# DEEP LEARNING-BASED STEREO MATCHING FOR HIGH-RESOLUTION SATELLITE IMAGES: A COMPARATIVE EVALUATION

X. He [1] , S. Jiang [1,2,*], S. He [3] , Q. Li [4] , W. Jiang [3] , L. Wang [1,2]

[1] School of Computer Science, China University of Geosciences, Wuhan 430074, China - jiangsan@cug.edu.cn
[2] Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China
[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China
[4] College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China

**KEY WORDS:** Satellite Image, Dense Matching, Deep Learning, Semi-global Matching

**ABSTRACT:**

Dense matching plays an important role in 3D modeling from satellite images. Its purpose is to establish pixel-by-pixel correspondences between two stereo images. The most well-known algorithm is the semi-global matching (SGM), which can generate high-quality 3D models with high computational efficiency. Due to the complex coverage and imaging condition, SGM cannot cope with these situation well. In recent years, deep learning-based stereo matching has attracted wide attention and shown overwhelming benifits over traditional algorithms in terms of precision and completeness. However, existing models are usually evaluated by using close-ranging datasets. Thus, this study investigates the recent deep learning models and evaluate their performance on both close-ranging and satellite image datasets. The results demonstrate that deep learning network can better adapt to the satellite dataset than the typical SGM. Meanwhile, the generalization ability of deep learning-based models is still low for the real application at recent time.

## 1. INTRODUCTION

Dense matching of stereo images is a classic problem in the field of photogrammetry and computer vision (Ji et al., 2019). Its core task is to establish the pixel-by-pixel correspondences between two images to recover the 3D information of the target(Geiger et al., 2010). Stereo dense matching has become the most crucial component in many tasks that range from localization tracking to 3D reconstruction (Li et al., 2023b, Jiang et al., 2023, Geiger et al., 2011, He et al., 2021). As the popularity and quality of satellite images continue to improve, stereo matching based on high-resolution satellite images has been widely used in various applications, such as 3D modeling of large-scale cities (Zhang et al., 2022, Facciolo et al., 2017, Huang et al., 2017). Thus, efficient and robust stereo matching becomes the key to applying high-resolution satellite images (Jiang et al., 2021).

Given a pair of rectified stereo images, the first step of stereo dense matching is to compute the disparity of each pixel in the reference image, which is further used to recover the depth and 3D information (Gu et al., 2020). The classic stereo matching algorithm consists of four steps: matching cost calculation, cost aggregation, disparity calculation, and disparity refinement (Scharstein and Szeliski, 2002). Traditional handcrafted stereo matching algorithms are divided into three categories: local matching, global matching, and semi-global matching (Zhong et al., 2017). Among them, semi-global matching (SGM) is a popular and effective method for global optimization, which approximates the path form of the two-dimensional optimal energy function by aggregating the one-dimensional path costs of multiple path directions in the neighborhood (Hirschmuller, 2007). It is widely used in the stereo matching of close-range, aerial, and satellite images (Humenberger et al., 2010, Spangenberg et al., 2013).

In recent years, the use of deep learning networks for stereo dense matching of satellite images (Zbontar et al., 2016, Li et al., 2019) has attracted widespread attention. Compared with traditional algorithms, deep learning-based dense matching has a significant improvement in terms of accuracy and completeness (Zhou et al., 2020, He et al., 2022). The end-to-end deep learning network uses CNN (convolutional neural networks) to integrate matching cost calculation, cost aggregation, and disparity calculation, understand a wider range of context information, and obtain a disparity map through a stereo regression model (Seki and Pollefeys, 2017, Zbontar and LeCun, 2015). By superimposing and combining the features obtained by the multi-layer network, the deep learning network can effectively obtain the geometric and context information of the stereo image (Zhang and Wah, 2017). In the dense matching method using deep learning, GC-Net creates a cost volume to represent the correspondence between the left and right images and uses 3D convolution to calculate the disparity map (Kendall et al., 2017). With greater accuracy than traditional methods, StereoNet simultaneously calculates at a very low-resolution cost using sub-pixel matching (Khamis et al., 2018). PSM-Net constructs a spatial pyramid pooling module and dilated convolution to gather context information and uses a stacked hourglass structure to standardize the cost volume to obtain a disparity map (Chang and Chen, 2018). In HSM-Net, a feature pyramid encoder creates a four-dimensional feature volume, then the decoder generates the necessary disparity map, particularly for high-resolution images (Yang et al., 2019). To improve the accuracy of disparity predictions in low-texture or textureless regions, AANet presents an intra-scale cost aggregation method based on sparse points, and uses the neural network layer (Xu and Zhang, 2020) to approximate the cross-scale cost aggrega-

*Corresponding author

tion algorithm.

