BUILDING SEGMENTATION BASED ON STEREO INFORMATION FROM SATELLITE IMAGES

M. Roux[1]*, L. Dumas[1]

[1] CS, 5 rue Brindejonc des Moulinais, Toulouse Cedex 5, France - (mathis.roux, loic.dumas)@csgroup.eu

Commission II, WG II/6

KEY WORDS: Semantic Segmentation, Convolutional Neural Network, Computer Vision, Stereo Vision, Stereo Segmentation, Stereo Matching, Building Segmentation

ABSTRACT:

Building semantic segmentation is key to many applications relying on 3D modeling of city buildings such as urban planning or business intelligence. Recent works have shown great improvements in this area thanks to artificial intelligence, but even state of the art neural networks encounter difficulties to generalize to buildings that are different from the training dataset. 3D modeling applications also requires the elevation information often retrieved from a pair of High Resolution satellite images. In this article, we show that using both images of a stereo pair as inputs to a neural network trained for building semantic segmentation achieves better results than using a single view. Especially, stereo training gives a greater ability to generalize. We show that using neural networks designed for disparity estimation performs well for building semantic segmentation from a pair of satellite views in epipolar geometry. We also discuss how radiometry and disparity both affect the definition of what a building is depending on the multi-view network architecture.

1. INTRODUCTION

Automatic acquisition of geospatial information is of great interest especially in the context of 3D modeling of cities, and more and more softwares like (Baillarin et al., 2020) require accurate segmentation pipelines. Artificial intelligence technologies have permitted great improvements for semantic segmentation in the field of computer vision (Hao et al., 2020) but also for remote sensing applications like building segmentation (Liu et al., 2020). One of the main challenge of artificial intelligence is generalization ability. In remote sensing, good generalization is often difficult to achieve : a network trained in a particular city will likely have difficulties when tested on a new city (Maggiori et al., 2017). The goal of this study is to improve building segmentation results and generalization ability of networks, especially for 3D modeling applications. These applications use stereo images in order to retrieve 3D information. More and more satellite missions allow the production of high resolution stereo images like WorldView and Pléiades, or the future CO3D mission (Lebègue et al., 2020) which is especially designed for 3D applications. Our idea to improve the generalization ability of segmentation networks is to use stereo images as input in order to give elevation clues to the network. The hypothesis tested in this study is that a stereo neural network could have a better characterization of a building by using elevation features extracted from stereo images.

A review of the different architectures of neural networks used for monocular semantic segmentation, multi-view semantic segmentation and stereo disparity estimation is presented section 2. Then the section 3 introduces the networks used, the dataset on which we trained them and the methodology we followed. Results are shown on section 4. The section 5 presents additional experiments that evaluates how much the stereo networks rely on disparity features to infer segmentation.

* Corresponding author

2. RELATED WORK

2.1 Monocular semantic segmentation

Semantic segmentation is the task of assigning each pixel of an image to the object class it belongs. In computer vision field, most semantic segmentation methods use deep learning. The networks used in remote sensing are often taken or inspired from the ones designed for computer vision. The classical network designed for semantic segmentation is U-Net (Ronneberger et al., 2015). It is a fully convolutional neural network composed of a downsampling feature extractor followed by an upsampler.

The convolutional neural networks designed afterwards follow the same pattern as U-Net with additional modules like pyramid spatial pooling (Zhao et al., 2017) or atrous convolutions (Chen et al., 2017).

Several articles compared these architectures over different remote sensing datasets of building segmentation as the Inria Labeling Dataset (Maggiori et al., 2017) or Potsdam and Vaihingen datasets from the ISPRS. (Hu et al., 2019) shows that U-Net, Deeplab and PSPNet have similar results. Networks used in this study will be similar to U-Net to keep a simple and efficient architecture.

2.2 Multimodal semantic segmentation

Results presented on the section 2.1 concerns segmentation on RGB images, or IRRG images for Vaihingen dataset. However remote sensing data can often be associated to additional data like multispectral images or Digital Surface Models (DSM). As our goal is to give elevation clues to the network, one possibility could be to associate the satellite image with a DSM. Several networks have been designed for multimodal feature fusion of remote sensing images. (Audebert et al., 2018) fuses IRRG

images with DSM, nDSM and NDVI data on the Postdam and Vaihingen datasets. The results show a significant improvement for vegetation class but not for building class. According to the study the quality of the DSM could be a problem. (Lei et al., 2021) uses attention mechanism for multi-modality and multi-scale fusion. IRRG images and nDSM are used from the same ISPRS datasets. The article show the efficiency of multi-modal fusion between nDSM and IRRG image.

Fusion of image and DSM is an interesting solution but could raise technical difficulties. First the quality of the segmentation relies on the quality of the DSM, and DSM often lacks accuracy particularly for small buildings. Then the image and the DSM should be aligned which means the image should be orthorectified in the exact same geometry as the DSM. We choose to use stereo images and rely on the neural network to extract elevation features instead of directly giving the DSM to the network because it is a much simpler pipeline.

