IMPROVING PAIRWISE DSM WITH 3SGM: A SEMANTIC SEGMENTATION FOR SGM USING AN AUTOMATICALLY REFINED NEURAL NETWORK.

L. Dumas[1]*, V. Defonte[1], Y. Steux[1], E. Sarrazin[2]

[1] CS, 6 rue Brindejonc des Moulinais, Toulouse, France – loic.dumas@csgroup.eu
[2] Centre National d'Etudes Spatiales (CNES), 18 avenue E. Belin, Toulouse, France

Commission II, WG II/4

KEY WORDS: Stereo-Matching, Semantic Segmentation, Refinement, Interactive Learning, Disparity, Optical Satellite Imagery,

ABSTRACT:

The amount of very high resolution optical satellite images at our disposal is continuously increasing. Besides, associated satellite programs often come with high revisit rates and geometric properties that allow for either opportunistic or by-design 3D stereo reconstruction. Digital Surface Models (DSM) computed from these satellite images offer new possibilities. In the past, the high revisit rate has largely benefited glacier monitoring studies. Now, DSM with increased resolution provided on urban areas can be used for smart city applications as well. However, most of these require 3D modeling of buildings with level of details ranging from 0 to 2. This is where the need for better reconstructed buildings inside DSM arises. Indeed, building edges and corners tend to be smoothed and softened by the stereo matching step of a DSM computation pipeline. This undesired behavior can mostly be linked to the difficult task of optimizing the Disparity Space Image, thus finding good balance between smoothing untextured areas while conserving sharp discontinuities where needed. In this paper, we show how the optimization can benefit from an input building semantic segmentation. We also provide a method to create it from a very high satellite image in epipolar geometry using a convolutional neural network. To help our network generalize well on unseen areas we propose an interactive learning method based on clicked annotations. Eventually, we show that annotations can be automatically created, hence removing the need for an operator and making our solution suitable for operational conditions.

1. INTRODUCTION

Very High Resolution (VHR) satellite images often are provided by agile satellites that allow for 3D reconstruction from stereo acquisitions. Data collected from WorldView2 and 3, Pléiades or Pléiades Néo programs, to name a few, are of great interest to help comprehend our environment and its evolution. Besides, devoted stereo missions like CO3D (Melet et al., 2020) are making a statement on the importance of 3D data for many applications. These applications benefit from the ability to constantly refresh the Digital Surface Model (DSM) with newly acquired stereo pairs. The high revisit rate might open the door for 3D time series. Plus, the increasing DSM resolution obtained from satellite imagery allows to focus on urban areas, where new challenges arise from smart city to business intelligence applications. Monitoring the expansion of a city requires a level 1 of detail (LOD1) for 3D buildings representations while evaluating solar rooftop potential would need LOD2 representations.

Many 3D reconstruction pipelines have recently been designed to cope with satellite stereo pairs (De Franchis et al., 2014, Qin, 2016) and focus on scalability for operational conditions (Michel et al., 2020). However, resulting DSMs still present noisy data and overly smooth transitions near building discontinuities. One intuitive way to improve DSM quality is to combine multiple stereo pairs. Multiple views methods mostly depict two categories that we could call *True MultiView* (TMV) and *MultiView Stereo* (MVS). In remote sensing community, MVS seems to be the most efficient method (Zhang et al., 2019). In fact, authors of (Facciolo et al., 2017) and (d'Angelo et al., 2019), respective winners of the 2016 IARPA Challenge (Bosch et al., 2016) and 2019 Data Fusion Contest (Bosch et al., 2019) used MVS based methods. The idea is to compute pairwise DSMs and then merge them together instead of building a single point cloud from all images at once.

To collect multiple satellite views, a large period of time is required as evidenced by the datasets provided for 3D reconstruction challenges (Bosch et al., 2016, Bosch et al., 2019) where more than a year separates first from last acquisitions. Therefore inconsistencies between pairwise DSM are likely to appear and complicate the DSMs merging step. Semantic segmentation (d'Angelo et al., 2019, Qin et al., 2019) and uncertainty map (Qin et al., 2022) can help merge inconsistent DSMs but it still consists in choosing one out of several plausible elevations for the time period observed. For this reason, MVS methods might not be suitable for scene monitoring as they hide changes that end-users might want to analyze.

Another approach for DSM refinement is to use neural networks to post process the DSM. In (Bittner et al., 2018) DSMs are still computed from MVS methods but the amount of images, and with it the acquisition period, is reduced. As so the DSM quality is reduced too, hence the need for a post processing refinement step. Though the network architecture evolved between the works of (Bittner et al., 2018) and more recent contributions from (Stucker and Schindler, 2022) the idea is still to learn how to make DSM ressemble rasterized LiDAR or City GML 3D models. The main drawback from this approach being that such ground truths are tree free models. So as the network learns to create very sharp building edges, it also learns to discard the vegetation. Plus, so far generalization ability has only been tested across two cities, namely Berlin in Germany and Zurich in the German-speaking part of Switzerland (Stucker and Schindler, 2022).

* Corresponding author

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-2-2022-167-2022 | © Author(s) 2022. CC BY 4.0 License.

In this paper, we choose to focus on improving pairwise DSMs as we target monitoring applications and operational conditions. More precisely we aim to improve the Disparity Space Image (DSI) from which a DSM is computed. We review the state of the art stereo matching methods and propose a new contribution to the stereo matching pipeline of CARS (Michel et al., 2020) that is able to provide sharper building edges by using building semantic segmentation prior. The segmentation is computed using a Convolutional Neural Network. To achieve good results and ensure generalization on unseen areas and possibly different sensors, we train our network using an interactive learning method based on local annotations (points). We then demonstrate the feasibility of automatically computed annotations to remove the need of an operator. Eventually, we compare our proposal against state of the art methods.

