EXPERIMENT ON PRODUCING DISPARITY MAPS FROM AERIAL STEREO IMAGES USING UNSUPERVISED AND SUPERVISED METHODS

L. Zhu[1]*, E. Hattula[1], J. Raninen[1], J. Hyyppä[2]

[1] National Land Survey of Finland (NLS) - lingli.zhu@nls.fi, (emilia.hattula, jere.raninen)@maanmittauslaitos.fi
[2] Finnish Geospatial Research Institute FGI, National Land Survey of Finland (NLS) - juha.hyyppa@nls.fi

Commission IV, WG IV/4

KEY WORDS: aerial images, disparity map, deep learning, OpenCV, supervised machine learning, unsupervised machine learning.

ABSTRACT:

Recent advancement in hardware and software provides the possibility of realizing full automation in stereo-image tasks. This paper investigated disparity map generation from aerial images with different methods: unsupervised method and supervised methods. The datasets were from aerial stereo dense matching benchmark dataset for deep learning in ISPRS 2021: Vaihingen dataset and the WHU MVS/Stereo Dataset released in the CVPR 2020. Two neural networks: GC-net and PSMnet have been trained with the Vaihingen dataset and the WHU MVS/Stereo Dataset. With unsupervised methods, stereo block matching(StereoBM) and Stereo Semi-Global Matching (StereoSGM) methods from the OpenCV were studied. We selected seven image pairs from the Vaihingen dataset and six image pairs from the WHU dataset for testing and evaluation. Difficulty scenes such as textureless areas, reflective surfaces, and repetitive patterns were also included in our study. The performance from different methods was compared by both visualization and quantitative means. The advantages and disadvantages are presented.

1. INTRODUCTION

Although study on the generation of disparity maps has been for decades, from the latest statistics of Web of Science (WebOfScience, 2022), the research on disparity maps is still increasing yearly, especially in the field of artificial intelligence (AI). With recent advances in AI, machines are gaining the ability to learn, improve, and execute repetitive tasks precisely, especially with deep learning techniques. More and more researchers started to explore the new methods in the fields of deep learning to produce disparity maps. The advantages of utilizing the deep learning technique lie in its intelligence, meaning that the learning process is done by the machine. Without demanding professional knowledge from humans, solutions can be obtained. As the numerous research on deep learning techniques for stereo images in recent years, the advantages and disadvantages of the traditional and the new technology for stereo image tasks should be investigated.

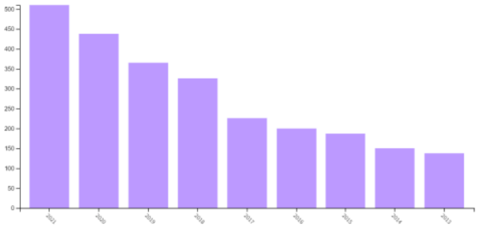

Figure 1. Statistics of publications on the topic 'disparity map' in Web of Science. Horizontal axis: publication year; Vertical axis: number of publications with the topic 'disparity map'.

What is a disparity map? A disparity map shows the pixel locational differences between the left and the right images when a 3D scene is projected perspectively on the stereo images. It presents the relative depth while a depth map demonstrates the absolute depth by giving information about the distance between the lenses of the stereo camera and the focal length of the camera. The conventional methods for disparity map generation were primarily focused on the unsupervised methods such as area-based matching methods (Bracewell, 1965; Briechle & Hanebeck, 2001) and feature-based matching methods (Schmid & Mohr, 1997; Baumberg, 2000; Caspi et al., 2006), in which professional knowledge of stereo geometry and image matching, and a deep understanding of terms in the field were required. For example, the knowledge of the epipolar geometry, epipolar planes, and epipolar lines in stereo images, is needed. In an epipolar geometry, when a stereo camera takes images of a 3D scene from two distinct positions, the projection of a 3D scene on two images is constrained. The relations between the projected points on each image (left or right image) follow the rules: the image point, the observed 3D point, and the perspective center of the camera are aligned. An epipolar plane consists of three points: the perspective centers (the optical center of a camera) of a stereo camera and the observed point from a 3D scene. An epipolar line is the intersected line between the epipolar plane and the left or right image plane. The steps in traditional methods for producing a disparity map include i) image rectification; ii) resampling epipolar lines; iii) estimating the disparity map. Optionally, the users manually select at least five pairs of corresponding points from stereo images to estimate the parameters of rotation, scale, and translation. After image matching, a disparity map can be computed. The difficulties in producing a disparity map might rise in the following context: i) illumination inconsistency and noise presentation on images; ii) inconsistency of specular reflection on images; iii) perspective distortion and perspective

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-561-2022

inconsistency when cameras are close to the objects; iv) textureless surfaces; v) transparent objects; vi) repeating textures on surfaces; vii) occlusion and depth discontinuity. In the condition of aerial images, case iii) rarely happens.

The core of disparity map generation with an unsupervised method is to find the corresponding points between the left and right images. That is image matching. In 2005, Hirschmüller (2005) proposed the semi-global matching (SGM) method for image matching. The algorithm has been utilized in various applications such as aerial image matching to driver assistance systems. It supports pixel-wise matching for maintaining sharp object boundaries and fine structures and can be implemented efficiently on different computation hardware (Hirschmüller, 2005). Later on, the algorithm has been developed further to utilize mutual information (Hirschmüller, 2008) and increase memory efficiency (Hirschmüller et al., 2012).