For this purpose we have to find or design a network able to process multiple views.

2.3 Multi-view segmentation

Multi-view segmentation have not been heavily studied in computer vision. Some networks have been designed to process medical data (CT scans or MRI) for semantic segmentation from the multiple views (Mortazi et al., 2017), but this type of architecture is not adapted to satellite images.

INSTR (Durner et al., 2021) is a classless instance segmentation network that works with stereo RGB images. The network predicts both the segmentation on the left image and the disparity between the two images. The ablation study have shown an improvement of 8 to 11% of IoU for stereo segmentation compared to monocular segmentation.

The Track 2 of the 2019 Data Fusion Contest (Bosch et al., 2019) have gathered several researchers around challenges of disparity estimation and semantic segmentation of stereo images. The dataset given is made of stereo images in epipolar geometry took by WorldView-3 above the cities of Jacksonville and Omaha (USA). The network used for stereo segmentation by the winner of this contest (Chen et al., 2019) is called DFSN. This network is different from all the ones presented above because it integrates a disparity estimation network called PSMNet (Chang and Chen, 2018). DFSN has three convolutional encoders : one for the left image, one for the right image, and the one of PSMNet that processes a cost volume with 3D convolutions. The three feature maps obtained are concatenated and given to the decoder. The results of DFSN have been compared with a monocular network : SSN-RGB. DFSN improves the results of the segmentation by 1.5% of IoU for all classes and by 0.2% of IoU for the specific building class. This improvement cannot justify the use of a heavy disparity estimation network in our case. However we could search for lighter solutions among disparity estimation networks.

2.4 Disparity estimation networks

The majority of networks that process multi-view images are intended for disparity estimation and some of them could be adapted to semantic segmentation. Disparity estimation networks can fall into two categories. The simplest ones process the image directly using 2D convolutions, like DispNetSimple (DispNetS) or DispNetCorr (DispNetC) (Mayer et al., 2016). However the best performances are obtained by another type of networks that process the cost volume using 3D convolutions, like GC-Net (Kendall et al., 2017), PSMNet (Chang and Chen, 2018) and GA-Net (Yang et al., 2019). The networks using volumetric methods gives very good results on classical disparity estimation benchmarks like Middlebury, but are computationally expensive and not easily adaptable to segmentation tasks. On the contrary, networks like DispNetS and DispNetC are very similar to U-Net so their architectures are ready to be used for semantic segmentation.

Studies have shown that disparity estimation networks can be efficient in remote sensing : (Xia et al., 2020) has obtained good performances with GA-Net (Yang et al., 2019) over satellite images even from a network only trained on synthetic data.

To our knowledge networks like DispNetS or DispNetC have not been tested on stereo satellite images. We will use these networks for our study because they seem adapted for both semantic segmentation and stereo processing.

3. PROPOSED METHODOLOGY

3.1 Dataset

For the training of our stereo networks, we need stereo images in epipolar geometry. We process stereo-rectification on images coming from different sources.

The first product used is a stereo image of Montpellier in France acquired by Pléiades satellite (De Lussy et al., 2005), with a panchromatic band (480-830nm) at 50cm resolution and 4 spectral bands (Red, Green, Blue, NIR) at 2m resolution. The height of the camera is 695 kilometers with respect to ground. The B/H ratio is 0.375, the elevation angle is 77° for the left image and 75° for the right image. The estimated ground sampling distance is 0.517m horizontally and 0.510m vertically on the raw panchromatic image. The image has been pansharpened and only the first 3 bands have been used (RGB). The image used has a size of 17000×22000 pixels, which represents a 93 square kilometers area centered around the middle of the city.

Figure 1. Pléiades (PHR) image of Montpellier with the associated building labels

We use OpenStreetMap labels as ground truth for the images. We follow the method described in (Cournet et al., 2020) to transform both the stereo pair and the labels into epipolar geometry. The images are stereo-rectified with the software Chaîne Automatique de Restitution Stéréographique (CARS) (Michel et al., 2020). Then direct localization is performed on the labels to transform them from orthographic geometry to epipolar geometry. The labels obtained correspond to the roofs of the buildings. The parts of the roofs occluded by trees are still classified as roofs. An example image with building labels is presented at

figure 1 and the left diagram at figure 3 summarizes the pipeline employed to produce the images.

In addition to Montpellier dataset, we add images from US3D (Foster et al., 2020). US3D is the dataset used for the 2019 Data Fusion Contest, but has been complemented in 2020 with additional images from Jacksonville and Omaha, along with new images of Atlanta from the SpaceNet4 (Weir et al., 2019) contest. The images of Jacksonville and Omaha are acquired by WorldView-3 and have a resolution of 31cm. The camera height is 620 kilometers. The images of Atlanta are acquired by WorldView-2 and have a resolution of 50cm. The camera height is 772km. The B/H ratio ranges from 0.08 to 0.55 as there is many points of view available from the dataset.