2. RELATED WORK

2.1 Stereo matching

Although stereo matching pipelines used in (De Franchis et al., 2014, Qin et al., 2019, d'Angelo et al., 2019, Michel et al., 2020) may differ, they all feature the common matching steps as defined in the taxonomy of (Scharstein and Szeliski, 2002):

1. Disparity Space Image (DSI) computation
2. DSI optimization
3. Disparity Map computation

For 3D applications, especially on urban areas, the similarity measure used for DSI computation must limit fattening effect. Hence, the non parametric Census Filter with Hamming distance (Zabih and Woodfill, 1994) is often used. However using the Census Filter creates a rather noisy DSI that requires an optimization. The most popular optimization method is the Semi-Global Matching (SGM) (Hirschmuller, 2005) designed by H. Hirschmuller. SGM is often part of 3D reconstruction pipelines albeit with some adjustments (Facciolo et al., 2017, Qin, 2016, Michel et al., 2020). In fact, SGM-based stereo matching pipelines are top performers on both the IARPA Challenge of 2016 (Bosch et al., 2016) and the 2019 Data Fusion Contest (Bosch et al., 2019).

The idea is to optimize the DSI on a given number of directions, thus creating as much DSIs as there are directions. These DSIs are then summed up to compose the final and optimized DSI. To avoid undesired disparity jumps, two penalties named P1 and P2 are proposed (see EQ.1 with notations from (Hirschmuller, 2005)). P1 will penalize small disparity jumps while P2 will penalize larger ones. Then the disparity map is computed from the optimized DSI by using a classic Winner Take All (WTA) method. Ideally, this disparity map would display sharp transitions near discontinuities (building edges) and smooth ones on flat and slanted surfaces.

However, finding right values for P1 and P2 is far from an easy task. Incorrect values may lead to dissolved building edges in the disparity map and the DSM. As so, Hirschmuller introduced variable penalties that depend on image gradients. The rather strong underlying assumption being that depth discontinuities are linked to radiometric discontinuities and vice versa. Further studies compared different penalty functions (Banz et al., 2012) based on the same assumption. Unfortunately successive strong radiometric discontinuities may appear on agricultural fields while shadowed transitions between buildings and ground can lead to small radiometric gradients.

In (Scharstein et al., 2017) surface priors are used to enforce accommodation to slanted surfaces that are mainly present on datasets designed for autonomous vehicles. Penalties are no longer set depending on disparity jumps, rather P1 and P2 are used to penalize divergence from the surface assumption. Obviously the result depends on the ability to extract correct surface priors. Hence our proposed method adapts SGM equation to add geometric priors in the form of building semantic segmentations that clearly identifies discontinuities near building edges. With this proposition arises a new challenge that is the creation of this building semantic segmentation.

2.2 Building semantic segmentation neural networks

Semantic segmentation is an important task within computer vision and remote sensing community for it helps to better comprehend the world we live in. Using Convolutional Neural Network (CNN) promising results have been reached even in the field of satellite imagery (Kussul et al., 2017). However, the major challenge remains to find a solution that will be robust enough to face the rich visual variations of landscapes and man-made structures that satellite imagery can unveil around the globe. Though the amount of images is large enough to train deep CNN, ground truth labels are still very sparse. Challenges (Demir et al., 2018) provide some but those are often not enough for the network to generalize well to unseen areas or, in the case of building semantic segmentation, unseen architectures. To overcome this limitation, new machine learning paradigms have been designed to reduce the volume and sometimes the quality of ground truth labels required to train such CNNs. Among these is interactive learning that allows the network to learn continuously. Assuming one cannot train a robust enough network to the task of building semantic segmentation then interactive learning provides hope in the form of an interaction between the machine and the operator. Thus, the benefit of this method comes at the price of a human-in-the-loop. Indeed, the operator will be asked to provide ground truth annotations or validations that will help the network improves its inner parameters and refines its future predictions.

The first framework designed for this purpose is the work of (Xu et al., 2016). The authors proposed a method called DIOS that concatenates two additional channels to the input image. These channels respectively contain positive and negative annotations for a mono-class task. Both channels then act as a binary mask. Originally thought for binary classification, DIOS has been extended to multi-classes segmentation and aerial data with the work of (Lenczner et al., 2020). The author proposes a framework named DISCA, allowing the operator to smoothly interact with the network by providing left and right clicks to create the annotated masks. This annotation process results in local points (on clicked areas). The number of points, their sizes and locations greatly affect the network learning improvements. In (Benenson et al., 2019) the authors investigate many aspects of this human-machine collaboration and give insights to reduce the burden of this interaction and demonstrate large-scale feasibility.

Although the work of (Benenson et al., 2019) constitutes a big leap forward towards quick and efficient annotations, it still is incompatible with constraints of an image ground segment. Indeed, it would require one operator to interact with the network in operational conditions. Hence, we propose an algorithm to create automatic annotations from sparse elevation clues that are deduced from early steps of the 3D reconstruction pipeline.

3. PROPOSED ALGORITHMS

The main idea of our proposal (illustrated in FIG. 1) is to create a building semantic segmentation from the left epipolar image and use it to optimize the DSI. We then propose a slight modification of SGM optimization method to incorporate the semantic segmentation. We refer to this new SGM-based optimization as Semantic Segmentation for SGM (3SGM). We propose the use of a LinkNet (Chaurasia and Culurciello, 2017) to create the building semantic segmentation. We train the network on a rather limited dataset (see section 4.2.1) and call this trained network the Initial Neural Network (INN). Because it does not generalize well to unseen architectures, we improve it using annotations and the DISCA interactive learning framework from (Lenczner et al., 2020). Aiming for our solution to be used in operational conditions, we propose to automate the creation of annotations by using sparse elevation clues deduced from a rough Disparity Space Image (DSI), that is the non optimized DSI. We call ARNN (Automatically Refined Neural Network) the INN refined with automatic annotations. Combined together, the 3SGM method and the ARNN constitute our proposed stereo matching pipeline. They are individually and jointly evaluated section 4.

