# Reconstruction of 3D Environments from Satellite Images by AI and Computational Geometry for Exploitation in Mixed Reality

**Dr. Yuliya Tarabalka, Dr. Nicolas Girard, Dr. Sebastien Tripodi,**
**Cedric Larrosa, Dr. Jean-Philippe Bauchet and Guillemette Fonteix**
LuxCarta Technology
460 Avenue de la Quiera – Voie K – Bat 119B
06370 Mouans Sartoux
FRANCE

ytarabalka@luxcarta.com / ngirard@luxcarta.com / stripodi@luxcarta.com / clarrosa@luxcarta.com / jpbauchet@luxcarta.com / gfonteix@luxcarta.com

**Christian Zanca, Fabien Lavignotte and Gregory Smialek**
CS-GROUP
Les hauts de la Duranne,
370 rue René Descartes, 13290 Aix-en-Provence, FRANCE

christian.zanca@csgroup.eu / fabien.lavignotte@csgroup.eu / gregory.smialek@csgroup.eu

**Vincent Madelain**
LuxCarta International
245 route des Lucioles – Bat A, 06560 Valbonne, FRANCE

vmadelain@luxcarta.com

## *ABSTRACT*

*The multiplication and increasing availability of high-resolution satellite imagery sources allows, thanks to artificial intelligence and computational geometry, the increasingly rapid reconstruction of faithful 3D cartographic environments adapted to the needs of simulation for troop training devices, in particular for mixed reality visualization. We have developed an operational automatic pipeline, which enables automatic generation of digital terrain models and orthoimages from multi-stereo satellite images. The generation of additional 3D vector assets needed for the geometric description of masks (buildings, trees) are also available from multi-stereo imagery but as well from a simple ortho-image. Furthermore, our pipeline allows for both recognition of roof shapes and automatic texturing of buildings using a hybrid approach which marries artificial intelligence and procedural modelling. Provision of optimized 3D tiles format (OGC standard promoted by CESIUM) generated in an automatic way enables massive dissemination of the generated information in various visualization engines. Finally, the exploitation in the context of mixed reality (Microsoft HoloLens 2) integrating the virtual objects in a real scene, allows the calculation of the occultations of the scene on-site. These advances constitute a breakthrough technology for the rapid and cost-effective generation of large-scale terrains, allowing the necessary precision for automatic scene generation in simulation (Unreal Engine 5).*

## 1.0 INTRODUCTION

The latest-generation satellite sensors acquire large volumes of very-high resolution images on a global scale and with permanent availability. Therefore, the satellites play an important role in both military and civilian intelligence and surveillance systems, notably for controlling the information in the phases of situation assessment, preparation and action. They contribute to the economy of resources by allowing a better

concentration of efforts to obtain the maximum military effectiveness. Thus, it is important and urgent to develop methodologies for automatic processing of satellite images, to enable a rapid identification of relevant information from big data of geospatial archive imagery and a correlation of different data for deriving richer insights.

One of strategic applications is the reconstruction of faithful 3D cartographic environments. The solutions allowing to reconstruct large earth surfaces within a short time lapse are crucial, both for the Earth monitoring, for the mission preparation and control, and for generating models of realistic geophysical environments which would be used to carry out definition studies and design future systems with the support of simulation.

This article presents an operational automatic pipeline, which enables reconstruction of 3D earth environments from stereo satellite images. These environments are composed of digital terrain models and orthoimages, enriched by 2D and 3D vector assets for both semantic and geometric description of scene objects, such as buildings, trees, roads and water surfaces. Furthermore, the proposed pipeline enables automatic LOD2 (level of details with detailed roof planes) reconstruction and texturing of buildings using a hybrid methodology, which marries artificial intelligence (AI) and procedural modelling.

The reconstructed 3D environments are further streamed in the optimized 3D tiles format, enabling massive dissemination of the generated information in various visualization engines (Cesium, Unreal Engine, etc.). Finally, we illustrate the exploitation of the generated data in the context of mixed reality, using the HoloLens demonstrator. Integrating the virtual objects in a real scene allows the calculation of the occultations of the scene on-site.

## 2.0 AI-ASSISTED GEODATA GENERATION FROM SATELLITE IMAGES

Reconstructing 3D environments in an accurate way consists in reconstructing 3D objects present at the surface of the Earth but also placing them at a world position in an accurate manner. It means being able to manage raw satellite imagery to allow both terrain reconstruction (civil satellites offer today 30-cm spatial resolution, allowing a fine level of details in the reconstructed scene) and aligning the 3D data with the reference (either image or ground control points), which can be very useful for an update. The LuxCarta's automatic chain takes advantage of the raw satellite images to reconstruct 3D models with low horizontal and vertical errors with respect to the selected reference. Some key performance indicators are given in this section to validate the pipeline.