OpenCV (Open Source Computer Vision Library) was built in 2008. It is an open-source library that includes several hundreds of computer vision algorithms (OpenCV, 2022). There are functions in OpenCV to support depth estimation from stereo images. These functions include epipolar geometry estimation and constraint, and different stereo image matching algorithms (including the SGM method) for stereo reconstruction.

In recent years, using deep learning methods for obtaining depth maps from stereo images has been highlighted. Deep learning models have been successful in learning representations directly from the raw data and are effective for understanding semantics (Kendall et al., 2017). The learning-based method can introduce global semantic information such as specular and reflective priors for more robust matching (Yao et al.,2018). The rich training data help in dealing with the difficulty of matching ambiguity caused by occlusion, varying lighting conditions, or textureless regions (Wang, et al. 2021). Laga et al. (2020) gave a comprehensive review of deep learning methods for stereo depth estimation. The authors categorized deep learning methods into two groups: one is the traditional stereo-matching technique, and another is an end-to-end trainable framework. The traditional method is composed of three modules: a feature extraction module, a feature matching, and cost aggregation module, and a disparity/depth estimation module. Each module is trained independently from the others. The end-to-end trainable methods were grouped into two types. One type is to tackle the regression problem with a large amount of training data, another type is to break the traditional pipeline into differentiable blocks. GC-Net, HRS Net, MVSNet, PMS Net, and PLUMENet are examples of convolutional neural networks (CNNs) with an end-to-end trainable framework.

GC-Net was introduced in 2017 by Kendall et al.. It reasons about geometry by forming a fully differentiable cost volume and incorporates context from the data with a 3-D convolutional architecture to reduce the mismatch in ambiguous regions so that depth estimation is improved (Kendall et al., 2017). PMSNet was proposed by Chang & Chen (2018). It employed a pyramid stereo matching network and presented a stacked hourglass 3D CNN to extend the regional support of context information in cost volume. Yao et al. (2018) proposed the MVSNet (Multi-View-Stereo network). It inferred a depth map from multi-view images. One reference image and several source images were input into the network. The differentiable homography warping operation encodes camera geometries in the network to build the 3D cost volumes from 2D image features and enables the end-to-end training. It utilized a variance-based metric that maps multiple features into one cost feature in the volume (Yao et al., 2018). HRS Net was presented by Yang et al. in 2019. The authors proposed a hierarchical stereo matching architecture to extract multi-scale features from high-resolution images. Besides, asymmetric augmentation techniques were introduced to increase the amount of training data. The algorithm can be run efficiently in real-time (Yang et al., 2019).

In deep learning methods, some experiment was based on open-source datasets, such as KITTI stereo (KITTI stereo dataset, 2022) and Middlebury stereo (Middlebury stereo datasets, 2022). KITTI stereo datasets were collected from two video cameras mounted on the roof of a car. The Middlebury stereo image pairs were taken in indoor scenes under controlled lighting conditions. Ready aerial stereo image datasets still seem to be quite scarce, but at least some can be found, for example, the stereo image dataset of Vaihingen: Aerial Stereo Dense Matching Benchmark introduced by the ISPRS in 2021 (ISPRS2021 benchmark, 2022), and the synthetic aerial dataset: the WHU dataset (Liu & Ji, 2020).

Our experiment was focused on obtaining disparity maps i) from two sets of aerial images with two algorithms from the OpenCV; ii) from two sets of aerial images with deep neural networks: GC-Net and PSM net; The results from unsupervised methods and supervised methods are evaluated with both visual and quantitative analysis. Their advantages and disadvantages are discussed.

2. MATERIALS

2.1 The WHU MVS/Stereo Dataset

The WHU Stereo Dataset was provided by Wuhan University, China (Liu & Ji, 2020). It is a synthetic aerial dataset for large-scale Earth surface reconstruction. It was generated from a 3D digital surface model produced from thousands of real aerial images and refined by manual editing. The dataset covers an area of 6.7 x 2.2km2 over Meitan county, Guizhou Province in China. The virtual aerial image was taken at 550 m above the ground with 90% heading overlap and 80% side overlap. The ground resolution is 10cm (Liu & Ji, 2020). The dataset contains dense and tall buildings, sparse factories, mountains covered with forests, bare ground, and rivers. The aerial images are 8-bit RGB images and depth/disparity maps are 16-bit.

The WHU Stereo image dataset was divided into training, validation, and test sets, respectively. Each dataset contains both stereo image pairs and corresponding disparity maps with ground truth. The image size in each of the sets was 768×384 pixels. There were 8,316 RGB image pairs in the training set, 1694 image pairs in the test set, and 924 image pairs in the validation set. Fig. 2 shows an example of the WHU dataset.

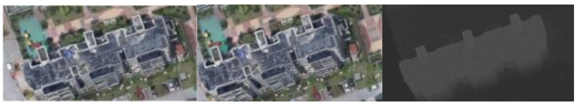

Figure 2. The WHU dataset for training, test, and validation. From left to right: stereo-left image, stereo-right image, and disparity map.