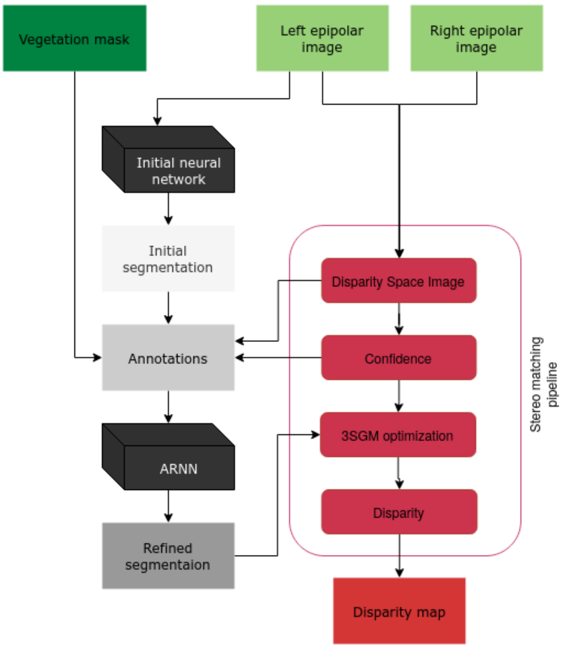

Figure 1. Our stereo pipeline: 3SGM optimization with labels from an Automatically Refined Neural Network (ARNN).

3.1 3SGM: Semantic Segmentation for SGM

To introduce the building semantic segmentation inside the original SGM equation (EQ. 1) we simply stop the optimization of a given path every time this path crosses a building edge as illustrated on FIG. 2. Then the optimization starts again from the very next pixel.

$$L_r(p,d) = C(p,d) + min \begin{cases} L_r(p-r,d), \\ L_r(p-r,d-1)+P_1, \\ L_r(p-r,d+1)+P_1, \\ L_r(p-r,i)+P_2, \end{cases} \tag{1}$$

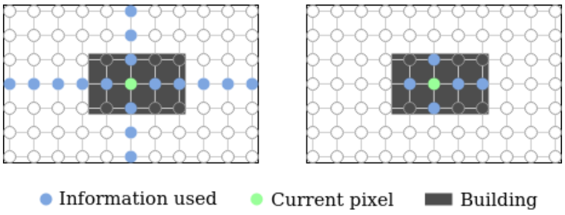

Figure 2. Illustration of classic SGM (left) and our implementation (right) with 4 directions. History of previous pixels along the path is discarded when a building edge is met.

This methodology can be resumed by EQ. 2, where the history of the current path is discarded when the segmentation values of two consecutive pixels differ. In the meantime, we keep fixed values for both penalties and for every stereo pairs and we use no relation between image radiometric information and the penalties. Optimizing the DSI per segment obviously requires a strong confidence in the segments quality. Nevertheless, having a look at extreme possibilities, it is easy to see that in the worst case scenario no optimization is done. This would be when the semantic segmentation always alternates between building and ground labels. On the other hand, if the segmentation only contains a single segment then the classic SGM optimization is performed.

$$L_r(p,d) = C(p,d) + \beta * min \begin{cases} L_r(p-r,d), \\ L_r(p-r,d-1)+P_1, \\ L_r(p-r,d+1)+P_1, \\ L_r(p-r,i)+P_2, \end{cases} \tag{2}$$

with $\beta = (Seg(p-r) == Seg(p))$.

3.2 ARNN: Automatically refined Neural Network

The chosen neural network architecture (LinkNet) only takes RGB as input channels. Hence we can assume it will try to detect building pixels mainly based on roof and neighborhood colors. While this shall not prevent the network from detecting building edges and corners, we believe there is a strong possibility for the network to entirely miss buildings of unseen architecture.

To prevent this from happening, we use interactive learning following the DISCA workflow (Lenczner et al., 2020) to refine the initial neural network (INN) and with it, its predictions. In the work of (Lenczner et al., 2020) annotations are manual clicks provided by an operator. To remove the need for a human-in-the-loop we propose to automate the annotation process. A parametrizable disparity threshold is applied to a disparity map computed from a DSI that has yet to be optimized. The threshold helps differentiate between ground and roof pixels. Disparities above the threshold are translated to building annotations. We then filter the building annotations to remove false positive on high vegetation areas using a simple vegetation mask (NDVI indicator) computed from the left epipolar image (see FIG. 1). We also remove annotations associated to wrongly matched pixels according to the confidence map presented in (Sarrazin et al., 2021). Because wrong matches tend to occur near building edges, we are left with building annotations mostly located in the middle of building roofs. We assume these are enough for the network to learn new architectures and propagate this knowledge to building edges as shown by the work of (Benenson et al., 2019).

Since the INN has no trouble labelling the ground and because ground is often the most common area, we limit the number of ground annotations we make. The only pixels annotated as ground are the ones that have been assigned a building label by the INN but whose disparities are below the disparity threshold. The underlying idea is that ground disparities are mostly correct matches, albeit noisy ones, while building disparities, especially near building edges, might belong to incorrect matches due to fattening effect.