### 2.1 Automatic chain

Fig. 1 describes the proposed chain for large-scale 3D reconstruction of earth scenes in LOD1/LOD2 (CityGML). The input is a set of raw high-resolution satellite images, with the associated RPC models provided by the vendors. Our pipeline can process images acquired by different satellites with different spatial resolutions, such as Worldview, Pleiades, GeoEye or Spot. In this paper, we validate the pipeline by processing images at a spatial resolution of 50 cm/pixel. This section describes automatic reconstruction of 3D models in LOD1, where objects are represented by a set of polygons with the associated height. Section 3 describes automatic LOD2 reconstruction and texturing of buildings. Our chain consists of five main parts (ref. Fig. 1):

- Extraction of the semantic information (Part 1),
- Adjustments of the Rational Polynomial Coefficients (RPC) models using a reference (Part 2),
- Extraction of the height/elevation information (Parts 3, 5),
- 2D/3D object reconstruction (Parts 4, 6),
- Generation of 3D environments (Part 7): this part is described in Section 3.

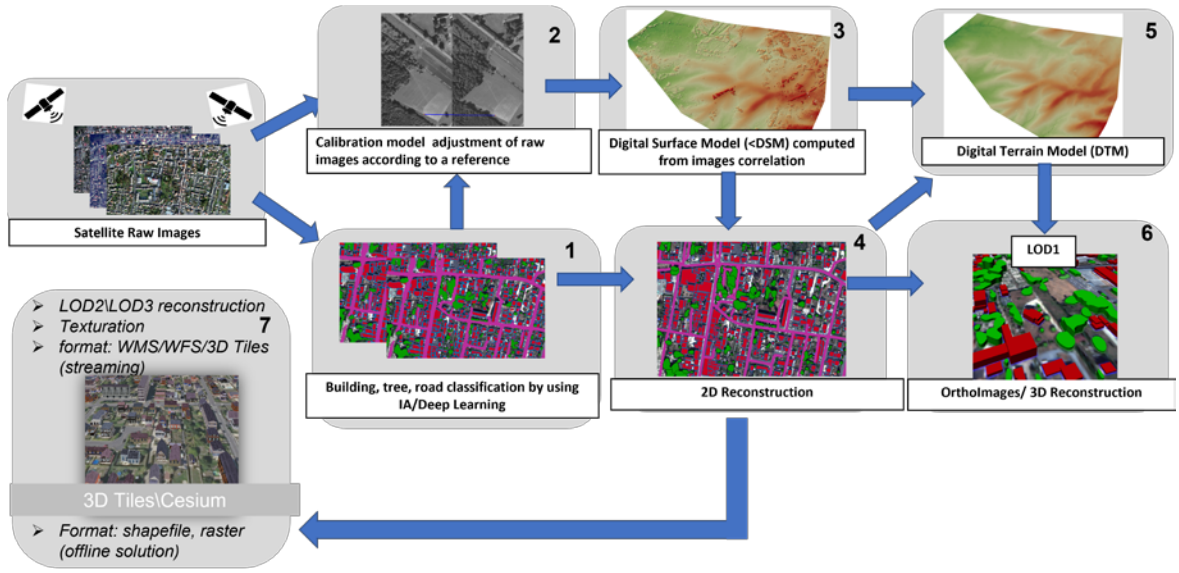


**Figure 1: Automatic chain for 3D reconstruction of Earth environments.**

### 2.1.1 Extraction of the semantic information

We extract semantics such as buildings, trees, water, and roads, using deep learning. To perform a pixel-wise classification for each class (See Fig. 4), a separate neural network model U-Net [1] with a ResNet-101 [2] encoder was trained using the data from LuxCarta archives. 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.
Our learning database consists of hundred cities around the world and manually digitalized or corrected for an accurate ground truth. More details about classification models can be found in [3].

### 2.1.2 Adjustments of RPC models

The satellite raw images (level 1B or 2A) are not rectified by the terrain; it means there is no relation between the world coordinates and pixel coordinates of the image. To solve this problem, vendors of satellite imagery provide in addition an RPC model (polynomial model) allowing this relation. However, even if the accuracy of these models has been continuously increasing, errors of an order of several meters are still regularly present. To adjust an RPC model, our method takes at the input a reference image or a set of ground control points (GCPs), and uses deep learning-assisted approach to find a common ground key points in the raw images and a reference; then ratio polynomial coefficients are adjusted accordingly, by optimizing the data alignment on the ground.