2.2 ISPRS Vaihingen dataset

The ISPRS Vaihingen Aerial Stereo Dense Matching Benchmark 2021 (ISPRS2021 benchmark, 2022): the Vaihingen dataset, was provided by the German Society for Photogrammetry, Remote Sensing, and Geoinformation. The dataset contains aerial images with a depth of 11 bits and a ground sample distance (GSD) of 8 cm. The aerial images were acquired with a fly height of 900m and a focal length of 120mm. Both forward and side overlaps were 60%. It included

RGB stereo image pairs of size 1024×1024 pixels, and same size disparity images for labels (see Fig. 3). Only the original training set of the ISPRS dataset included the reference depth maps. All training, validation, and test sets were formed from the original ISPRS training set. This way, the final training set used training included 449 stereo image pairs and labels, a validation set of 68 image pairs and labels, and a test set of 68 image pairs and labels.
The reference depth maps were produced by Lidar point clouds, consisting of points. They were interpolated twice with a 5×5 average window and once with a 3×3 average window for more consistent disparity image labels that were then used for training, validation, and test metrics.

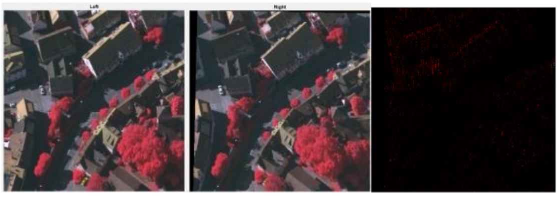

Figure 3. Stereo images and disparity image labels (reference data) from the ISPRS Vaihingen dataset.

3. METHODS

A disparity image for a set of stereo images is defined as an image where each pixel denotes the distance between the pixel in image one to its matching pixel in image two. The core problem in computing a disparity map is to find corresponding points from a stereo image pair. Image matching is essential for disparity map generation. In this experiment, we employed unsupervised methods including stereo block-matching (SteroBM) and stereo semi-global block matching (StereoSGM) for our experiment: OpenCV with Matlab 2022 (Mathsworks, 2022), and supervised methods with end-to-end deep learning frameworks: GC-Net and PSMNet.

3.1 OpenCV

In OpenCV, two image matching methods were provided: StereoBM and StereoSGM. With stereoBM algorithm, parameters of the 'numDisparities' and 'blockSize' should be preset. 'numDisparities' defines the disparity search range. The default minimum disparity is 0 while the maximum disparity needs to be set by the users. 'blockSize' defines the size of an image block. When blockSize =1, it is at pixel level. The matching is implemented pixel by pixel.
With the StereoSGM algorithm, the parameters include 'minDisparity', 'numDisparities', 'SADWindowSize', 'disp12MaxDiff', 'preFilterCap', 'uniquenessRatio', 'speckleWindowSize', 'speckleRange', and 'fullDP'.
SADWindowSize--- Size of block;
disp12MaxDiff ---Maximum allowed difference for disparity check;
preFilterCap---Truncation value for the prefiltered image pixels;
uniquenessRatio---a threshold to filter out unreliable pixels from estimated disparity values
speckleWindowSize --- Maximum size of smooth disparity regions to consider their noise speckles and invalidate.
speckleRange --- Maximum disparity variation within each connected component.

3.1.1 Stereo Block Matching: Since the images have been rectified, the epipolar lines are parallel to the baseline of the stereo camera. The search range is constrained along the epipolar line, which is in one dimension. Assume that a pixel in the left image is located at (x, y), its corresponding pixel in the right image will be (x+d, y), where d is the distance between its location in the left image and the right image. A window in the right image slides along the epipolar line and compares the contents of that window with the reference window in the left image. The matching cost is computed by Sum of Square Distances block-matching (SSD) or Normalized Correlation (NC).

3.1.2 Stereo Semi-Global Block Matching: The original idea of SGM is to perform line optimization along with multiple directions and compute an aggregated cost by summing the costs to reach pixel p with disparity d from each direction. The number of directions affects the run time of the algorithm. In OpenCV, the class 'StereoSGBM' modified the original SGM algorithm (Hirschmuller, 2008) from i) using five directions as default to reduce memory consuming; ii) employing block instead of pixel matching; iii) Mutual information cost function is not implemented. Instead, a simpler Birchfield-Tomasi sub-pixel metric from Birchfield & Tomasi (1998) is used. iv) some pre- and post-processing steps from K. Konolige's algorithm (2010) are included, for example, pre-filtering using the sobel filter and post-filtering employing uniqueness check, quadratic interpolation, and speckle filtering (OpenCV, 2022).

3.2 GC-Net

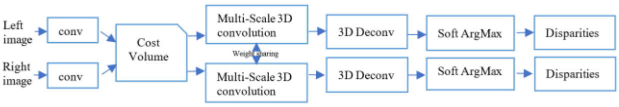

Figure 4. Architecture overview of GC-Net (Kendall et al., 2017).