At present, deep learning models show clear benefits over traditional stereo dense matching algorithms. To evaluate the performance in the high-resolution satellite datasets, we cannot just rely on the results in close-range datasets (Li et al., 2023a). Therefore, this paper studies several classic deep learning models and evaluates the effectiveness of deep learning algorithms on several datasets, including SceneFlow (Mayer et al., 2016), KITTI 2015 (Menze and Geiger, 2015), US3D (Bosch et al., 2019, Le Saux et al., 2019), and WHU-Stereo (Li et al., 2023c).

## 2. EVALUATED DENSE MATCHING METHODS

### 2.1 The Workflow of Dense Matching

The disparity map records the disparity value of each object point in the image coordinate system (Hartley and Zisserman, 2003). After epipolar rectification, dense matching methods try to find the corresponding point from the right image as much as possible for each pixel in the left image. The result is stored by the disparity map of the left image (Mühlmann et al., 2002). The workflow of dense matching can be divided into 4 steps (Kendall et al., 2017). First, matching costs are calculated to measure the correlation between the pixel to be matched and the candidate pixel. The classic matching cost includes the brightness difference of pixel values, correlation coefficient, and mutual information, etc. These costs are calculated pixel by pixel in the search region using a specific similarity metric based on gray value, gradient, or information entropy, within an image block. Second, matching cost aggregation is then executed, which is usually implemented as the weighted sum of all matching costs in the neighborhood of the matching pixel. Matching cost aggregation has been simplified in traditional algorithms like the Semi-global matching and the GraphCut (Boykov and Jolly, 2001). Calculating the disparity value is the third step. The preferred outcome is the disparity value obtained by minimizing the energy function with the lowest matching costs, which is followed by the optimization of disparities. In general, the disparity value is then refined with a series of post-processing techniques, including the left-right consistency check, median filter, sub-pixel enhancement, etc.

### 2.2 Dense Matching Methods

It is challenging to achieve the mathematical optimum because classical matching algorithms at each stage adopt empirical methods rather than strict mathematical models, such as design features, measures, aggregation methods, etc., and they also have made varying degrees of simplification, such as considering the matching cost of pixels in the neighborhood independently. Current research attempts to see whether deep learning algorithms can overcome the restrictions. To evaluate their performance, this study has chosen six typical algorithms in this field, including the handcrafted SGM algorithm (Hirschmuller, 2007), and five deep learning prediction networks, i.e., GC-Net (Kendall et al., 2017), StereoNet (Khamis et al., 2018), PSM-Net (Chang and Chen, 2018), HSM-Net (Yang et al., 2019), and AANet (Xu and Zhang, 2020). The details are listed as follows.

**2.2.1 Semi-Global Matching** The matching cost of SGM is calculated by computing the Hamming distance of the census transformation values of the two pixels corresponding to the left and right images (Hirschmuller, 2007).The SGM method then uses the global energy optimization strategy to identify the best disparity for each pixel to minimize the global energy function over the whole image based on the calculated matching cost. The 1D matching costs are evenly aggregated from all directions for each pixel, and the 1D minimum costs are added up for all pathways. The winner-take-all (WTA) method is used to calculate disparity, and each pixel chooses the disparity value that corresponds to the lowest aggregation cost value as the final disparity. Finally, disparity optimization is employed to handle incorrect value areas that need to be repaired as well as faults in disparity pictures. Common methods include the removal of peaks, left-right consistency check, and discontinuity preserving interpolation.