4. EXPERIMENTS AND RESULTS

In this section, we evaluate the combination of the 3SGM optimization method and the ARNN presented section 3. First (4.1) we evaluate 3SGM using ground truth building labels obtained from Open Street Map (OSM). Then on subsection 4.2 we compare the ARNN against the INN and a Manually Refined Neural Network (MRNN), that is the INN refined using the DISCA interactive learning workflow and manual annotations. Eventually, section 4.3, we display disparity maps and DSMs to compare our stereo matching pipeline, e.g. with 3SGM and ARNN, against state of the art method.

In subsections 4.1 and 4.3, we use CARS (Michel et al., 2020) and its embedded stereo matching framework, Pandora, for our experiments. Both are open source software [1][2] and part of the CO3D (Melet et al., 2020) image processing chain. CARS rectifies the input stereo pair into epipolar geometry. Then we use different stereo matching pipelines created with Pandora to compute and optimize the DSI and evaluate our 3SGM method. To observe and quantify the impact on DSM quality we use CARS to triangulate the disparity maps and create the DSMs.

4.1 3SGM evaluation

To observe how 3SGM behaves, we use a dataset on Montpellier (France) where Pléiades stereo acquisitions, LiDAR, and OSM labels are available. For quantitative evaluation on the Disparity Maps we use 83 stereo pairs of size 1840x1840 created with CARS rectification step. Ground Truth disparities are computed using LiDAR and the process detailed in (Cournet et al., 2020). The same methodology is applied to project OSM labels into epipolar geometry so we can use them with 3SGM.

In TAB. 1 are shown the results of five different stereo matching pipelines. Four of them are combinations of either Census Filter (Zabih and Woodfill, 1994) or MCCNN (Zbontar et al., 2016) matching costs with either SGM optimization (see EQ. 1)) or our 3SGM optimization (see EQ. 2). The last pipeline uses Census matching costs and the More Global Matching optimization presented in (Facciolo et al., 2015).

Census matching costs are computed on 5x5 windows. For MCCNN we use the plugin for Pandora[3] trained on the Middleburry dataset with an 11x11 patch size. Results and analysis of this MCCNN training and its generalization to satellite images are presented in (Defonte et al., 2021). For the MGM optimization we use the original code [4] to make sure we emulate the s2p default stereo matching pipeline. For all optimizations of Census matching costs, the penalties are set to CARS and s2p default values (P1=8; P2=32). Penalties used for the optimization of MCCNN matching costs are set according to the original paper of Zbontar & LeCun (Zbontar et al., 2016).

Results show MCCNN performs slightly better than Census, which is consistent with the work of (Defonte et al., 2021). We can also notice that less disparity errors are produced when using 3SGM with OSM labels. Eventually, using MGM optimization on top of Census matching costs decreases errors with a magnitude higher than a pixel, however the mean, standard deviation and 70 percentile seem to reveal the presence of more outliers.

Methods	% Error $\leq 1px$	Mean error	Std error	70 p
CENSUS with SGM	64.02	1.61	2.64	1.31
CENSUS with 3SGM	66.46	1.50	2.51	1.19
MCCNN with SGM	65.44	1.49	2.34	1.24
MCCNN with 3SGM	**67.24**	**1.40**	**2.27**	**1.12**
CENSUS with MGM	66.20	1.60	2.79	1.23

Table 1. Disparity errors (in pixels) on Montpellier (France). All pixels are considered (no rejection criteria). Optimization with 3SGM uses OSM labels.

To make sure these improvements are still visible in the output DSMs, we use CARS to triangulate the disparity maps created for TAB. 1. On FIG. 3 and 4 we choose to present meshed DSM to better observe building edges and corners. We show CARS default stereo matching pipeline (Michel et al., 2020) along with the pipeline from (Defonte et al., 2021) and the MCCNN with our 3SGM optimization.

On FIG. 3 we can see that using 3SGM edges are sharper and building shapes are less complex. Interestingly enough, building labels also prevent 3D reconstruction pipelines from completely missing buildings, as evidenced by the elongated one in the middle of the scene. We believe this case indicates that penalties used for the optimization of MCCNN costs, though very efficient on untextured areas (Defonte et al., 2021), tend to be too high for urban scenes. This further demonstrate the difficult task of choosing adequate penalties values for operational conditions.

Consequently we observe that sharper building edges help to better recover the correct elevation of the main streets (see FIG. 4). We can also notice that using Census instead of MCCNN with the same optimization method visually gives better results on building edges and streets. We believe this is another demonstration of too high penalties values for MCCNN. Higher values help remove noise on roofs and ground areas thus improving overall metrics as seen TAB. 1 and in (Defonte et al., 2021). The drawback being a tendency to blur building edges and discourage sharp discontinuities. 3SGM then seems to remove noise while still allowing sharp discontinuities where needed.

4.2 ARNN evaluation

4.2.1 Experimental setup We start by training the INN (a LinkNet (Chaurasia and Culurciello, 2017)) with two combined datasets listed in the TAB. 2. The first one is made of 1496 rectified stereo pairs created from the Track3 US3D WorldView-3 dataset (Bosch et al., 2019). These products cover cities of Atlanta, Jacksonville and Omaha (USA). The second one is composed of 327 rectified stereo pairs acquired from Pléiades products on Montpellier (France). Images are tiled into patches

[1] https://github.com/cnes/cars
[2] https://github.com/cnes/pandora
[3] https://github.com/cnes/pandora_plugin_mccnn
[4] https://github.com/gfacciol/mgm