Geometry and Context Network (GC-Net) is a deep learning architecture introduced by Kendall et al. in 2017 for the estimation of 3D geometry from stereo images (Kendall et al., 2017). Fig. 4 shows the network architecture. The left and right images were fed to the network with a number of 2-D convolutional operations. Each of the convolutional layers is followed by batch normalization and a rectified linear non-linearity. After learning a deep representation, the stereo matching cost was computed through 2-D convolutional operations. 5×5 convolutional filter is then applied with the stride of two for subsampling the input, after which eight residual blocks are appended. The residual blocks consist of two 3×3 convolutional filters in series. Parameters between the left and right towers of the GC-Net are shared to effectively learn corresponding features (Kendall et al., 2017). The highlighted parts in this network architecture are i) incorporating context directly from the data; ii) employing 3-D convolutions to filter the cost volume; iii) using a differentiable soft argmin function to regress sub-pixel disparity. The network architecture was implemented with PyTorch (PyTorch, 2022).
The model was trained using the absolute error between the ground truth disparity d_n, and the model's predicted disparity d_{pn} for pixel 'n'. 'N' is the number of pixels. This supervised regression loss is defined (Kendall et al., 2017):

$$Loss = \frac{1}{N}\sum_{n=1}^{N}\left\|d_n - d_{pn}\right\|_1 \qquad [1]$$

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-561-2022 |

3.3 PSM Net

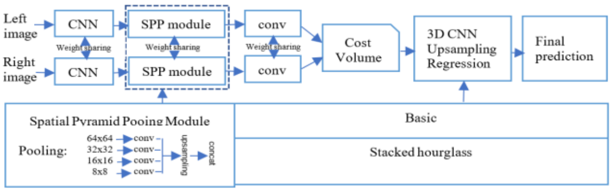

Figure 5. Architecture overview of PSM Net (Chang & Chen, 2018).

The pyramid stereo matching network (PSMNet) was introduced by Chang & Chen in 2018. It is an end-to-end deep learning framework, without postprocessing. Fig. 5 depicts the architecture of the PSMNet. The left and right stereo images are fed to two weight-sharing pipelines consisting of a CNN for feature maps calculation, a spatial pyramid pooling (SPP) module for feature harvesting by concatenating representations from sub-regions with different sizes, and a convolution layer for feature fusion.

The image features are then used to form a 4D cost volume, which is fed into a 3D CNN for cost volume regularization and disparity regression (Chang & Chen, 2018). Different from other deep learning frameworks, it exploited a pyramid pooling module for global context information in stereo matching and presented a stacked hourglass 3D CNN for extending the regional support of context information in cost volume.

Because of the disparity regression, the smooth L1 loss function was adopted to train the PSMNet. The loss function of PSM Net is defined (Chang & Chen, 2018) as

$$L\left(d_n, d_{np}\right) = \frac{1}{N}\sum_{n=1}^{N} smooth_{L1}\left(d_n - d_{np}\right) \qquad [2]$$

Where

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & , if\ |x| < 1 \\ |x| - 0.5, otherwise \end{cases} \qquad [3]$$

3.4 Evaluation methods

We employ the Mean-squared error (MSE) to indicate the accuracy of the disparity maps.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{R_i} - \mathrm{P_i})^2 \qquad [4]$$

Where n --- amount of measured points, i --- the ith point, $\mathrm{R_i}$ --- reference value, $\mathrm{P_i}$ --- predicted value

4. RESULTS AND ANALYSIS

In our experiment, we tested supervised methods and unsupervised methods for disparity map generation with two datasets: the ISPRS Vaihingen and the WHU datasets. The supervised methods include the GC-Net and PSM Net, and the unsupervised methods employ the StereoBM and StereoSGM methods from the OpenCV.

Image rectification has been implemented in both datasets. The images have the size of 1024 x 1024 pixels in the Vaihingen dataset and the size of 384 x768 pixels in the WHU dataset. The test images were selected from multiple scenes including buildings with diverse roof structures, vegetation-covered areas, shadows, reflective areas (water), textureless areas (agricultural field), repetitive pattern textures, and so on. The reference data for evaluation were taken from the validation data provided by the ISPRS and the WHU. OpenCV algorithms were conducted with Matlab 2022. Different parameter settings were tested in both StereoBM and StereoSGM methods so that the best results were obtained. GC-Net and PSM Net models were trained in a supercomputer environment of the CSC, a Finnish company offering ICT solutions (CSC, 2022).

Two GC-Net models were trained with two different datasets. Both models were trained with 256×512 pixel stereo image pairs randomly cropped from the original image pairs, thus parameters W is 512 and H is 256. The maximum disparity used was D=160. The batch size used was 4. The optimizer was the RMSprop with a learning rate of 0.001 and alpha, which is a smoothing constant, was 0.9. The loss used was L1-Loss. Before training, each image was normalized so that the pixel intensities range from −1 to 1. Best models were saved based on the lowest validation losses. Model accuracy was evaluated based on the percentage of disparities with an error of less than 3 pixels. Training accuracy, the percentage of disparities with error less than 3 pixels was 0.988. Testing accuracy was almost as good, 0.983.

The stacked hourglass architecture of PSMNet was implemented using Tensorflow 2.8.0 and Python version 3.7.9. All models were end-to-end trained with Adam (β1 = 0.9, β2 = 0.999). As a data preprocessing, the values of RGB images were normalized between 0 and 1. During training, the images were randomly cropped to size height = 384, width = 512, and randomly flipped vertically and horizontally. The maximum disparity was set to 96 for the WHU dataset and 192 for the ISPRS Vaihingen. The models were trained with a constant learning rate of 0.001. The batch size was set to 4 for the training. The models were trained for 30 epochs with the WHU dataset and 90 epochs with the ISPRS Vaihingen dataset. Training the model for the WHU dataset took about 52 hours and about 15 hours for the ISPRS dataset in the CSC Puhti with one GPU. The WHU model got 0.0253 3-pixel error and 0.38 loss. Vaihingen model got 0.2428 3-pixel error and 1.16 loss.