**2.2.2 GC-Net:** GC-Net proposes a new deep learning architecture to solve the end-to-end stereo matching problem while using the deep convolutional network formula to explicitly reason about geometry and semantics using a deep convolutional network formulation, to understand global semantic context knowledge, rather than relying solely on local geometry (Kendall et al., 2017). As shown in Figure 1, the left and right stereo images in GC-Net go through a series of 2D convolutions to form a unary feature of shared parameters, which is then cascaded with the feature map under each disparity in the right image. Furthermore, the unary feature is encapsulated into a four-dimensional cost volume. The context information in the data is then combined using the deep convolution encoder-decoder network architecture, and after getting the multi-scale features, the regularization of the cost volume in the disparity dimension is accomplished using the defined soft argmin function. GC-Net train the model with supervised learning using ground truth depth data, the supervised regression loss is defined in Equation 1, where N is the labeled pixels, loss value is the absolute error between the ground truth disparity $\hat{d}_i$ and the models's predicted disparity $d_i$.

$$Loss = \frac{1}{N}\sum_{i=1}^{N} ||d_i - \hat{d}_i|| \tag{1}$$

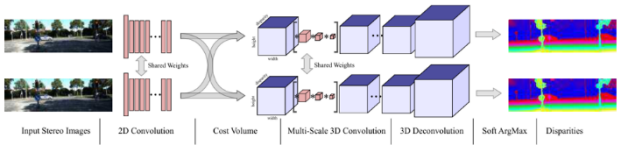



Figure 1: Network structure of GC-Net.
(Kendall et al., 2017)

**2.2.3 StereoNet:** Although some encoder-decoder networks solve the stereo matching problem end-to-end without postprocessing and show good performance on various benchmarks, the proposed methods require vast amounts of processing power and memory. StereoNet applies edge-aware filtering stages in a multi-scale manner to deliver high quality output and uses a very low resolution cost volume to accomplish the real-time function. StereoNet provides a course disparity estimate by extracting image features between input image pairs through a Siamese network with shared weights, matching features along scan lines and constructing a cost volume. Finally, a single-pass optimization is used to upsample the disparity output to full resolution, recovering thin structures and small objects. This is accomplished by hierarchically optimizing the disparity output with an edge-preserving refinement network. Similar to GC-Net, Stereo is trained in a fully supervised manner using groundtruth-labeled stereo data, and hierarchical loss functionis

minimized by Equation 2, where $d_i^k$ is the predicted disparity at pixel I at the k-th refinement level, and $\hat{d}_i$ is the groundtruth disparity at the same pixel. Finally, $rho(.)$ is to approximate a smoothed L1 loss.

$$Loss = \sum_k \rho(d_i^k - \hat{d}_i) \tag{2}$$

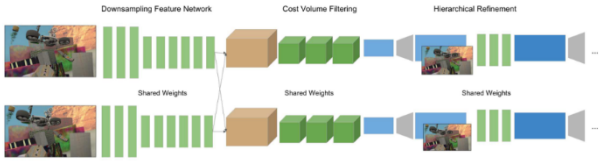



Figure 2: Network structure of StereoNet.
(Khamis et al., 2018)

**2.2.4 PSM-Net:** Due to the lack of using context information to find the corresponding relationship between occlusions and textureless regions when performing feature extraction through the Siamese network, PSM-Net proposes a spatial pyramid pooling module (SPP) and a stacked hourglass structure to realize the aggregation of global context information at the level of the entire image (Chang and Chen, 2018). As shown in Figure 3, the left and right stereo images are input into two weight-shared CNN channels to calculate feature maps, and the pyramid pooling module employs four-scale average pooling in conjunction with dilated convolution. While enlarging the receptive fields, pixel-level features are extended to region-level features with different receptive field scales, which are used to form a reliable cost volume for disparity estimation. To maximize the utilization of the global context information, the cost volume is then sent using an hourglass encoder-decoder system with intermediate supervision layers and is repeatedly modified by several fine-to-coarse and coarse-to-fine operations. The three major hourglass networks that make up the stacked hourglass structure each produce a disparity map, which leads to three outputs and three losses. Each loss value is obtained by Equation 3 and 4, where $d_i$ is the predicted disparity, and $\hat{d}_i$ is the groundtruth disparity, and N is the number of labeled pixels. The overall loss value is then produced by adding the three loss values in weighted fashion.

$$Loss = \frac{1}{N} \sum_{i=1}^{N} smooth_{L_1}(d_i - \hat{d}_i) \tag{3}$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 0 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{4}$$