### 2.1.3 Extraction of the height/elevation information

If a pair of satellite images is available, it is possible to correlate them to get the elevation information for each pixel of the image. The resulting set of estimated elevations for every image pixels composes a Digital Surface Model (DSM, see Fig. 2). As shown in [4], two main approaches exist to compute a disparity map between two stereo images: based on semi-global matching (SGM) and deep learning. Due to the difficulty to build a ground-truth, we retained the method based on SGM. We modified the original SGM algorithm [5], for both boosting the performance by using the GPU and managing complex area as the occlusion zones

or shadows (details are explained in [6]).

An example of a DSM is shown in Fig. 5. To get the height information for each object (e.g. building, tree), one strategy consists in reconstructing a digital terrain model (DTM), which gives an elevation of each position at the ground over the sea level; then deriving heights by computing difference between DSM and DTM (see Fig. 2). The DTM is a valuable component of the final 3D model, since it allows to properly position each object in the simulated 3D world.

The DTM extraction from DSM is a well-known research problem [7]. We propose an original and efficient approach based on a physical simulation under constraints and a GPU implementation to solve this problem. The physical simulation can model the deformation of the terrain in a nonlinear way and respect some constraints. These constraints allow for example to keep the relief under the large forests and at the same time avoid removing the terrain, when compared to the traditional interpolation approaches. To get these constraints, a deep analysis of the DSM is done based on slope, accumulation flow, valley, hill detection, as well as AI-based object detection (see Fig. 2). An example of DTM is shown in Fig. 5.

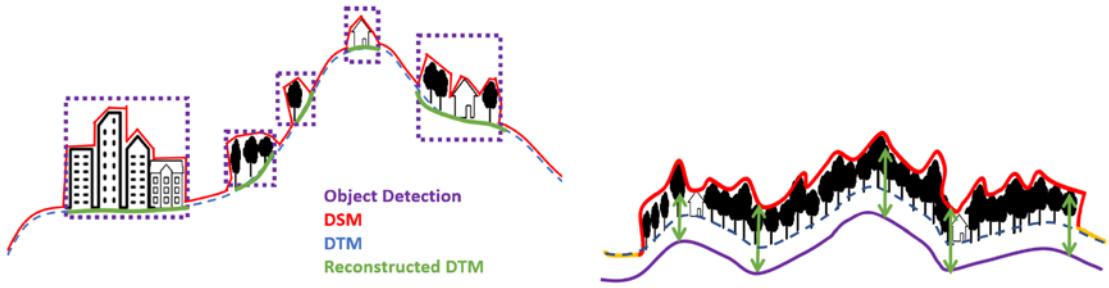


**Figure 2: DSM vs DTM (left) and constraints on the DTM (right).**

### 2.1.4 2D/3D object reconstruction

Object reconstruction in 2D and 3D (building, road, tree, water) consists in vectorizing pixelwise classification results in pixel, yielding vector objects into world coordinates, by considering appropriate regularity, simplification and smoothness constraints. Using the computed DTM and adjusted RPC models allows to establish a direct and accurate relation between pixel coordinated and world coordinates. We have developed optimization algorithms, which allow to reconstruct objects with the optimal ratio complexity/data fidelity.


**Figure 3: Problem of alignment between a reference and a raw image and automatic alignment.**

## 2.2 Validation

This section validates some key performance indicators of the LuxCarta's chain on the dataset acquired over the city of Mourmelon-le-Grand, France:

- RPC model adjustment. Without this step, the satellite image in Fig. 3 has several pixels shift with respect to the reference. Our automatic adjustment yields an average error < 0.5 m when compared to the reference.
- Fig. 4 (top-right) illustrates results of pixelwise classification. Table 1 shows key performances indicators on the validation dataset, where an accurate ground truth has been manually digitalized and not used in the learning.

**Table 1: Accuracy of pixelwise classification.**

| | Overall accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Buildings | **0.998** | **0.838** | **0.722** | **0.775** |
| Roads | **0.94** | **0.58** | **0.60** | **0.59** |
| Trees | **0.98** | **0.90** | **0.91** | **0.90** |

- Elevation and height extraction. Fig. 5 shows comparison between the reconstructed DSM and DTM and the RGE_ALTI (French large scale reference for altimetry). We obtain an average elevation error < 0.28 m.
- Reconstruction in 2D and LOD1. Fig. 4 shows results of vectorization and height assignment (DSM-DTM) for the 3D LOD1 reconstruction. We obtained an average vertical error < 0.82 m.