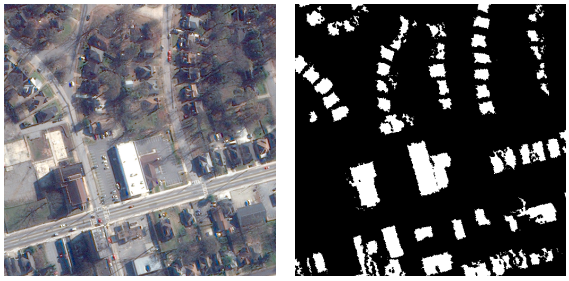


Figure 2. WorldView-2 image of Atlanta from the US3D dataset with the associated building labels

These images were already associated with building labels and in particular the roof masks. The method employed to produce the classification is unknown to us. The parts of buildings occluded by trees are not classified as roof, but taking this into account seems to have produced noisy labels as we can see figure 2. This dataset is still valuable for the training despite the poor quality of the classification. The images were not rectified into epipolar geometry on the official website of the dataset [1] so we process stereo rectification on these images with CARS to transform them into epipolar geometry. Label maps included in the dataset were in the same geometry as the images so we apply the same rectification grids to transform them into epipolar geometry.

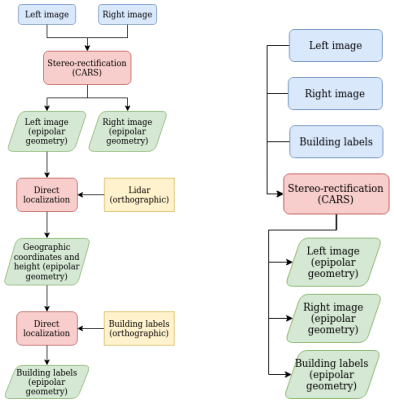


Figure 3. Pipelines used for the registration of building labels against satellite images : PHR image (left) and US3D dataset (right)

In the end, all the images are cropped into patches of size 512×512 pixels, without overlap. The number of patches are summarized in Table 1

[1] https://ieee-dataport.org/open-access/urban-semantic-3d-dataset

City	Number of patches	Area covered (km^2)
Atlanta	1700	111.4
Jacksonville	700	8.8
Omaha	700	8.8
Montpellier	1300	85.2

Table 1. Quantity of data by city

The area covered on Jacksonville and Omaha is very small but the resolution of the images is higher and there is two stereo images from each part of the ground (with different point of views). It explains that the number of patches is still balanced compared to Atlanta or Montpellier.

We divide the dataset into a training set of 4000 patches and a validation set of 400 patches. No area covered in the training set is present on the validation set. The test dataset will be composed of images from different cities : Toulouse (France) and London (UK).

3.2 Stereo network

Among the networks presented in section 2, we chose the DispNet neural networks (Mayer et al., 2016) for our task of stereo segmentation. DispNet is the simplest and lightest network architecture that permits to work with stereo images. It shows good results for the task of disparity estimation. Even if the network has not been tested on segmentation tasks, its architecture is very close to segmentation networks so the network can easily be adapted for our problem.

DispNetS and DispNetC are direct adaptation of optical flow estimation networks FlowNetSimple and FlowNetCorr (Dosovitskiy et al., 2015) whose architectures are represented on figure 4.

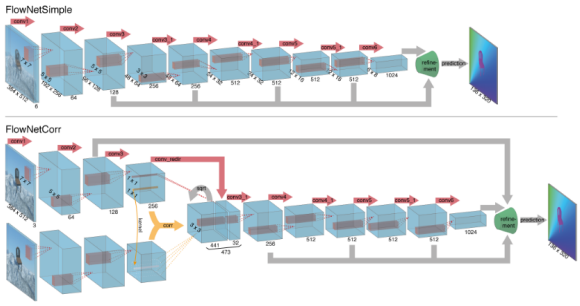


Figure 4. FlowNet architectures (Dosovitskiy et al., 2015)

At the time of publication of FlowNet and DispNet networks, residual connections and residual blocks (He et al., 2016) were not used or at least not common in convolutional neural networks architectures. FADNet (Wang et al., 2020) is a recent network using DispNetS and DispNetC architecure to build a more efficient network. The architecture of the networks are globally the same but each standard convolutional layer is replaced by two residual blocks. Residual connections permit to make the networks deeper and improve the performances. We use the source code of DispNetS and DispNetC PyTorch modules of the git repository of FADNet for our work [2].

DispNetS is a simple U-Net for which the two images I_L (left image) and I_R (right image) of the stereo pair in epipolar geometry are concatenated along the channel axis at the input of the

[2] https://github.com/HKBU-HPML/FADNet

network. With $I_L = (R_L, G_L, B_L)$ and $I_R = (R_R, G_R, B_R)$, the input stereo image is $I = (R_L, G_L, B_L, R_R, G_R, B_R)$.