**2.2.5 HSM-Net:** Many deep stereo networks do not execute well on high-resolution images because of memory or speed limitations. In order to address this issue, HSM-Net produces four-dimensional feature volumes of various resolutions, from coarse to fine, using a high-resolution encoder to calculate the image's features (Yang et al., 2019). The decoder decodes the feature volumes and produces a high-quality disparity map while considering the running time. First, after the features of the left and right images are obtained by the feature encoder, a four-level feature volume pyramid is built based on the differences between potential matching descriptors along horizontal scan

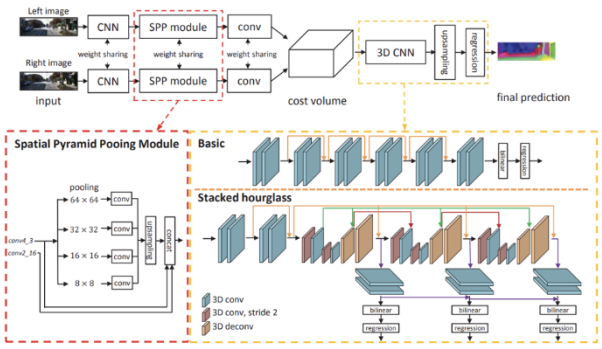



Figure 3: Network structure of PSM-Net.
(Chang and Chen, 2018)

lines, and the spatial and disparity resolutions of each level increase sequentially. Second, the feature volume is filtered by six 3D convolutional blocks in the decoder, a volumetric pyramid pooling operation is applied, and the minimum-scale feature volume in the feature pyramid is upsampled to a higher spatial resolution through trilinear interpolation, merging with the following feature volume in the pyramid. At this time, the disparity can be calculated based on the feature volume of the current scale to generate a three-dimensional cost volume, which takes the least amount of time to generate, a more precise disparity map can be recalculated by the final feature volume. A natural loss is a softmax distribution such in GC-Net, over candidate disparities at the current pyramid level. The final loss value is determined by Equation 5, where $L_1$ is the loss on the finest level, and $L_4$ represents the loss on the most coarse level. Figure 4 depicts the HSM-Net network structure.

$$Loss = L_1 + \frac{1}{2^2} + \frac{1}{2^4} L_3 + \frac{1}{2^6} L_4 \tag{5}$$

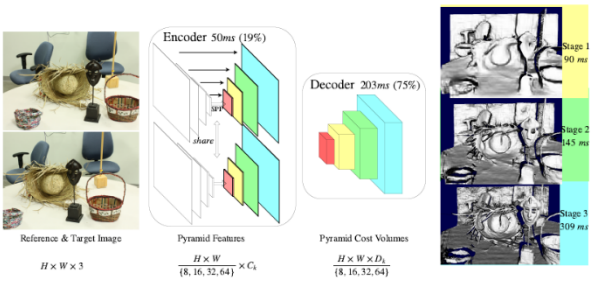



Figure 4: Network structure of HSM-Net.
(Yang et al., 2019)

**2.2.6 AANet:** A majority of dense matching networks are based on 3D convolution that causes high memory consumption and cubic computational complexity. To increase running speed and keep accuracy at the same level, AANet seeks to replace the widely used 3D convolution (Xu and Zhang, 2020). An intra-scale and a cross-scale cost aggregation module are included in AAModules in AANet for this purpose. According to Figure 5, after extracting the downsampling pyramid from a particular pair of left and right images using the shared feature extractor, create a multi-scale 3D cost volume by correlating the left and right image features at the appropriate scales. Then the original cost volume is aggregated by six stacked AAModules, where each AAModule consists of three intra-scale cost

aggregation (ISA) and one cross-scale cost aggregation (CSA) modules. The ISA module is a representation method based on sparse points, which realizes efficient and flexible cost aggregation and alleviates the well-known edge-fattening issue at disparity discontinuities. The CSA module introduces multi-scale interaction in the traditional cross-scale cost aggregation, and the final cost volume is obtained by adaptively combining the cost aggregation results performed at different scales. Finally, the predicted low-resolution disparity layers are upsampled to the original resolution using a refinement module. The corresponding loss function is defined as Equation 6, where V(p) is a binary mask to denote whether the ground truth disparity for pixel p is available, $\hat{d}_i$ is the ground truth disparity and $d_{pseudo}$ is the pseudo ground truth. As Equation 7, the final loss function is a combination of losses over all disparity predictions where $\lambda_i$ is a scalar for balancing different terms.