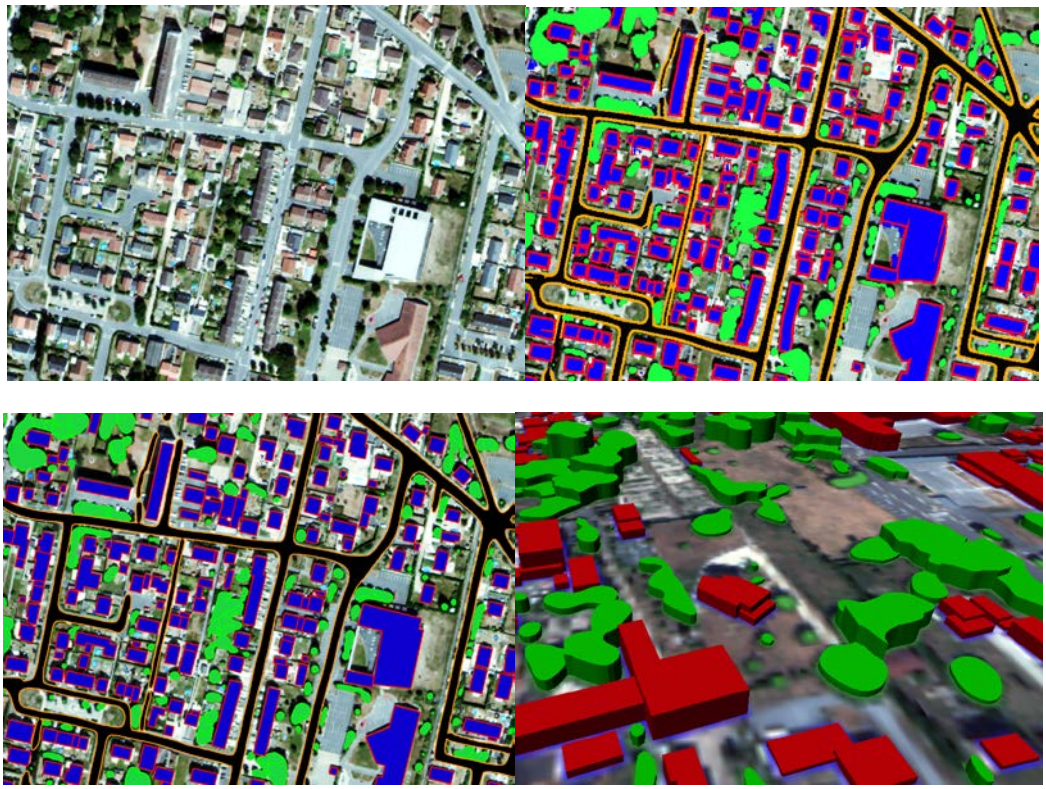

**Figure 4: Pixelwise classification (top), 2D and 3D reconstruction (bottom).**

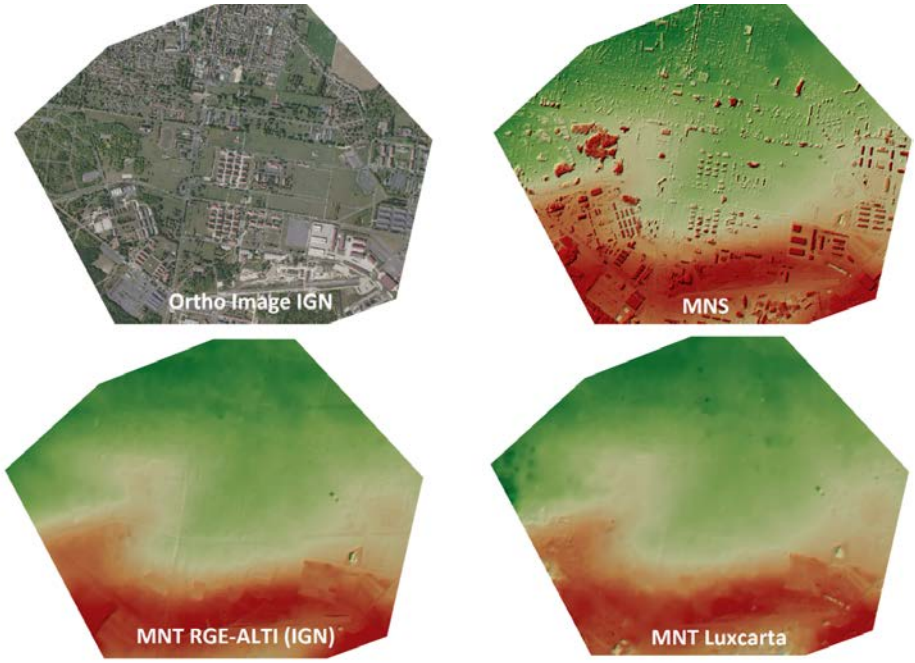


**Figure 5: DSM and DTM reconstruction.**

## 3.0 GENERATION OF 3D ENVIRONMENTS

We explain here part 7 of our automatic chain whose overview is below:

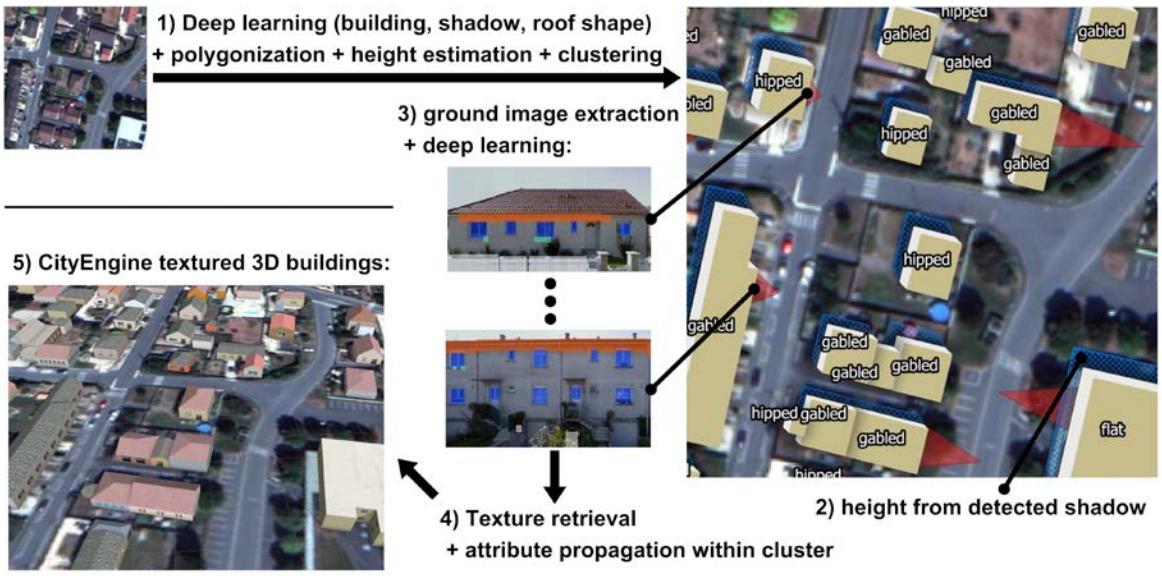


**Figure 6: Overview of part 7 of our chain.**

This part of the chain was developed as an independent fully automatic pipeline for the extraction of 3D textured buildings in LOD2 from a single orthoimage as input. However, it can also use the DTM and heights generated previously to provide better building height precision. It can also optionally use street level images for more accurate generation of building facades.

The principle of our approach is to procedurally generate the textured 3D buildings. This approach guarantees clean and uncluttered geometries. It also allows to use good resolution textures in an efficient way (one texture pack for the whole area rather than one texture per building). From a polygon delimiting the base of a building, we defined grammatical rules of generation to build a 3D model with windows, roof, etc. These rules accept some parameters as input, such as the height of the building, the textures to be used, the number of windows, floors, etc. When these parameters are generated randomly, we would get a probable 3D city but not faithful to that city in the real world. To bring fidelity to the generation of 3D buildings, we add constraints to the random procedural generation by fixing the parameters whose values we can extract from the orthoimage, from the DTM if available, and from the street level images if available.

The first step is to extract all the semantics from the input ortho image. Our neural network extracts building rooftops, ground shadows and roof shapes (flat, gabled, hipped, skillion, mansard, round, dome, silo, other). The second step consists in estimating the height of each polygon thanks to the detection of the building shadows (if heights are not available yet). We then apply a clustering algorithm to group the polygons belonging to the same aesthetic group of buildings, considering building area, height, roof shape, roof color, and local density. The third step consists in extracting facade information from street level images (if available) using another deep learning model based on U-Net [1]. This information is the representative color of the wall, the size and spacing of the windows.

