DTM extraction from a DSM works equally as well on 30 m DSMs or high resolution DSMs (e.g. coming from lidar). In this paper we have illustrated our method with the AW3D30 DSM as we propose a pipeline which has to be operational anywhere in the world. We validate our method with this data on all of New Zealand by using as reference 22000 GCPs (Ground Control Point). We obtain an RMSE of 3.49 m. According to the 30 m resolution of AW3D30 and the number of GCP, the RMSE validates our approach. An advantage of our approach is the possibility to easily add constraints such as taking into account the GCPs to force the physics simulation to reach these GCPs without creating artefacts. In this use case the RMSE is 0.8 m.

Image/area name	image size	# buildings	sun elevation α	Time ↓	avg. err. ↓
Leon, Mexico 1	36001 × 38384	76753	47°	1min30s	1.18m
Leon, Mexico 2	24056 × 25196	156059	51°	3min	1.09m
San Luis Potosi, Mexico	15132 × 17476	4485	43°	30s	1.45m
Île-de-France 1	44276 × 114477	300107	55°	4min	1.58m
Île-de-France 2	42445 × 147147	486010	61°	8min	2.18m
Île-de-France 3	3960 × 39727	40668	55°	40s	1.68m
Tijuana, Mexico 1	25312 × 22520	156891	36°	3min	0.84m
Tijuana, Mexico 2	24000 × 23600	93419	39°	1min40s	0.78m
Toulon, France	43364 × 32944	89261	42°	1min40s	1.89m

Table 1. Test results for the building height estimation step on satellite images with a GSD of 50cm. Times only include the height estimation itself (excluding building extraction) on an Intel i9-10850K CPU (single-thread).

8.2 Conclusion

We present an automatic 3D reconstruction pipeline for worldwide very-large-scale urban scenes from satellite mono-image. In particular, our pipeline generates a complete 3D scene: land use, water vector, road vector, 3D buildings, 3D trees and finally the 3D Digital Terrain Model extracted from a DSM. Our 3D scenes are refined in terms of geometry accuracy and regularity, and enriched by various kind of building types as well as vegetation. The quality of our 3D reconstruction suits many valuable Remote Sensing applications such as urban planning for telecommunication and the simulation market. Moreover, based on our massive experiments and production experiences, our pipeline is robust to almost all type of geographical styles with very different urban appearance and structures. In the whole pipeline, particularly the DeepLearning segmentation module, we evaluated each key step during our experiments, which proves the generality and robustness of our method. As future work, we will extend the pipeline to textured 3D recon-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2022-263-2022 | © Author(s) 2022. CC BY 4.0 License.

struction with rooftop type information in order to reconstruct 3D urban and rural scenes with richer and more accurate geometry details in a realistic way.

REFERENCES

Bodansky, E., Gribov, A., Pilouk, M., 2002. Smoothing and compression of lines obtained by raster-to-vector conversion. D. Blostein, Y.-B. Kwon (eds), *Graphics Recognition Algorithms and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 256–265.

Budninskiy, M., Liu, B., Tong, Y., Desbrun, M., 2017. Spectral Affine-Kernel Embeddings. *Comput. Graph. Forum*, 36(5), 117–129. https://doi.org/10.1111/cgf.13250.

Duan, L., Desbrun, M., Giraud, A., Trastour, F., Laurore, L., 2019. Large-scale dtm generation from satellite data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Fonteix, G., Swaine, M., Leras, M., Tarabalka, Y., Tripodi, S., Trastour, F., Giraud, A., Laurore, L., Hyland, J., 2021. Marrying Deep Learning and Data Fusion for Accurate Semantic Labeling of SENTINEL-2 Images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 101–107.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

Japan Aerospace Exploration Agency, ALOS World 3D 30m, 2021. Distributed by OpenTopography. `https://doi.org/10.5069/G94M92HB`. Accessed: 2022-01-06.

McFeeters, S. K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International journal of remote sensing*, 17(7), 1425–1432.

Mousa, A.-k., Helmholz, P., Belton, D. et al., 2017. New DTM extraction approach form airbone images derived DSM. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42.

NASA Shuttle Radar Topography Mission Global, 2013. Distributed by OpenTopography `https://doi.org/10.5069/G9445JDF`. Accessed: 2022-01-06.

Pepe, M., Costantino, D., Alfio, V. S., Vozza, G., Cartellino, E., 2021. A Novel Method Based on Deep Learning, GIS and Geomatics Software for Building a 3D City Model from VHR Satellite Stereo Imagery. *ISPRS International Journal of Geo-Information*, 10(10), 697.

Rahman, M. A., Wang, Y., 2016. Optimizing intersection-over-union in deep neural networks for image segmentation. *International symposium on visual computing*, Springer, 234–244.

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

Tasar, O., Tarabalka, Y., Alliez, P., 2019. Incremental Learning for Semantic Segmentation of Large-Scale Remote Sensing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(9), 3524-3537.

Tripodi, S., Duan, L., Trastour, F., Poujad, V., Laurore, L., Tarabalka, Y., 2019. Automated Chain For Large-Scale 3d Reconstruction Of Urban Scenes From Satellite Images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., Zhang, S., 2018. Urban land use and land cover classification using novel deep learning models based on high spatial resolution satellite imagery. *Sensors*, 18(11), 3717.

Zheng, X., Hanjiang, X., Gong, J., Yue, L., 2016. A virtual globe-based multi-resolution TIN surface modeling and visualizetion method. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2, 459-464.