4.1 With ISPRS Vaihingen dataset

The test images selected from the ISPRS Vaihingen dataset include diverse scenes with challenges in shadows, occlusions, reflection, featureless areas, and area covered with vegetation. With the GC-Net, the best training loss achieved with the dataset was 4.41, while the validation loss was 2.27. The testing loss was 3.70. High training loss can be accounted to the training, validation, and test set sizes, as the training set was notably larger than the two other sets. Training accuracy, the percentage of disparities with error less than 3 pixels, was 0.738. Testing accuracy was even better, 0.796. From the loss curves in the left image of Fig. 5 can be seen that the validation set included easier samples for the model than the training set resulting in smaller validation loss in comparison to training loss. The validation data is scarce but well represented in the training dataset, leading to the model performing well on it. It can also be noticed that there is more spiking in the training loss. It might be affected by the size of the training data. The smaller the training data are, the less smooth the training loss curve is. However, it is good to consider that spiking in the curves can also be caused by label inaccuracies, as well as the optimizer and the used learning rate. Spiking of the training loss curve can also indicate unrepresentative data in some cases, but as the spiking is not very intense, it can be caused by the dataset size and label inaccuracies as mentioned.

ISPRS Vaihingen PSMNet model used smooth L1-loss with beta = 1.0 and the missing values in the disparity masks were ignored. Both the training loss and validation loss (see the right image of Fig. 6) had many spikes suggesting that a smaller

https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-561-2022 | © Author(s) 2022. CC BY 4.0 License.

learning rate could have helped with the training. Validation loss had two large spikes between 20 and 40 epochs. Validation loss was lower than training loss during most of the epochs. Validation loss achieved the smallest result at epoch 79.

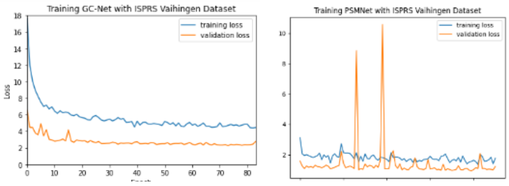

Figure 6. Training loss and validation loss in ISPRS Vaihingen dataset with the GC-Net and PSM Net.

The test results with the ISPRS Vaihingen dataset was shown in Fig. 7. As references, the first two rows are the stereo images. Then the disparity maps were generated by different methods: StereoBM, StereoSGM, GC-Net, and PSM Net. For the StereoBM method, the block size affects the resultant noise level. When a small size block is used, more detailed information can be acquired, but at the same time, more noise is presented. A balance between the details and a cleaned map should be kept. Besides, the disparity range needs to be set properly. In the StereoBM algorithm, the value of the disparity range should be the multiples of 16, while in the StereoSGM algorithm, it should be the multiples of 8 and the value shouldn't be over 128. For the ISPRS Vaihingen datasets, the StereoSGM shows better results than the StereoBM method. The StereoSGM method is more robust and produces smoother buildings and less noise in the results.

Disparity maps from GC-Net and PSM Net show high smoothing buildings. The completeness of the building roofs with supervised methods beats the unsupervised methods. The results from the supervised methods were clean and no holes were presented. However, when we check the results carefully from the supervised methods, it can be seen that there were cut prints shown in each disparity map. The reason is that the size of training data is fixed in both networks. When training data use a size of 256 x 512 in GC-Net, the resultant disparity map is in a size of 1024 x 1024. Thus, the results were combined from multiple images.

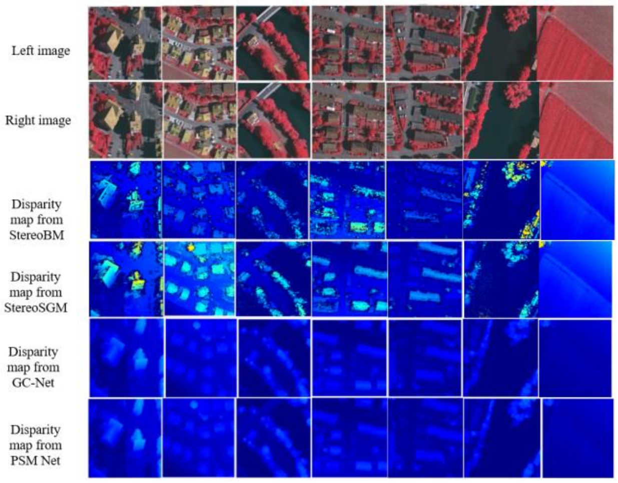
Left image
Right image
Disparity map from StereoBM
Disparity map from StereoSGM
Disparity map from GC-Net
Disparity map from PSM Net

Figure 7. The test results with the Vaihingen stereo images. The corresponding photo no. from left to right: 0003, 0045, 0016, 0031, 0038, 5007, 0024, 0048.

4.2 With the WHU dataset

The selected images from the WHU dataset include more high and complex buildings. Besides, a scene with repeated textures was also selected to test the performance of different methods.