$$L_i = \sum_p V(p) * smooth_{L_1}(d_i, \hat{d}_i) + (1 - V(p)) * smooth_{L_1}(d_i, d_{pseudo}) \quad (6)$$

$$Loss = \sum_{i=1}^{N} \lambda_i * L_i \quad (7)$$

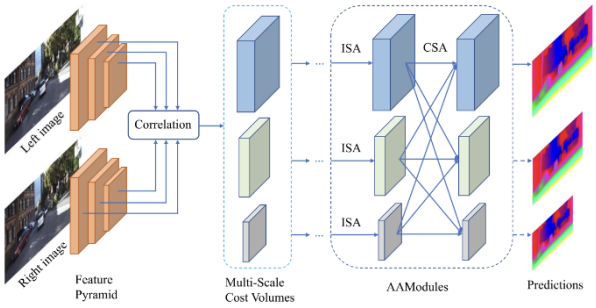



Figure 5: Network structure of AANet. (Xu and Zhang, 2020)

## 3. EVALUATION METRICS AND DATASETS

In order to evaluate the performance of each stereo matching algorithm under different datasets, our experiments include two evaluation metrics: EPE and D1. We also employ four distinct dataset types—SceneFlow, KITTI 2015, US3D, and WHU-Stereo, and split each dataset's data into a training set (80%) and a validation set (20%). The details are as follows.

### 3.1 Evaluation Metrics

We choose endpoint error (EPE) and 3-pixel error ratio (D1) as the evaluation indicators of the comparison method. EPE is the average of the Euclidean distances between the predicted value and the true value. D1 refers to the percentage of error points in all effective pixels on the basis that the difference between the predicted value and the real disparity value exceeds 3 pixels, which is considered an error.

$$EPE = \frac{1}{N} \sum_{k \in T} |\hat{d}_k - \widetilde{d}_k| \quad (8)$$

$$D1 = \frac{1}{N} \sum_{k \in T} [|\hat{d}_k - \widetilde{d}_k| > t] \quad (9)$$

where $\hat{d}_k$ = ground-truth disparity
$\widetilde{d}_k$ = estimated disparity
$N, T$ = number and set of labelled pixels in the image
$t$ = threshold of erroneous disparity

### 3.2 Datasets

**3.2.1 Sceneflow:** Flyingthings3d, Monkka, and Driving are the three subsets that make up the SceneFlow dataset. The majority of the items in Flyingthings3D fly in random 3D trajectories. Monkaa contains non-rigid and soft joint movements, as well as visually challenging hair. Driving is mainly naturalistic and dynamic street scene. The dataset offers dense disparity maps as ground truth data and consists of a total of 35,454 training images and 4,370 test images with an image size of 960*540 pixels. Figure 6(a) is an example graph from Sceneflow with the disparity map colored for easier detail detection.

**3.2.2 KITTI 2015:** KITTI 2015 is a real street view dataset for driving cars. It includes 200 test picture pairs without ground truth disparity and 200 training stereo image pairs with sparse ground truth disparity acquired using LiDAR. The average image size is 1240*376 pixels.The sample graph of KITTI 2015 is shown in Figure 6(b), where the disparity map is colorized for picking up more detail.

**3.2.3 US3D:** The US3D dataset is a large-scale remote sensing image dataset proposed for multiple tasks, including stereo semantic stereo, multi-view semantic 3D reconstruction, single-view height estimation and point cloud semantic segmentation. For stereo matching, 4292 RGB image pairs and publicly available ground truth disparity maps are provided, and the image size is 1024*1024 pixels. In Figure 6(c), a sample graph in US3D is displayed. The images, collected from the WorldView-3 satellite, cover the cities of Jacksonville and Omaha in the United States.