DispNetC processes feature extraction on the two images separately with the same convolutional encoder E to produce two distinct feature maps $\mathbf{f} = E(I_L)$ and $\mathbf{g} = E(I_R)$ in $\mathbb{R}^{H \times W \times F}$. Then a correlation layer fuses the results between the left feature map $\mathbf{f}$ and the right feature map $\mathbf{g}$ shifted from $\mathbf{f}$. With $D = d_{max} - d_{min}$ the range of possible shifts between $\mathbf{f}$ and $\mathbf{g}$, the correlation volume $\mathbf{C} \in \mathbb{R}^{H \times W \times D}$ is defined as :

$$\mathbf{C}_{ijk} = \sum_{h=0}^{F} \mathbf{f}_{ijh} \mathbf{g}_{i(j+k+d_{min})h} \tag{1}$$

The correlation volume produced is then processed by 2D convolutions which directly squeezes the disparity dimension. In the DispNetC version of the FADNet article (Wang et al., 2020) and git repository, the dimensions (H, W) of the feature maps at the correlation stage are $1/8$ of the original image (as shown on figure 4). Applied to a satellite image acquired by Pléiades, it gives feature maps of 4 meters resolution. Small houses could easily be lost at this level of downsampling, so it can be relevant to process the correlation volume C at a higher resolution ($1/4$ or $1/2$).

Also, the disparity range D chosen in (Wang et al., 2020) is 20 pixels at $1/8$ resolution, which covers disparities up to $20 \times 8 = 160$ pixels at the original resolution. This choice was relevant because the dataset used in (Wang et al., 2020) can have disparities up to 192 pixels. However satellite images have a smaller disparity range and disparities over 30 pixels are quite rare. We prevent disparities higher than 60 pixels on our version of DispNetC because increasing too much the disparity range and the volume depth for just a few high buildings could decrease the global performances of the network as (Wang et al., 2020) shows. We also have to deal with both positive and negative disparities in our dataset contrary to (Wang et al., 2020).

The networks were designed for disparity estimation but it is not difficult to adapt the architecture for segmentation. The only things we change is the activation function and the loss function. The activation function was a ReLU (Rectified Linear Unit) and we replace it by a sigmoid function in order to get results between 0 and 1 (the probability of each pixel to represent the roof of a building). The loss function was a L1 loss for disparity, and we replace it by the sum of Jaccard distance J and binary crossentropy BCE. Let $A \subset \mathbb{N}^2$ be an array of the size $N \times N$ of an input image, $pr : A \to [0, 1]$ the output of the network and $gt : A \to \{0, 1\}$ the ground truth of the building labels, the loss function is defined as :

$$J(pr, gt) = 1 - \frac{\sum_A pr \times gt}{\sum_A pr + gt - pr \times gt} \tag{2}$$

$$BCE(pr, gt) = -\frac{1}{N^2} \sum_A gt \log(pr) + (1 - gt) \log(1 - pr) \tag{3}$$

$$loss(pr, gt) = J(pr, gt) + BCE(pr, gt) \tag{4}$$

With these modifications the network is ready to be trained for segmentation. Training has been done with PyTorch framework. The optimizer used is Adam, the learning rate starts at 4×10^{-4} and decreases linearly until 0 during 30 epochs. The network is trained by batch of 8 images. All these hyperparameters have been manually fine-tuned. The weights of the network are randomly initialized. The weights retained at the end of training are not necessarily the ones of the last epoch but the ones of the epoch that has produced the best results on the validation dataset. Instance normalization is used as preprocessing of images in order to improve generalization abilities.

3.3 Validation method

The main concern of this study is about the advantage of stereo processing over monocular processing. The problem about comparing a monocular network trained with monocular images and a stereo network trained with stereo images is that we cannot be sure about whether the difference observed is due to the contribution of stereo information or caused by differences in the architecture of the network (number of parameters or number of layers). Therefore we decide to compare stereo training with monocular training on the exact same stereo architectures : DispNetS and DispNetC. To perform monocular training on a stereo architecture, the left image of each stereo pair is duplicated to form a fake stereo pair with twice the same image. These fake stereo pairs can be used as input of DispNetS or DispNetC to perform monocular training.

A real monocular network is still tested along with stereo networks ; it is the DispNetS network with 3 channels as input instead of 6, which is a simple U-Net network.

4. EXPERIMENTS

4.1 Results on validation dataset

Five trainings have been made. Both DispNetS and DispNetC have been trained once with duplicated monocular images and once with real stereo pairs. The left images of the two training datasets are exactly the same. U-Net have been trained with simple monocular images. The results of these trainings on the validation dataset of 400 images from Atlanta, Jacksonville, Omaha and Montpellier are given in the table 2. The metric used is IoU (Intersection over Union) which is the intersection of predicted buildings and ground truth divided by the union of predicted buildings and ground truth.