The fourth step is the extraction of roof, wall, and window textures from the semantic data estimated by the previous steps. Each polygon cannot have a street level image of its corresponding facade, on the one hand because it may not be available, and on the other hand because processing as many facade images as

buildings would slow down the automatic chain considerably. To feed facade information to all the polygons that do not have a corresponding facade image, for each cluster we propagate the facade information extracted for some buildings of the cluster to all the other buildings of the same cluster. Since clusters are computed to group similar buildings (in terms of size, height, roof color, etc.), this propagation of information by cluster allows to follow the architectural style division of the area. Finally, the fifth step applies our grammatical rules for generating textured 3D buildings from all the semantic data extracted so far. These rules define, for example, how to generate the geometry of roofs, windows, as well as their texture.

To illustrate the extraction of facade semantics, Fig. 7 (left) shows an example of a street level image that we could process. It is a georeferenced photosphere of a viewpoint close to a facade of interest. The façade rectification step consists in reprojecting this image on the theoretical facade (the one extracted by neural network on the orthoimage). This gives the facade image (Fig. 7 (right)). A neural network then segments this image, Fig. 7 (right) shows the detection of windows and other objects.

**Figure 7: (Left) Example ground image. (Right) Example of an orthorectified photosphere to the facade on interest, overlayed with object extraction.**

Fig. 8 shows example results on the reconstruction of Mourmelon-le-Grand, visualized in Unreal engine v5. The trees come from another specific automatic chain of ours. The terrain is overlaid with the input ortho image.

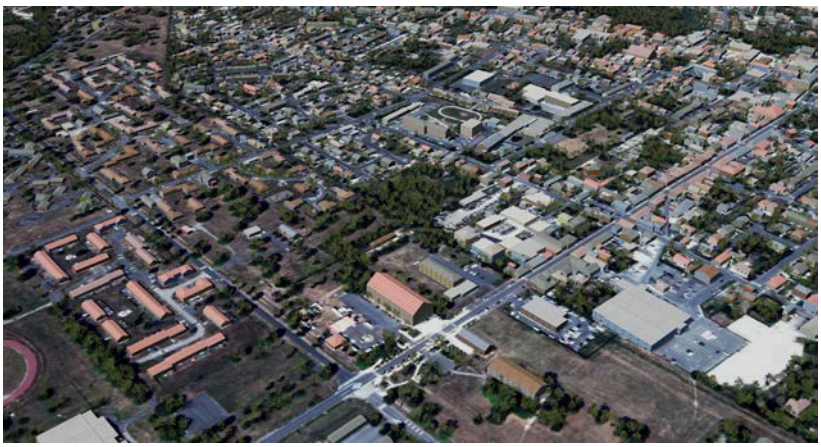

**Figure 8: Visualization in Unreal Engine v5 of automatic reconstruction of Mourmelon-le-Grand, France.**

Finally, we compare the ortho image and our 3D rendering from top view (Fig. 9). We observe that we have very few building detection errors, the majority of the roofs have the right type with a proper generated 3D shape and their texture represents well the color of the roof as exist in the satellite imagery.

On Mourmelon-le-Grand (0.7 km², 2377 buildings) the whole pipeline runs in a few minutes (on a desktop computer with a GTX 1080 Ti GPU) when ingesting only a single orthoimage. When using a stereo pair and street level images (for 318 facades) for more fidelity, the run time is around 50 minutes (including download time of the street level images from a remote server).

## 3.1 Streaming in 3D Tiles format

The output of our automatic pipeline is formatted into 3D Tiles format to be streamed and visualized through the Cesium client (see Fig. 10). 3D Tiles is an OGC specification designed specifically for the streaming and rendering of massive, heterogenous 3D geospatial content. This format is notably built on glTF to ensure fast and lossless streaming of the tiles containing massive tileset. Cesium is a platform designed for 3D geospatial visualization. CesiumJS and Cesium for Unreal have been used for visualizing 3D Tiles streamed either from the Cesium Ion hosting service or a local server.

For usage in mixed reality application or in real-time simulation visualisation engine, high performance is required, with a high and stable framerate, at 60 Hz or even 120 Hz. To optimize the performance and quality of visualization, the 3D Tiles generation method must be carefully designed to build an optimal imagery pyramid. The choice of grouping for buildings, the tile initial grid or the level of the pyramid have multiple influences on the performance result, and the choice of these criteria can depend on the target engine and device.