The GC-Net model trained with WHU MVS/Stereo Dataset was trained for 36 epochs, after which the validation loss didn't seem to decrease. The best validation loss was achieved during epoch 32 and the final model was saved. A model trained with the ISPRS Vaihingen stereo image set, on the other hand, was trained for 84 epochs, after which the validation loss did not decrease any more. The lowest validation loss was achieved after epoch 74, and the final model was saved. The ISPRS Vaihingen model needing more epochs for training was due to the dataset's larger image size (1024×1024 pixels) in comparison to the WHU dataset's image size (768×384 pixels), considering the random cropping data augmentation method used (256×512 pixel image crops). With the GC-Net, the best training loss achieved with the dataset was 0.22, while validation loss remained higher, 0.76. The testing loss was 0.35. From the left image of Fig. 8 can be seen that the validation loss curve stays on top of the training loss curve. It also indicates that the WHU dataset's division for training, validation, and test sets seems successful.

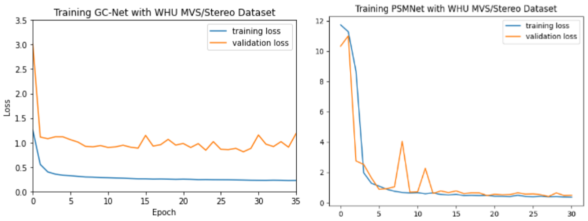

Figure 8. Training loss and validation loss in the WHU dataset with the GC-Net and PSM Net.

The WHU PSMNet model used smooth L1-loss with beta = 1.0. Training loss drops steadily with the training epochs. Validation loss spiked at epochs 9 and 12 but dropped again in the next epoch to a smaller value than before the spikes. Training loss was below validation loss during most of the epochs. Validation loss achieved the smallest results at epoch 27 and started to rise slightly after that. With the PSM Net, the WHU model got 0,023 3-pixel error and 0.38 loss.

From Fig. 9 can be seen that, with high buildings, both StereoBM and StereoSGM performed quite well. Compared to unsupervised methods, the edges of the roofs in the results of supervised methods show clearer and sharper. In the water area, both StereoBM and StereoSGM performed better than the supervised methods. With the WHU dataset, there were smooth, no holes, and less noise presented from unsupervised results. It was better than the results from the ISPRS Vaihingen datasets. In the WHU dataset, the size of images is much smaller than the ISPRS images. A WHU image encompasses fewer objects. It also indicates that unsupervised methods are good for small and simple scenes, while supervised methods are suitable for images containing large and complex scenes. With repeated textures, both supervised and unsupervised perform well. In addition, the cut prints were visible on the supervised results.

Fig. 7 and Fig. 9 demonstrated the results visually. Table 1 shows the results of the quantitative evaluation. The resultant disparity maps were compared to the reference data (given as validation data for supervised methods). The measurement was performed with the mean square error. From the quantitative evaluation results can be seen that, overall, supervised methods showed better quantitative accuracy than unsupervised methods.

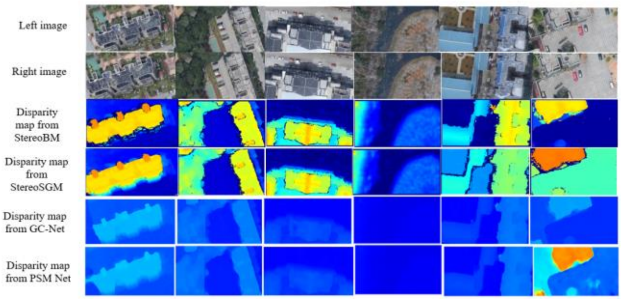
Left image
Right image
Disparity map from StereoBM
Disparity map from StereoSGM
Disparity map from GC-Net
Disparity map from PSM Net

Figure 9. The test results with the WHU stereo images. From the upper to lower: left image, right image, disparity map from StereoBM, StereoSGB, GC-Net, PSM Net The corresponding photo no. from left to right: 3003, 8006, 2005, 4001, 6002, 5007.

However, the reference disparity maps were tailored for the validation of the deep learning methods. The reference images were in the data type of 'uint16'. Instead, disparity maps produced from the unsupervised method were in a type of 'single', which was not accurate when compared to 'uint16'. Furthermore, with the WHU dataset, the training, test, and validation images were extended to 256 times more in the data type of uint16 in order to prevent accuracy loss. Due to these reasons, we keep in mind that the results are shown in Table 1 as a reference. It might not be very accurate. The relative accuracy between StereoBM and StereoSGM, and between GC-Net and PSM Net can be an indicator. Between the StereoBM and StereoSGM, the results relied on the input datasets. In the supervised methods, the GC-Net performed slightly better than the PSM Net. For the ISPRS Vaihingen dataset, the GC-Net was slightly better, while with the WHU dataset, it was the opposite. For the photos of 0024 and 0048, one contained reflective water area, another was textureless field. It can be noticed that the MSE values from the GC-Net and PSM Net were extremely big. With the unsupervised methods, they were normal.

Disparity map / Dataset	Photo No.	StereoBM (MSE:cm)	StereoSGM (MSE:cm)	GC-Net (MSE:cm)	PSM Net (MSE:cm)
ISPRS Vaihingen dataset	0003	252.13	141.73	74.24	77.15
	0016	180.21	149.58	20.66	20.97
	0024	149.84	162.86	244.67	252.54
	0031	200.00	140.47	24.00	23.34
	0038	120.72	83.77	18.51	21.14
	0045	203.06	112.78	28.51	28.80
	0048	146.21	148.25	136.36	145.53
Average		178.88	134.21	78.14	81.35
	2005	31.72	144.97	1.23	1.26
WHU dataset	3003	58.10	373.14	6.68	17.09
	4001	11.65	62.34	0.46	0.25
	5007	73.00	428.52	2.67	3.52
	6002	67.35	197.51	2.44	4.39
	8006	103.43	341.85	5.31	7.02
Average		57.54	258.06	3.13	5.59

Table 1. Accuracy evaluation results.