**3.2.4 WHU-Stereo:** Similar to the US3D dataset, WHU-Stereo is an open-source dataset used to match stereo pairs of high-resolution satellite pictures. Among the 1981 epipolar rectification stereo image pairings in WHU-Stereo, 1757 pairs offer ground truth information derived from aerial LiDAR point clouds. The disparity map, which covers six Chinese cities, is saved as a 16-bit float value in the dataset, which is made up of panchromatic band pictures with a 16-bit depth and a 1024*1024 pixel size. Figure 6(d) is a sample graph in WHU-Stereo.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

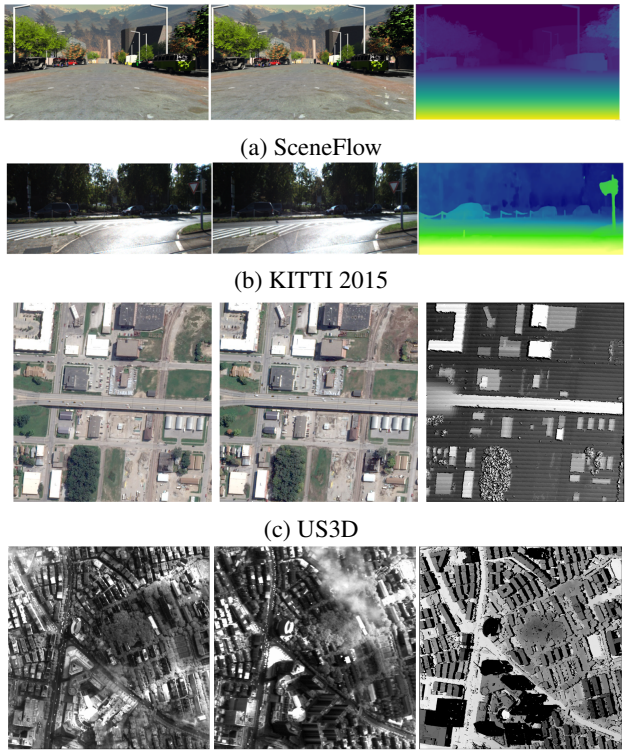
To comprehensively evaluate the performance of stereo matching algorithms on high-resolution satellite images, we designed two types of experiments. The first type of experiment is to use KITTI 2015, US3D, and WHU-Stereo three datasets to compare the dense matching performance of SGM and the deep learning network. The second category is to verify the generalization ability of the deep learning network. Without any fine-tuning, the pre-trained HSM-Net network model on the training dataset is applied immediately to the target dataset, and the model's degree of deterioration is assessed and compared.

| Datasets | SGM | GC-Net | StereoNet | PSM-Net | HSM-Net | AANet |
|---|---|---|---|---|---|---|
| KITTI 2015 | 45.082 | 2.732 | 2.09 | 1.528 | 1.344 | 1.706 |
| US3D | 38.534 | 1.819 | 1.796 | 1.515 | 1.710 | 1.354 |
| WHU-Stereo | 10.426 | 3.17 | 3.615 | 2.753 | 9.634 | 7.765 |

Table 1: EPE of deep learning models and SGM method on 3 datasets

| Datasets | SGM | GC-Net | StereoNet | PSM-Net | HSM-Net | AANet |
|---|---|---|---|---|---|---|
| KITTI 2015 | 93.0 | 8.7 | 11.5 | 6.2 | 7.4 | 10.6 |
| US3D | 45.6 | 14.0 | 13.6 | 9.5 | 12.3 | 15.6 |
| WHU-Stereo | 71.2 | 35.0 | 39.8 | 29.6 | 55.3 | 51.0 |

Table 2: D1 of deep learning models and SGM method on 3 datasets



(a) SceneFlow

(b) KITTI 2015

(c) US3D

(d) WHU-Stereo

Figure 6: Sample images in the four datasets. The left images, right images, and disparity maps are listed from left to right.

The traditional SGM algorithms are available in OpenCV and are easy to implement with Python. As for deep learning methods, we implement our approach in PyTorch and use Adam ($\beta_1$ = 0.9, $\beta_2$ = 0.999) as optimizer. These models are all trained on by an NVIDIA GeForce RTX 3080 on Windows 10 OS, while the SGM method is also implemented on Windows 10 OS. We used the pre-trained model on the SceneFlow dataset for 10 epochs when testing on the KITTI 2015 dataset, then finetuned it on the KITTI 2015 dataset for 100 epochs. In the process of training and fine-tuning, the image crop size is set to 256*512, the learning rate is set to 0.001, the maximum disparity search range is set to [0,192], limited by the graphics card memory, and the batch size is set to 4. For the satellite image dataset, the image in the US3D dataset is cut to 256*512 pixels during training, the disparity search range is [-96,96], the learning rate is set to 0.001, and 100 epochs are trained from scratch. When training on the WHU-Stereo dataset, the image is also cut to 256*512 pixels, the disparity search range is set to [-128,64], and a total of 120 epochs are trained. The initial learning rate is set to 0.001, and as the training progresses, the learning rate is reduced by half every 10 epochs.