Network	Training	IoU (%)	Loss
U-Net	Monocular	64.5	0.513
DispNetS	Monocular	63.8	0.515
	Stereo	67.9	0.449
DispNetC	Monocular	64.1	0.513
	Stereo	**69.8**	**0.432**

Table 2. Metrics comparison between monocular and stereo training on validation dataset

The advantage of stereo networks on the validation dataset is clear. It may be more relevant to compare the results according to the loss metrics because it is the one that is optimized. We observe that the three monocular trainings have given very close scores regarding the loss. Stereo trainings show significant improvements regarding the IoU or the loss: stereo improved the loss by 12% for DispNetS and by 15% for DispNetC. The advantage of DipNetC over DispNetS was expected because of

the better ability of DispNetC to perform stereo matching, but DispNetS reaches good results which shows it manages to extract and use stereo information without any module dedicated to it.

4.2 Results on Toulouse dataset

To test the generalization abilities of our network, we begin by using an image quite similar to the ones of the training dataset : a Pléiades (PHR) image of Toulouse. The B/H ratio of the stereo image is 0.395 which is close to the Montpellier image. The image has been pansharpened and stereo-rectified by CARS as the Montpellier one from the training dataset. Both DispNetC networks have been tested on Toulouse images : the network trained with monocular images is tested with monocular images from Toulouse, and the network trained with stereo pairs is tested with stereo pairs of Toulouse. In this section the evaluation is only qualitative because the ground truth is not precisely calibrated against the image.

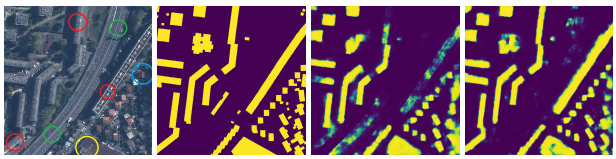


Figure 5. Advantage of stereo training for DispNetC network over an image of Toulouse (1). From left to right : image, ground truth, monocular prediction and stereo prediction

For the patch shown figure 5, the monocular network misses 3 buildings (red circles) or at least has low confidence about them, while the stereo network detects them well. Both networks get false positives on the elevated road but at different locations (green circles). Both also have difficulties labeling the buildings on the right of the image that are shadowed (blue circle). Little buildings are slightly better labeled by the stereo network. By observing attentively, we can note false positives on the bottom building (yellow circle) for the stereo inference. It may be caused by changes of disparity on top of this building that disturb the stereo network.

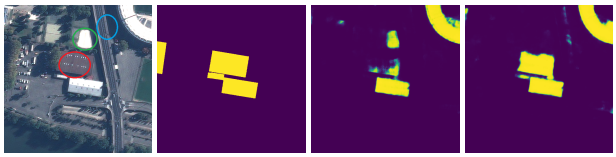


Figure 6. Advantage of stereo training for DispNetC network over an image of Toulouse (2). From left to right : image, ground truth, monocular prediction and stereo prediction

The second patch shown figure 6 shows a building (red circle) detected by the stereo network but not by the monocular network. The white zone on top-middle of the image is a volleyball court (green circle) and is labeled as building by the monocular network. These two differences show a clear advantage for the stereo network on difficult cases of building detection and supports the underlying idea that stereo information gives effective clues about the presence or absence of buildings.

However the stereo network gives a few false positives on the elevated road (blue circle). It shows that despite a global improvement given by stereo, a stereo network could have limits and some disadvantages.

Both monocular and stereo networks label the football stadium on the top-right corner of the image as a building while it is not labeled as building by OpenStreetMap. It shows the ambiguity about the definition of building.

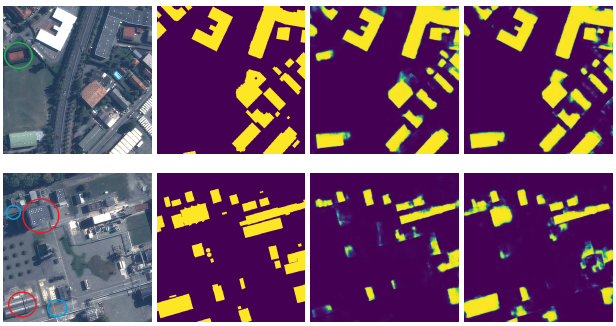


Figure 7. Advantage of stereo training for DispNetC network over other images of Toulouse (3, 4). From left to right : image, ground truth, monocular prediction and stereo prediction

The first image of figure 7 shows a tennis court (green circle) labeled as building by the monocular network which is avoided by the stereo network. On the second image the color of the buildings is very close to the color of the ground which makes buildings very difficult to detect for both networks, but the stereo network manages to detect two big buildings (red circles) not recognized by the monocular network. However the stereo network also shows some false positives (blue circles) not present for the monocular result.