**Figure 9: Comparison between orthoimagery ( top) and nadir view of final 3D reconstruction (bottom).**

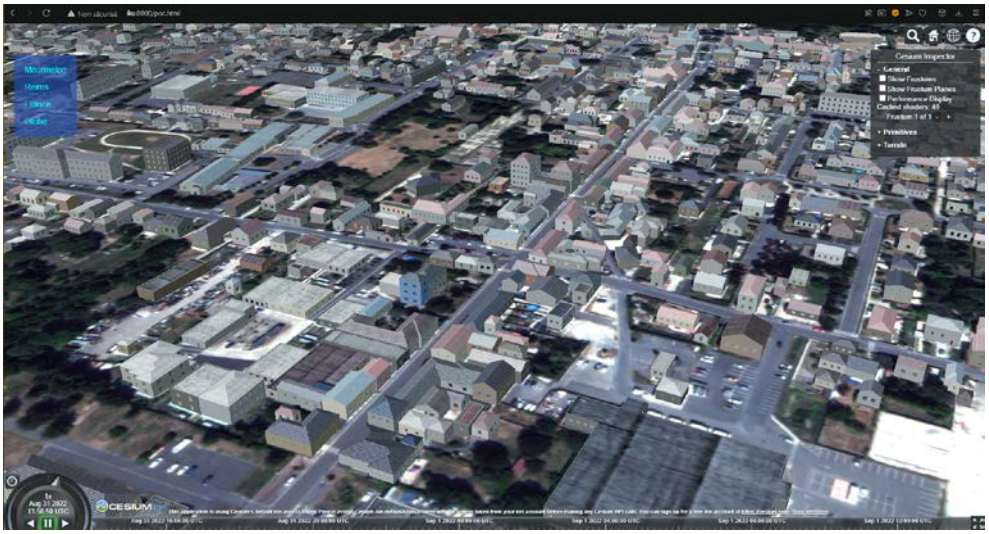

**Figure 10: Final 3D reconstruction streamed with Cesium through an internet browser.**

# 4.0 EXPLOITATION IN THE CONTEXT OF MIXED REALITY

## 4.1 Purpose of the demonstrator

The objective of this demonstrator is to display the database generated from the automatic chain described above, in a Microsoft HoloLens 2, in order to compare it to the reality on the ground of the city of Mourmelon-le-Grand. This comparison will focus on two points:

- Check the correspondence of the volumes and appearances of the buildings.
- Check that the volumes from automatic chain data allow masking.

## 4.2 Demonstrator scenario

An operator equipped with a HoloLens 2 will be placed on a specific point of the Mourmelon-le-Grand site. After calibration, at the launch of the software, the operator will be able to view a 3D database of the city of Mourmelon-le-Grand generated from the automatic chain data. This visualization will be done in overprint of the reality visible through the headset. In addition to the automatic chain database, the

HoloLens will display virtual vehicles in this environment.

The operator will then be able to disable the graphical display of the automatic chain database, but this will still be taken into account by the HoloLens software to calculate the masking. This masking consists of hiding a virtual vehicle (example: jeep, helicopter, ...) when it is supposed to be located behind a building, in order to prevent the virtual vehicle from being displayed in front of the building permanently (see Figs. 11-12).

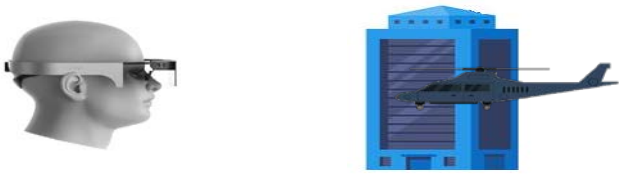

**Figure 11: No masking.**

**Figure 12: With masking.**

## 4.3 Data to be analyzed

The points to be analyzed for this demonstrator are:

- The accuracy of the operator's placement in the real world and its correspondence in the virtual.
- The accuracy of the correspondence between the volumes of real and virtual buildings.
- The precision of the correspondence between the textures of virtual buildings and reality.
- The proper masking of volumes by virtual buildings

## 4.4 Technical difficulties and choices

The HoloLens v2 must be used outdoors (using filters).

3 challenges need to be addressed:

- Avoid saturating the cameras with too much brightness.
- Precisely position the carrier with latitude and longitude coordinates in WGS 84.
- Properly manage the orientation of the operator's head to match the virtual world with the real world.

For the software CS Group naturally turned to the use of the Unreal Engine game engine for the development of the HoloLens 2 software.

## 4.5 Results

This application found that:

- The volumes of automatic chain data can be considered accurate if we ignore the architectural specificities of the buildings (roof decorations, arches, etc.).
- In good conditions, the masking of buildings is successful, and the effect created when elements pass behind is functional (see Fig. 13).