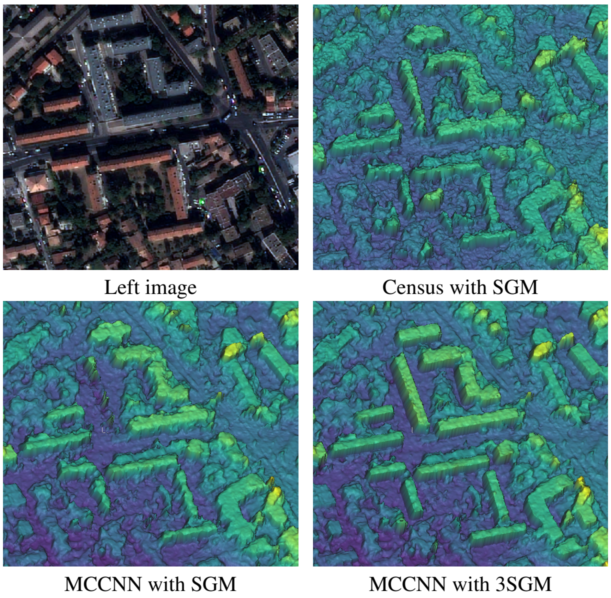

Figure 3. Meshed DSMs on Montpellier (France). DSM building edges and corners are better reconstructed with 3SGM.

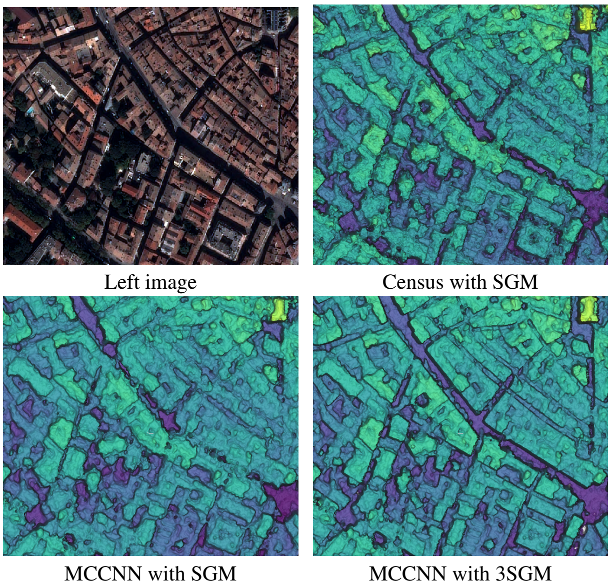

Figure 4. Meshed DSMs on Montpellier (France). Our proposed 3SGM optimization helps recover main streets elevations.

of size 1024x1024. For WorldView-3 products, ground truth building labels are provided by the 2019 Data Fusion Contest (Bosch et al., 2019) while for Pléiades products they are created using OSM labels and the methodology exposed in (Cournet et al., 2020) to obtain the labels in epipolar geometry.

We use the stochastic gradient descent (SGD) optimizer with a base learning rate of 0.1 divided by 10 when the loss has stopped improving. The network is trained with Dice Loss, for 70 epochs with a batch size of 8.

Then we refine the INN over the city of London (UK). We use 6 Pléiades patches of 2500 x 2500 pixels for which we compute semantic segmentation with the INN. Each segmentation is then annotated twice: manually and automatically. For manual annotations, we limit the operator to a 20 minutes time-frame resulting in 150 annotations each on average. Automatic annotations on the other hand are computed as explained in subsection 3.2. Eventually we obtain two refined networks: the MRNN and the ARNN.

Areas	Samples	Sensor	Median B/H
Atlanta	1117	WorldView 3	0.22
Jacksonville	194	WorldView 3	0.24
Omaha	185	WorldView 3	0.25
Montpellier	327	Pléiades	0.68

Table 2. List of products used to train the LinkNet INN.

4.2.2 Results

We use qualitative observations and quantitative measures to evaluate the impact of annotations and in particular to assess our ARNN.

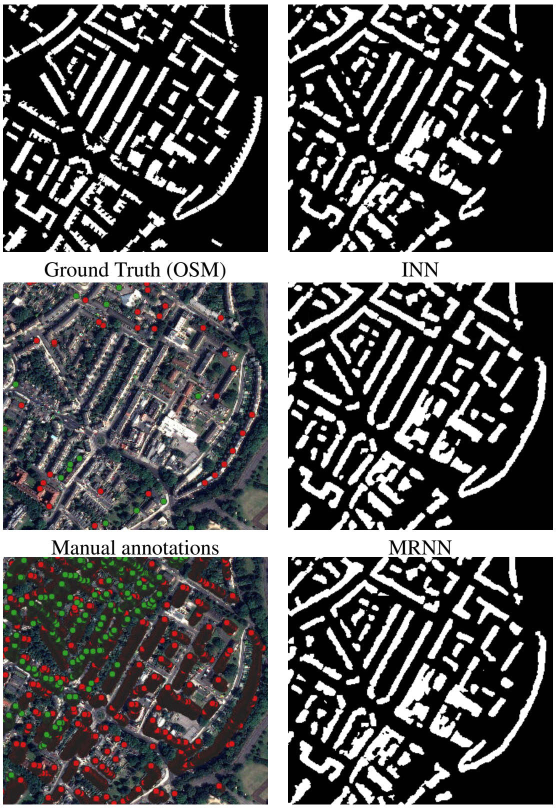

Figure 5. Annotations produced by the manual and automatic annotation strategies: red points indicate ”Buildings” class while green ones indicate ”Not buildings” class.

On Fig. 5, we display the building semantic segmentation generated with the INN, the MRNN and the ARNN from Pléiades product on London (UK). Though we only showcase one illustration, we have observed similar behavior for every cities considered far from the training dataset. We can see that both manual and automatic annotations improve the initial segmentation map. Not only building shapes are better identified but more buildings are well detected. Visual comparison of both refined labels shows no relevant difference, whereas much more annotations are created with the automatic strategy. This seems to indicate that few annotations are required to improve the model accuracy which is on par with the conclusions of (Benenson et al., 2019). It should also be noted that some automatic annotations are wrongly identified as buildings. However this does not seem to impact the accuracy of the refined network, presumably because only a few mistakes are made.