The analysis of these images permits to better understand the quantitative results given on the validation dataset at the last section because it shows which cases give an advantage to the stereo network. It also validates the efficiency of our stereo network over an unknown city, even if the environment is close to the one of training.

4.3 Results on London dataset

Our networks have then been tested on a Pléiades (PHR) image of London. This test dataset is different from the training dataset. The network has never seen any image from United Kingdom and the buildings architecture there is different from the one of France or United States. More importantly, the azimuth viewing angle is reversed from the images in the training dataset. It means the shadows of the buildings are not in the same direction between the training images (at the left of buildings) and testing images (at the right of buildings). It could significantly increase generalization difficulties especially for monocular network that seems to use shadows to estimate the elevation of buildings. The B/H ratio of the London stereo image is 0.388 which is similar to the PHR image of Montpellier in the training dataset and shouldn't cause generalization difficulties linked to stereo configuration.

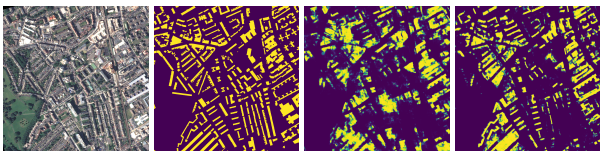


Figure 8. Advantage of stereo training for DispNetC network over an image of London (1). From left to right : image, ground truth, monocular prediction and stereo prediction

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-2-2022-299-2022 | © Author(s) 2022. CC BY 4.0 License.

The results on figure 8 show an expected lack of generalization for the monocular network. Besides, stereo training gives much better results than monocular training. The difference between the two trainings is much more significant for London than for Toulouse, which could mean that the main advantage of stereo matching is the generalization ability it permits. The improved generalization ability of stereo networks could be explained : defining what is a building according to its height above ground is more relevant and more generalizable than characterizing it according to radiometric information. The color of buildings varies according to the city, and spectral information also depends on the acquisition method of the image.

More results on London are shown below. We can notice some buildings are missing from the OpenStreetMap labels on the second ground truth.

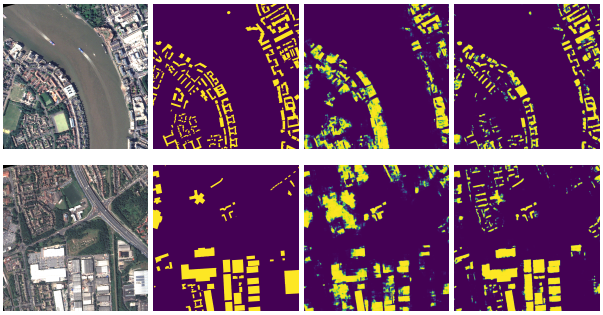


Figure 9. Advantage of stereo training for DispNetC network over other images of London (2, 3). From left to right : image, ground truth, monocular prediction and stereo prediction

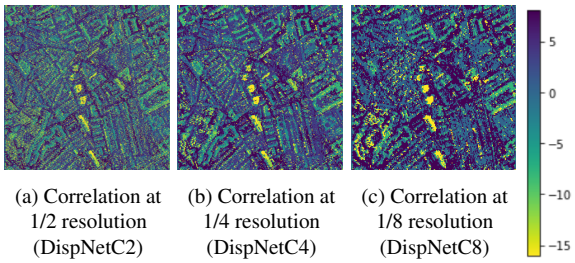
These results show a significant advantage for stereo training on this particular network, with this particular training and testing dataset.

5. DETAILS ANALYSIS

This section presents our attempts to understand the advantages of stereo segmentation over monocular segmentation. As explained in the section 3.2, the DispNetC network uses a correlation layer to manually perform a stereo matching of the image, unlike DispNetS which extracts the stereo clues by direct regression only. DispNetC shows slightly better results and is easier to study thanks to the explicit construction of the cost volume. We analyse the quality of the stereo matching performed by DispNetC and test how much the network use it. We compare DispNetS to DispNetC.

5.1 Analysis of the cost volumes produced by DispNetC

As explained in section 3.2, the correlation layer follows several downsampling layers. In the original paper of DispNet the authors used 3 downsampling layers which gives an image of 1/8 of the original resolution as input of the correlation layer. We can study the effect of the downsampling level on the quality of the cost volume. It is difficult to visualize the cost volume because its representation is three-dimensional. Thus, we decide to show the disparity map induced by this cost volume instead of the cost volume itself. We compute the disparity map from the cost volume by applying a max function over its disparity axis between -16 and 8 pixels (which are the usual disparities for the image used). The disparity maps are shown on figure 10. The image taken as example for the visualisation of cost volumes is the same as the one shown on figure 8.