5. DISCUSSIONS

5.1 Factors that affect the results

In the original GC-Net experiments, disparity D=192 was used [4]. In our GC-Net experiments, the used value was D=160, which can affect the resultant disparity accuracy decreasingly. Also, the used data augmentation methods for GC-Net tests were scarce as only random cropping was used. It might not affect greatly the performance of the model trained with the WHU dataset due to the large dataset size, but the effects are more visible with the model trained with the smaller ISPRS Vaihingen dataset. By increasing the amount of used data augmentation, the performance of that model could be most likely increased. From the loss plots of the models (Fig. 6 and Fig. 8) can be found slight overfitting from the end of training both models, as the validation loss can be noticed to start to increase. Better and more versatile data augmentation methods could also delay when the models start to overfit. In addition, it is good to take the quality of used labels into account. The WHU dataset had ready-accurate disparity labels, while the disparity labels used for the ISPRS Vaihingen dataset were self-made with a rough interpolation that left them a bit uneven. With more delicate interpolation methods the quality of disparity labels could be more polished leading to better performance. This, and the small size of the ISPRS Vaihingen dataset were the main reasons for the lower performance of the model in comparison to the model trained with the WHU dataset. However, from the test outputs of the ISPRS Vaihingen model can be noticed that the model is able to perform reasonably well, though the interpolated labels have some holes and roughness in them. The outputs don't have any holes and look smooth.

With PSM Net, both models were trained only once. By training the models multiple times, some variation could be removed from the results. The results are affected by multiple factors, for example, the choice of loss function and optimizer, the amount of training data, the quality of training data, choice of model architecture, batch size, preprocessing of the data, for example, normalizing and cropping, and PSM Net has an intuitive design, and it can be used to predict the disparity map of aerial stereo images. However, it requires good training data and lots of memory and time during training. Recently, Huang et al. (2021) proposed a method to improve the PSM Net performance. In the paper, the authors mentioned that the PSM Net needs a long run time due to the algorithm dealing with too many parameters. In our test, training with the WHU dataset took 52 hours. In addition, the authors also pointed out that the PSM Net performed not well in the cases of reflective areas and the areas with repetitive pattern textures. We also evidenced such scenes in our experiment.

Unsupervised methods are not as robust as the supervised methods. There are many parameters needed to be set properly in order to achieve a good result. With the StereoBM method, 'BlockSize' defines the width of the search window. If it is set to too small, more details but a lot of noise will present. With the StereoSGM method, there were more parameters affecting the results. Experience and knowledge are important for obtaining a good result. From Table 1 can be seen that in photo no. 5007, the result from the StereoSGM showed a big error compared to the one from the StereoBM. Photo 5007 contains a scene with repetitive textures.

On the positive side, when many parameters are available. It also indicates that the user can control the noise level by setting proper values. And with different scenes such as textureless areas, reflective surfaces, and repetitive patterns, parameters are adjustable. The users' experience is valuable for unsupervised methods.

In addition to the factors affected by the methods, from the given datasets, the ISPRS dataset has an overlap of 60% in both forward and side directions while the WHU dataset has 90% forward overlap and 80% side overlap. Visually, results from the WHU dataset were better.

5.2 The unsupervised vs. supervised methods

5.2.1 Efficiency/Processing time: With the supervised methods, training a model needs a lot of time. The model performance relies on the amount of training data. That is, if one wants to achieve a good performed model, a big amount and a good quality of training data are needed. Thus, training time becomes an issue that the users need to concern about. It is worth doing it when a huge amount of datasets need to be processed. And also, as long as the model is trained, it can be repeatedly utilized for the same tasks in the future.

5.2.2 Robustness: Learned by the machine, the supervised methods became more robust than the unsupervised methods. It is less demanding of user expertise. With the unsupervised methods, users need to have a good understanding of the parameters and test them for a good result.

5.2.3 Handling difficult scenes: Kendall et al. (2017) state that a number of the challenging problems for stereo algorithms would benefit from the knowledge of global semantic context that deep learning methods can utilize, rather than relying solely on local geometry. From our experiment can also be seen that, due to the unsupervised methods focused more on local information, it resulted in poor pixel continuation and surface smoothness for the scene from the ISPRS Vaihingen dataset. However, it performs well with a small scene. The WHU dataset is an example.

The result for reflective areas (water) shows that unsupervised methods presented more noise, but also more details. For areas with textures of the repetitive patterns or featureless areas, it seems that there are rooms for the supervised methods to be improved.

5.3 Future direction

Like Kendall et al. (2017) mentioned, the current state-of-the-art stereo algorithms often have difficulty with textureless areas, reflective surfaces, thin structures, and repetitive patterns, and stereo algorithms aim to mitigate these failures with pooling or gradient-based regularization (Hirschmuller, 2005; Geiger et al., 2010). This kind of approach often requires a compromise between smoothing surfaces and detecting detailed structures.