For the quantitative evaluation, we use three cities unseen by the INN during training phase, namely Toulouse, Rennes (France), and London (UK). The products have been acquired by Pléiades. TAB. 3 shows best results are obtained with the MRNN presumably because manual annotations are of better quality. The refinement over London helps improve the IoU on Rennes but results obtained on Toulouse remind us the network must be refined for every stereo pairs, hence our proposed pipeline (see FIG. 1). On another topic we can notice the relative difference between both refined networks that seems to validate our proposal for automatic annotations.

Sites	Neural network	IoU building	IoU not building
Toulouse	INN	59.20	**82.10**
B/H = 0.65	MRNN	**60.40**	81.49
$38km^2$	ARNN	59.12	81.65
London	INN	49.77	93.70
B/H = 0.66	MRNN	**55.54**	**93.72**
$14km^2$	ARNN	53.19	93.29
Rennes	INN	61.09	88.74
B/H = 0.60	MRNN	**64.09**	**89.00**
$18km^2$	ARNN	63.45	88.95

Table 3. Mean IoU obtained with and without annotations.

4.3 Combined 3SGM and ARNN evaluation on DSM

On subsection 4.1 we showed how using 3SGM with OSM can limit noise and sharpen building edges. We now focus on operational and more realistic conditions where OSM labels are not available or outdated. So we use the pipeline shown FIG. 1 and analyse the behavior of 3SGM with labels from the ARNN detailed in subsection 4.2.

For the quantitative results presented in TAB. 4, we use a 14km^2 area over London (UK) as it is where the ARNN perform worst. We use the same stereo pipelines and parameters as the ones previously detailed in subsection 4.1. First, we can observe in TAB. 4 that the results over London are consistent with the ones obtained on Montpellier (France) (see TAB. 1). Then we can notice that of the three pipelines based on 3SGM, the best results are obtained with labels inferred by the MRNN or ARNN. We assume this comes from the time lapse between Pléiades acquisition and OSM last update. Lastly, we can see that MRNN and ARNN give comparable results. This observation encourages us to believe the segmentation neural network can efficiently be refined to new areas and architectures without the need for an operator. Or at the very least, it can be refined for the purpose of 3D reconstruction with 3SGM optimization.

Methods	% Err. $\leq 1px$	Mean error	Std error	70p
CENSUS with SGM	57.40	4.10	7.85	2.68
CENSUS with MGM	59.74	4.20	8.07	2.57
MCCNN with SGM	63.43	2.30	4.89	1.64
MCCNN with 3SGM (OSM)	64.86	2.19	4.82	**1.49**
MCCNN with 3SGM (MRNN)	64.89	2.19	4.81	1.50
MCCNN with 3SGM (ARNN)	**64.99**	**2.18**	**4.78**	**1.49**

Table 4. Disparity errors (in pixels) on London (UK). All pixels are considered (no rejection criteria).

The visual results showcased in FIG. 6 and 7 are consistent with the overall metrics of TAB. 4. Both figures present DSMs obtained with MCCNN matching costs and different optimization methods. As anticipated, building edges and corners are better reconstructed with 3SGM whether labels are inferred with MRNN or ARNN. Yet, FIG. 7 exposes automatic refined segmentation flaw. Indeed, in the top right corner of the scene, part of the road is labelled as building. This is because this road is actually a bridge with a disparity higher than our manually set disparity threshold so that automatic building annotations have

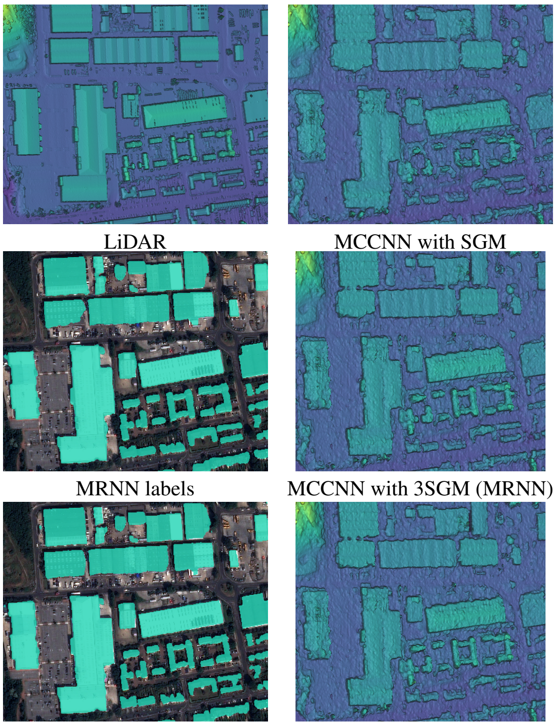
LiDAR
MCCNN with SGM
MRNN labels
MCCNN with 3SGM (MRNN)
ARNN labels
MCCNN 3SGM (ARNN)

Figure 6. DSMs computed on London (UK) with different optimizations. 3SGM method performs best.