(a) Correlation at 1/2 resolution (DispNetC2)
(b) Correlation at 1/4 resolution (DispNetC4)
(c) Correlation at 1/8 resolution (DispNetC8)

Figure 10. Effect of downsampling on the cost volume / disparity map of DispNetC

The original network DispNetC8 misses many details during the correlation as we can see on the figure 10, and many small buildings are completely lost on the cost volume. The network has no way to retrieve disparity features other than with this cost volume. It means the buildings that are too small to be taken into account in the cost volume at 1/8 resolution have no stereo information associated to them, which makes the use of stereo images useless for these buildings.

Besides, at 1/4 resolution many buildings appear on the disparity map and borders are better defined. The quality of the disparity map seems acceptable despite a little bit of noise, which means that the cost volume could be efficiently used by the network to induce a building segmentation even on small buildings.

At 1/2 resolution some borders are finer but no other buildings appear and the resulting disparity map is very noisy. Moreover, another problem of DispNetC2 is its runtime : most of the time is taken by the correlation layer because the number of operations needed for the correlation is multiplied by 8 over DispNetC4 and by 64 over DispNetC8 (the complexity is $\mathcal{O}(n^3)$ with n the width of the square image expressed in pixels).

5.2 Performance of stereo networks deprived of stereo information

The goal of this experience is to evaluate how much the stereo networks (trained with stereo images) use the right image and the stereo information induced by stereo for inference.

Three networks have been used : DispNetS, DispNetC4 and DispNetC8. They have been trained on the usual training dataset with stereo images and tested on the Montpellier dataset. For each network, two tests have been made : one with the classical stereo pair (left and right images), and one with two left images. Results are shown on table where the metric used is Intersection over Union.

Network	Left-Right IoU	Left-Left IoU
DispNetS	75.1	47.4
DispNetC8	74.8	50.4
DispNetC4	75.8	35.7

Table 3. Performance of stereo networks with and without the right image

As expected, stereo networks trained with stereo images give poor results when tested on fake stereo image (Left-Left) compared to real stereo images (Left-Right). It means that these networks use the right image and the depth-related information extracted from it to perform segmentation.

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-2-2022-299-2022 | © Author(s) 2022. CC BY 4.0 License.

The differences between DispNetS and DispNetC8 are too small to be interpreted, but the results of DispNetC4 are very interesting. It is the network that has the best IoU while not being very far from the scores of DispNetS. But more significantly, it is the network that has the worse results when deprived of the right image. It could mean than this network relies more on stereo clues than the other networks to perform the segmentation. This interpretation is supported by the disparity maps shown on figure 10 : DispNetC4 has a better cost volume so it is relevant that it uses more this cost volume than spectral information of the left image to infer the segmentation.

5.3 Effect of shifting left and right images on performance of stereo networks

This experiment is of the same kind of the last one and aims at analysing the behaviour of the networks when the stereo configuration is altered.

Stereo networks seems to generalize better than monocular networks as shown in section 4.3. The hypothesis made is that stereo networks rely more on disparity than radiometry which is more robust to new images with different environments. But the risk of our stereo networks is that it associates the presence of buildings with a simple range of disparities (between 3 and 8 pixels for example).

Variation on the B/H ratio on a new image could alter the disparities associated to buildings : a B/H twice as big would transform the range of disparities from $[3, 8]$ to $[6, 16]$ for the same building. Unfortunately the Pléiades images we used have similar B/H ratios (between 0.37 and 0.40), so the influence of the B/H ratio on the results have not been tested.

However a difference of disparity for similar buildings can occur without variation of B/H ratio. In case of undulating terrain, the ground elevation would vary over a single image and so the disparity associated to the ground, which would cause shifts of the disparities associated to buildings (for example from $[3, 8]$ to $[6, 11]$). Our networks need to work despite these shifts.

To test the consistence of the networks, horizontal constant shifts have been made between left and right images to simulate constant altitude change. By doing that, the disparity associated to buildings varies according to the shift. The results are on the figure 11.

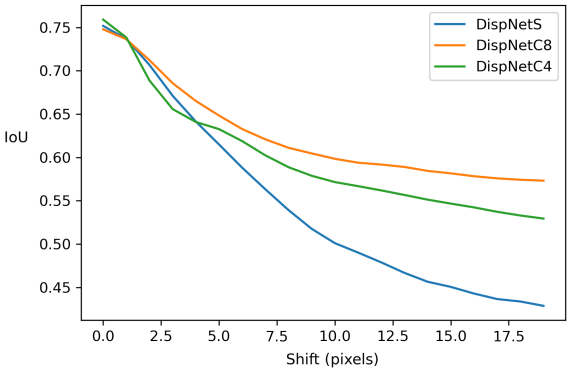


Figure 11. Effect of shift between left and right images upon inference score

This time, it is the DispNetS network that has the worse results when the shift is above 5 pixels. Once again it can be explained : DispNetS has no specific module for computing disparity and rely exclusively on convolutions to do so. Convolutions kernels are 7 by 7 on the first layer and then 3 by 3, so it does not allow the network to detect large disparities, whereas DispNetC could acknowledge every disparity level within its disparity range $[d_{min}, d_{max}]$ defined by the user (in our case the disparity range is $[-60, 60]$).