### 4.1 Traditional vs Deep Learning Network

To validate the effectiveness of each algorithm proposed in this paper, we first compare the traditional SGM with deep learning networks. The disparity maps computed from all testing samples of the three datasets in Tables 1 and 2 are used to compute EPE and D1. Each deep learning model has superior robustness for data in complicated contextual information and has higher accuracy compared to the conventional SGM method, as predicted. PSM-Net has produced comparatively the best results on the KITTI 2015, US3D, and WHU-Stereo data sets since different deep learning models have distinct network architectures and varying learning capacities.

In order to show the difference between the results of each model more intuitively, Fig 7 gives a visual example on KITTI 2015 test set. We also list the visualization results on the US3D and WHU-Stereo datasets in Figure 8 and Figure 9, respectively. From the table 1, table 2 and the visualization results, except for the WHU-Stereo dataset, the accuracy of the other two datasets is much higher than that of SGM. As shown in Figures 7 and 8, the disparity maps created by the methods of deep learning in the KITTI 2015 dataset are more comprehensive in repetitive areas as non-textured roads and sky, and clearer in detailed areas as car outlines and sign edges. As for WHU-Stereo and US3D datasets, compared with the SGM algorithm, the disparity maps predicted by deep learning models are smoother and more complete on building footprint. AANet employs the pre-trained model on Sceneflow dataset to predict the disparity maps on KITTI2015, and uses the prediction results as pseudo labels in pixels, but when training on US3D and WHU-Stereo without pseudo ground truth supervision. Additionally, there are a lot of holes in the disparity map obtained by the traditional SGM method, and the hole area needs to be filled through post-processing, while the deep learning method directly obtains the disparity map through end-to-end learning.

### 4.2 Generalization of Deep Learning Methods

Generalization learning is to transfer the trained model parameters to another new model to help training. The new model can use the learnt model parameters to accelerate and maximize learning efficiency based on the correlation of the data. The US3D and WHU-Stereo used in this paper belong to satellite building image datasets. However, the WHU-Stereo contains single channel grayscale images, and the US3D includes three channels RGB images, so Generalization learning cannot be completed using these two datasets. Table 3 shows the experimental results of the HSM-Net network model applied to the test data set after the pre-training model is obtained on the test dataset. HSM-Net has good generalization ability, using the KITTI 2015 dataset as the training set, and the 3-pixel error ratio of prediction on the US3D dataset is 52.8%.

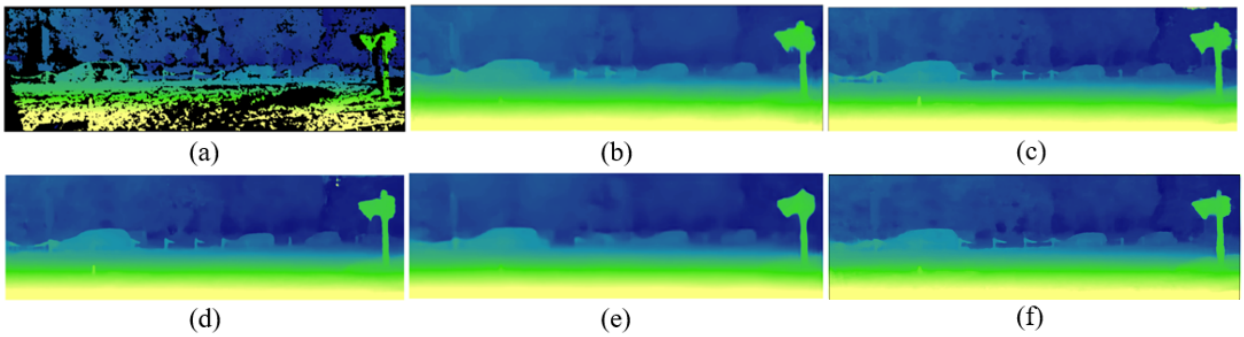



Figure 7: Results on KITTI 2015. From left to right and top to bottom are the disparity maps of SGM, GC-Net, StereoNet, PSM-Net, HSM-Net, and AANet.