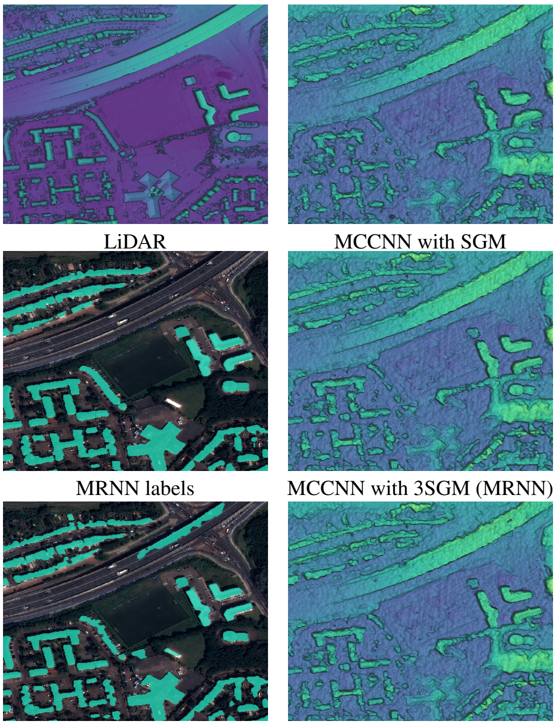
LiDAR
MCCNN with SGM
MRNN labels
MCCNN with 3SGM (MRNN)
ARNN labels
MCCNN with 3SGM (ARNN)

Figure 7. DSMs computed on London (UK) with different optimizations. ARNN errors do not impact 3SGM.

been added in this area. Though this prevent ARNN from reaching expected results, it does not degrade the DSM produced by 3SGM with ARNN labels. In fact, it could actually improve it as it helps identify what might be a strong discontinuity.

The next experiment measures the impact of the 3SGM optimization with ARNN on the DSM quality. We use the MCCNN matching costs for it gives the best results according to TAB. 4. We compare our results with the ones obtained with s2p configuration: a CENSUS matching costs on a 5x5 window followed by a MGM optimization (P1=8; P2=32) (Facciolo et al., 2017). We use CARS to rectify the images, triangulate the disparity maps, and rasterize the DSM in 2.5D. This way we make sure to observe only the impact of both stereo matching pipelines. Nevertheless and to the best of our knowledge, CARS and s2p rectification, triangulation and rasterization are very similar algorithms for small stereo pairs. In TAB. 5 results are obtained using a LiDAR as elevation ground truth. As expected based on the previous results, our proposal compares favorably against state of the art method for pairwise DSM reconstruction.

Methods	Mean error	STD error	RMSE error
DSM with s2p stereo pipeline	-0.55	3.58	3.61
DSM with our stereo pipeline	**-0.10**	**3.39**	**3.35**

Table 5. Evaluation of our stereo matching pipeline (MCNN with 3SGM and labels from ARNN) against state of the art method (s2p) on London (UK).

Eventually, on FIG. 8 and FIG. 9 one can observe meshed DSMs on two distinct areas. Along with the ground truth and the results obtained with the two methods compared in TAB. 5, we expose the DSM computed by CARS with its default stereo pipeline (CENSUS with SGM optimization). DSMs computed with 3SGM, and labels from the ARNN, not only are less noisy than with the classic, CARS default stereo pipeline, but also display sharper building edges. Noise reduction, as previously shown, can be due to the MCCNN matching cost and the high penalties values as suggested by earlier work (Defonte et al., 2021). Sharp discontinuities however are due to 3SGM optimization that relaxes the constraints imposed by high penalties near building edges. This visual impression corroborates with the profiles presented on FIG. 8 since our 3SGM optimization combined with ARNN labels allows discontinuities to appear between successive buildings.

5. CONCLUSIONS AND PERSPECTIVES

In this article, we have shown that pairwise DSM quality can be improved using 3SGM: a proposed SGM-based optimization guided by a building semantic segmentation. Because this proposal requires an input segmentation in epipolar geometry, we proposed the use of an automatically refined neural network (ARNN). We embedded the automatic refinement inside the stereo matching step so that the network can learn continuously. We then showed that DSM computed with 3SGM and labels from our ARNN are less noisy and contain sharper building edges even on geographical site not previously seen by the segmentation network.

Future work will focus on demonstrating how our DSMs can ease the creation of LOD0 to LOD2 building reconstructions. We will also focus on automating the process of setting the right disparity threshold to create annotations.

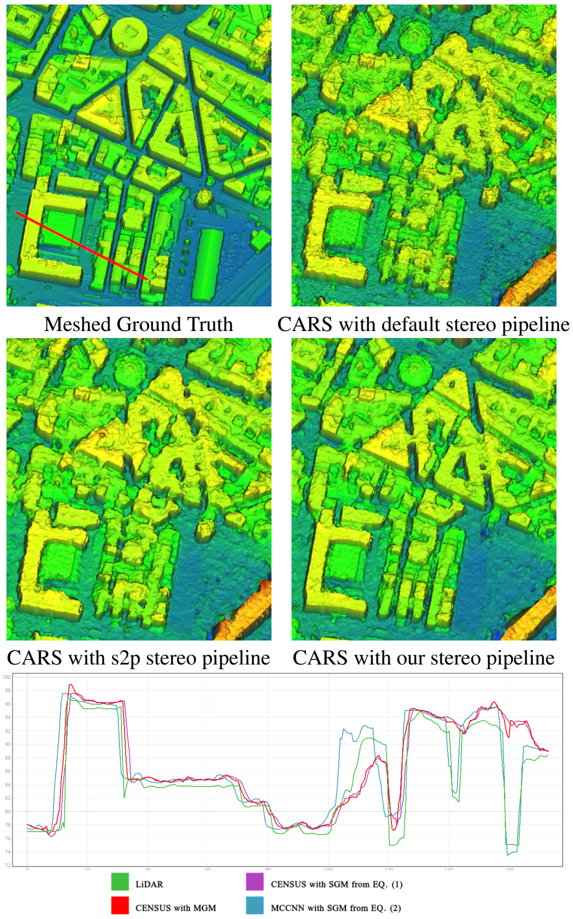
Meshed Ground Truth
CARS with default stereo pipeline
CARS with s2p stereo pipeline
CARS with our stereo pipeline

Figure 8. Comparison of meshed DSMs with different stereo matching pipelines. Both visual comparison and the profiles show our proposal, MCNN with 3SGM optimization and labels from an ARNN, provides the best results.