DispNetC4 has lower scores than DispNetC8 even with small shifts (between 2 and 5 pixels). It is the sign of a lack of generalisation ability of DispNetC4 for changes on ground elevation or building heights. We know DispNetC4 rely more on disparity than DispNetC8, which could mean that relying too much on disparity rather than radiometry to infer segmentation could lead to these generalization problems.

6. CONCLUSION

This study has shown that satellite stereo images could be used efficiently to perform semantic segmentation of buildings. In the specific context of this study, stereo networks generalized better than monocular networks when tested on cities different from the ones of the training dataset. Our explanation is that stereo networks can characterize a building according to its elevation instead of its color or shadow, which is a more universal feature. However we saw stereo networks can be sensitive to disparity shifts which can add new generalisation problems associated to the epipolar geometry.

To conclude, stereo segmentation is able to outperform monocular segmentation because it uses more relevant features to detect buildings. We believe stereo segmentation is a relevant choice for building segmentation when stereo images are available, especially in the context of 3D modeling.

Very few studies have been conducted on stereo segmentation and we hope these results will inspire new works on this subject. Stereo segmentation may be very efficient but its context of utilization have to be delimited with other studies. Researches on this subject could lead to a significant step in the area of computer vision and remote sensing.

ACKNOWLEDGEMENTS

We would like to thank the Centre National d'Etudes Spatiales (CNES) which provided PHR satellite images and an access to its HPC cluster with GPU cards to train the networks.

REFERENCES

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 20–32.

Baillarin, S., Brunet, P.-M., Lassalle, P., Souille, G., Gabet, L., Foulon, G., Romeyer, G., Ferrero, C., Huynh, T.-L., Masse, A. et al., 2020. Ai4geo: An automatic 3d geospatial information capability. *EGU General Assembly Conference Abstracts*, 11559.

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.

Chang, J.-R., Chen, Y.-S., 2018. Pyramid stereo matching network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, H., Lin, M., Zhang, H., Yang, G., Xia, G.-S., Zheng, X., Zhang, L., 2019. Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 4967–4970.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

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 Lussy, F., Kubik, P., Greslou, D., Pascal, V., Gigord, P., Cantou, J. P., 2005. Pleiades-hr image system products and quality-pleiades-hr image system products and geometric accuracy. *Proceedings of the International Society for Photogrammetry and Remote Sensing Workshop, Hannover, Germany*, 1720, Citeseer.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks. *Proceedings of the IEEE international conference on computer vision*, 2758–2766.

Durner, M., Boerdijk, W., Sundermeyer, M., Friedl, W., Marton, Z.-C., Triebel, R., 2021. Unknown Object Segmentation from Stereo Images. *arXiv preprint arXiv:2103.06796*.

Foster, K., Christie, G., Brown, M., 2020. Urban semantic 3d dataset.

Hao, S., Zhou, Y., Guo, Y., 2020. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406, 302–321.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J., Li, L., Lin, Y., Wu, F., Zhao, J., 2019. A comparison and strategy of semantic segmentation on remote sensing images. *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, Springer, 21–29.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-end learning of geometry and context for deep stereo regression. *Proceedings of the IEEE International Conference on Computer Vision*, 66–75.

Lebègue, L., Cazala-Hourcade, E., Languille, F., Artigues, S., Melet, O., 2020. CO3D, a worldwide one one-meter accuracy DEM for 2025. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 299–304.

Lei, T., Li, L., Lv, Z., Zhu, M., Du, X., Nandi, A. K., 2021. Multi-modality and multi-scale attention fusion network for land cover classification from VHR remote sensing images. *Remote Sensing*, 13(18), 3771.

Liu, Z., Chen, B., Zhang, A., 2020. Building segmentation from satellite imagery using u-net with resnet encoder. *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, IEEE, 1967–1971.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, 3226–3229.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T., 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4040–4048.

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.

Mortazi, A., Karim, R., Rhode, K., Burt, J., Bagci, U., 2017. Cardiacnet: Segmentation of left atrium and proximal pulmonary veins from mri using multi-view cnn. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 377–385.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Wang, Q., Shi, S., Zheng, S., Zhao, K., Chu, X., 2020. Fadnet: A fast and accurate network for disparity estimation. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 101–107.

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A. V., McPherson, S., Shermeyer, J., Kumar, V., Tang, H., 2019. Spacenet mvoi: A multi-view overhead imagery dataset. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 992–1001.

Xia, Y., d'Angelo, P., Tian, J., Reinartz, P., 2020. Dense matching comparison between classical and deep learning based algorithms for remote sensing data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43(B2), 521–525.

Yang, G., Manela, J., Happold, M., Ramanan, D., 2019. Hierarchical deep stereo matching on high-resolution images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5515–5524.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.