3D convolution has brought large improvement in cost volume regularization but at a cost of high computational time and runtime memory requirement. 3D scene reconstruction from aerial vehicles is in high demand and faces the above challenges to generate fine-grained city-scale reconstructions; The gap between depth maps and point clouds still exists and a unified framework for depth map based coherent scene reconstruction is needed. The number of available datasets and the diversity of data is not adequate (Wang et al., 2021).

In this experiment, we only tested two-view stereo pairs of images. Multi-view stereo images might improve the result since the redundant data can improve the accuracy and the results should be more reliable.

6. CONCLUSIONS

Our experiment explored the supervised methods and unsupervised methods for disparity map generation from two sets of aerial stereo images: the ISPRS Vaihingen dataset and the WHU dataset. We selected seven pairs of stereo images from the ISPRS Vaihingen dataset and six image pairs from the WHU dataset. With the diffult scene in shadows, reflective water area, textures with repetitive patterns, featureless area, and buildings with different heights and complex roof structures were tested. Overall, from quantative evaluation, the supervised methods showed better accuracy. In resultant disparity maps, building roofs had fewer holes and were smoother, and also the noise level was lower with the supervised methods. However, unsupervised methods, especially with the StereoBM method, perform well in small-size images where the local information was focused on.

The supervised methods require high-performance computational facilities and much time to train the models. As long as the model is trained, it becomes efficiency and user-friendly --- not requiring user expertise.

ACKNOWLEDGEMENTS

The work was supported by the National Land Survey's ATMU project (Advanced Technology for topographic Map Updating), which was funded by the Ministry of Finance in Finland (Valtiovarainministeriö) for the period of 1.1.2021-31.12.2022. Many thanks to the CSC (csc.fi) support on the high-performance computation for the research purpose of the project for free use.

REFERENCES

Baumberg, A., 2000. “Reliable feature matching across widelyseparated views”, In Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, pp. 774-781.

Birchfield, S. and Tomasi, C., 1998. A pixel dissimilarity measure that is insensitive to image sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Bracewell, R., 1965. "Pentagram Notation for Cross Correlation." The Fourier Transform and Its Applications. New York: McGraw-Hill, pp. 46 and 243.

Briechle, K., & Hanebeck, U. D., 2001. Template matching using fast normalized cross correlation. In Optical Pattern Recognition XII (Vol. 4387, pp. 95-102). SPIE. (March, 2001).

Caspi, Y., Simakov, D., & Irani, M., 2006. Feature-based sequence-to-sequence matching. International Journal of Computer Vision, 68(1), 53-64.
Chang, J. R., & Chen, Y. S., 2018. Pyramid stereo matching network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5410-5418).

CSC. ICT Solutions for Brilliant Minds. https://www.csc.fi/. Access date: 10th May, 2022.

Geiger, A., Roser, M., and Urtasun, R., 2010. Efficient large-scale stereo matching. In Asian conference on computer vision, pages 25–38. Springer.

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. CoRR (2017), abs/1703.04309.

KITTI stereo dataset. http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo Access date: 17th Jan, 2022

Konolige, K., 2010. Projected texture stereo. In 2010 IEEE International Conference on Robotics and Automation (pp. 148-155). IEEE. (May, 2010).

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semiglobal matching and mutual information. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 807–814. IEEE.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2):328–341.

Hirschmüller, H., Buder, M., & Ernst, I., 2012. Memory efficient semi-global matching. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 3, 371-376.

Huang, Z., Gu, J., Li, J., & Yu, X., 2021. A stereo matching algorithm based on the improved PSMNet. Plos one, 16(8), e0251657.

ISPRS2021 benchmark. https://github.com/whuwuteng/benchmark_ISPRS2021 Access date: 25th March, 2022.

Laga, H., Jospin, L. V., Boussaid, F., & Bennamoun, M., 2020. A survey on deep learning techniques for stereo-based depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Liu, J.; Ji, S., 2020. A Novel Recurrent Encoder-Decoder Structure for Large-Scale Multi-view Stereo Reconstruction from An Open 748 Aerial Dataset. CoRR 2020, abs/2003.00637.

Mathsworks. https://www.mathworks.com Accessed on 7th April, 2022

Middlebury stereo datasets. https://vision.middlebury.edu/stereo/ Access date: 17th Jan, 2022

PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., Eds.; Curran Associates, Inc., 2019; pp. 8024–8035.

Schmid, C. and Mohr, R., "Local grayvalue invariants for imageretrieval", IEEE Transactions on Pattern Analysis and MachineIntelligence, vol.19, no.5, pp.530-534, 1997.

Wang, C.; Zhang, L., Yuan, J., Xie, L., 2018. Kernel Cross-Correlator. The Thirty-second AAAI Conference On Artificial Intelligence. Association for the Advancement of Artificial Intelligence. pp. 4179–4186.

Wang, Y., Yang, B., Hu, R., Liang, M., & Urtasun, R., 2021.. PLUMENet: Efficient 3D Object Detection from Stereo Images. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3383-3390). IEEE. (January, 2021)

WebOfScience. https://www.webofscience.com/ Accessed on 3rd May, 2022

Yang, G., Manela, J., Happold, M., Ramanan, D., 2019. Hierarchical Deep Stereo Matching on High-Resolution Images. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Yao, Y., Luo, Z., Li, S., Fang, T., & Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European conference on computer vision (ECCV) (pp. 767-783).