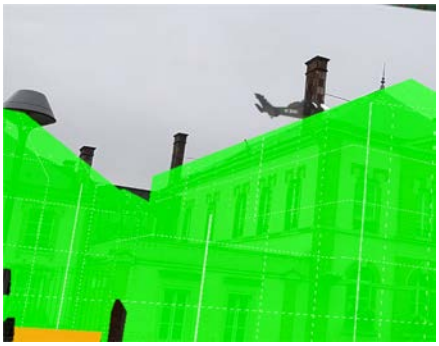

**Figure 13: Calibration and masking of buildings.**

- Matching virtual data with real data is tricky with a HoloLens, especially because of the need to geolocate virtual data.
- There is a lack of sensors/access to the system allowing their additions (No GPS, no compass).
- Inaccuracies of the HoloLens (sensors, restitution) can be added to that of the automatic chain data.
- HoloLens technology is not comfortable in outdoor use due to the brightness of the screens and difficulties interacting with the environment. The sensors are also dazzled and interacting with the environment can become very difficult.
- One could also imagine a version of a much more robust XR headset that could be used in the military context as HoloLens 2 is quite fragile.
- A rather limited field of view that allows holograms to be displayed only in the center of the eye.
- Quite limited power. Even though it is very fluid for small areas, the device has only 4 GB of RAM and a limited GPU which requires paying attention to the size of the 3D models, their quality as well as their textures.
- Some features that cannot be disabled. Indeed, the detection of the hands and the environment can sometimes be a problem, especially when holding an object, the HoloLens does not really know how to interpret the information and it can happen strange actions (Tremors, interface opening).
- For the environment, spatial mapping is between 0.85m and 3.1m, which virtually recreates the environment in which the user is located and can create artifacts compared to what the user would like to display on the HoloLens.
- The development kit used for the application is Microsoft's MRTK (Mixed Reality Tool Kit) which already implements the OpenXR plugin itself. The MRTK is a development kit complete enough to make mixed reality.

Finally, the application itself was very interesting, and allowed to compare the data automatically reconstructed by LuxCarta with the real buildings, and validate that the masking works correctly. On the other hand, the HoloLens 2 may not be the best mixed reality headset to use, or at least for the moment because it is not equipped with essential sensors (GPS and compass), and there are still many problems

related to the headset itself and its use, especially outdoors.

## 5.0 CONCLUSIONS

This paper has described a successful collaboration between LuxCarta and CS-Group, in the frame of the R&T project CREAS-MAP supported by DGA and AID France. An operational pipeline for automatic generation of 3D environments has been developed and validated by exploiting the generated scene in the context of mixed reality at ground scale.

The strength of our method is its ability to automatically generate 3D environments, including 3D terrains and clean and compact LOD2 textured buildings from minimal input data. Our pipeline can thus be used anywhere in the world and runs quickly. It is also flexible in the sense that it can generate a 3D environment from a single satellite orthoimage, a pair of images, or ingest additional information such as street level images. The reconstructed 3D environments can be further exported in the CDB and/or 3D Tiles format, enabling their direct use in different simulation environments.

## 6.0 ACKNOWLEDGEMENTS

The authors would like to thank DGA and AID France for funding this project. We thank DGA experts and architects for their deep involvement and fruitful exchanges during the overall duration of the project. Special thanks to College Technique de Référence de la Simulation (CTRS) which provided relevant use-cases to assess the pipeline products addressing both simulation for acquisition and operational requirements.

## 7.0 REFERENCES

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

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

[3] Tripodi, S. et al, (2022) BRIGHTEARTH: Pipeline for the on-the-fly 3D reconstruction of urban and rural scenes from one satellite images, ISPRS Annals.

[4] Le Saux, B N. Yokoya, R. Hansch, and M. Brown, (2019) 2019 IEEE GRSS data fusion contest: large-scale semantic 3d reconstruction," IEEE GRSM, pp. 33–36.

[5] Hirschmuller, H. (2008) Stereo processing by semiglobal matching and mutual information, IEEE TPAMI, vol. 30, no. 2, pp. 328–341.

[6] Tripodi, S. et al., (2020) Operational pipeline for large-scale 3D reconstruction of buildings from satellite images, in IGARSS.

[7] Mousa, A.-K., Helmholz, P., Belton, D. et al., (2017) New DTM extraction approach form airborne images derived DSM. International Archives of the Photogrammetry, Remote Sensing & SIS, 42.