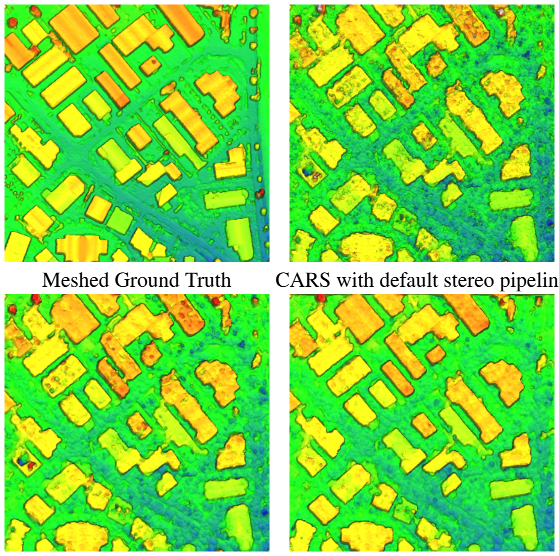
Meshed Ground Truth
CARS with default stereo pipeline
CARS with s2p stereo pipeline
CARS with our stereo pipeline

Figure 9. Comparison of meshed DSMs with different stereo matching pipelines. Our method based on MCCNN matching costs and 3SGM optimization with labels from an ARNN provides the best visual result.

ACKNOWLEDGEMENTS

The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest. In addition the authors are thankful to DELAIR and the ONERA organizations for having released and shared the DISCA workflow under the open source MIT license.

REFERENCES

Banz, C., Pirsch, P., Blume, H., 2012. Evaluation of penalty functions for semi-global matching cost aggregation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences [XXII ISPRS Congress, Technical Commission I] 39 (2012), Nr. B3*, 39number B3, Göttingen: Copernicus GmbH, 1–6.

Benenson, R., Popov, S., Ferrari, V., 2019. Large-scale interactive object segmentation with human annotators. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11700–11709.

Bittner, K., d'Angelo, P., Körner, M., Reinartz, P., 2018. Dsm-to-lod2: Spaceborne stereo digital surface model refinement. *Remote Sensing*, 10(12), 1926.

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G. D., Brown, M., 2019. Semantic stereo for incidental satellite images. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 1524–1532.

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery. *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, IEEE, 1–9.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 1–4.

Cournet, M., Sarrazin, E., Dumas, L., Michel, J., Guinet, J., Youssefi, D., Defonte, V., Fardet, Q., 2020. Ground truth generation and disparity estimation for optical satellite imagery. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 127–134.

De Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., Facciolo, G., 2014. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Defonte, V., Dumas, L., Cournet, M., Sarrazin, E., 2021. Evaluation of mc-cnn based stereo matching pipeline for the co3d earth observation program. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, 7670–7673.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 172–181.

d'Angelo, P., Cerra, D., Azimi, S. M., Merkle, N., Tian, J., Auer, S., Pato, M., de los Reyes, R., Zhuo, X., Bittner, K. et al., 2019. 3d semantic segmentation from multi-view optical satellite images. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 5053–5056.

Facciolo, G., De Franchis, C., Meinhardt, E., 2015. Mgm: A significantly more global matching for stereovision. *BMVC 2015*.

Facciolo, G., De Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 57–66.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2, IEEE, 807–814.

Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778–782.

Lenczner, G., Chan-Hon-Tong, A., Luminari, N., Saux, B. L., Besnerais, G. L., 2020. Interactive Learning for Semantic Segmentation in Earth Observation. *arXiv preprint arXiv:2009.11250*.

Melet, O., Youssefi, D., L'Helguen, C., Michel, J., Sarrazin, E., Languille, F., Lebègue, L., 2020. Co3d mission digital surface model production pipeline. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 143–148.

Michel, J., Sarrazin, E., Youssefi, D., Cournet, M., Buffe, F., Delvit, J., Emilien, A., Bosman, J., Melet, O., L'Helguen, C., 2020. A new satellite imagery stereo pipeline designed for scalability, robustness and performance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 171–178.

Qin, R., 2016. Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 77.

Qin, R., Huang, X., Liu, W., Xiao, C., 2019. Semantic 3d reconstruction using multi-view high-resolution satellite images based on u-net and image-guided depth fusion. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 5057–5060.

Qin, R., Ling, X., Farella, E. M., Remondino, F., 2022. Uncertainty-Guided Depth Fusion from Multi-View Satellite Images to Improve the Accuracy in Large-Scale DSM Generation. *Remote Sensing*, 14(6), 1309.

Sarrazin, E., Cournet, M., Dumas, L., Defonte, V., Fardet, Q., Steux, Y., Diaz, N. J., Dubois, E., Youssefi, D., Buffe, F., 2021. Ambiguity Concept in Stereo Matching Pipeline. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 383–390.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1), 7–42.

Scharstein, D., Taniai, T., Sinha, S. N., 2017. Semi-global stereo matching with surface orientation priors. *2017 International Conference on 3D Vision (3DV)*, IEEE, 215–224.

Stucker, C., Schindler, K., 2022. RESDEPTH: A deep residual prior for 3D reconstruction from high-resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 560–580.

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T. S., 2016. Deep interactive object selection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 373–381.

Zabih, R., Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. *European conference on computer vision*, Springer, 151–158.

Zbontar, J., LeCun, Y. et al., 2016. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1), 2287–2318.

Zhang, K., Snavely, N., Sun, J., 2019. Leveraging vision reconstruction pipelines for satellite imagery. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.