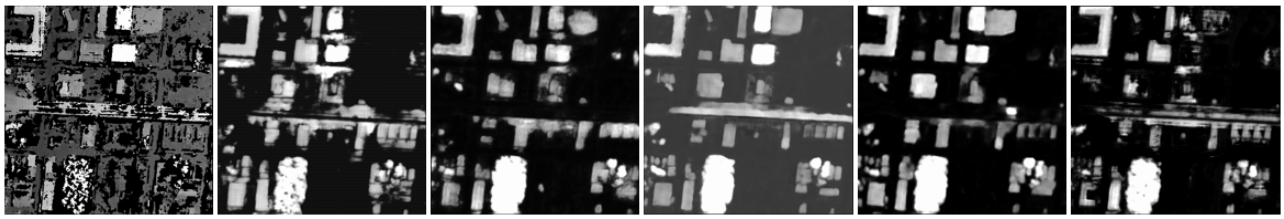


Figure 8: Results on US3D. From left to right are the disparity maps of SGM, GC-Net, StereoNet, PSM-Net, HSM-Net, and AANet.

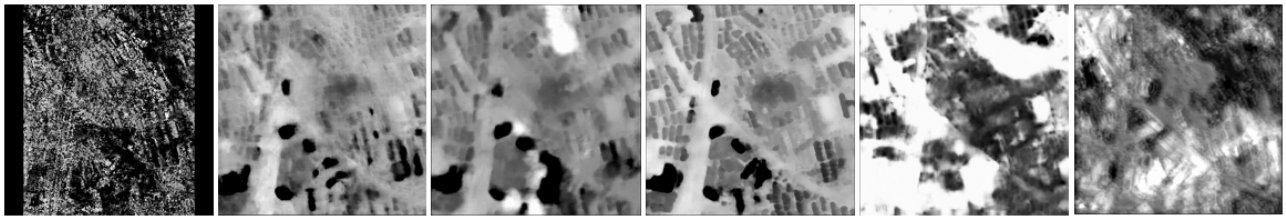


Figure 9: Results on WHU-Stereo. From left to right are the disparity maps of SGM, GC-Net, StereoNet, PSM-Net, HSM-Net, and AANet.

| Test dataset | Training dataset | |
|---|---|---|
| | KITTI 2015 | US3D |
| KITTI 2015 | 7.4 | 12.1 |
| US3D | 58.3 | 52.8 |

Table 3: The results of the HSM-Net pre-trained model on the target dataset (D1)/%

## 5. CONCLUSION

Using two convolutional computer vision datasets and two satellite image datasets, we thoroughly investigated the stereo matching method of deep learning based on the traditional stereo matching algorithm. We used the end point error (EPE) and the proportion of 3-pixel error (D1) as indicators to evaluate the chosen five representative deep learning networks. The outcomes demonstrate that the deep learning network can better adapt to the satellite dataset than the typical stereo matching method SGM. Moreover, the end-to-end matching deep learning network may acquire the predicted disparity map without the need for post-processing. The deep learning network's generalization capacity, however, is subpar. The accuracy significantly decreases when the model developed on the KITTI 2015 dataset is used to the US3D dataset, making the benefit over the conventional algorithm less clear.

## ACKNOWLEDGEMENTS

The authors would like to thank authors who have made their algorithms as free and open-source software packages, which is really helpful to the research in this paper. This research was funded by the National Natural Science Foundation of China (Grant No. 42001413) and the High-Resolution Remote Sensing Application Demonstration System for Urban Fine Management (Grant No. 06-Y30F04-9001-20/22).

## REFERENCES

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.

Boykov, Y. Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, 1, IEEE, 105–112.

Chang, J.-R., Chen, Y.-S., 2018. Pyramid stereo matching network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5418.

Facciolo, G., De Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite im-

ages. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 57–66.

Geiger, A., Roser, M., Urtasun, R., 2010. Efficient large-scale stereo matching. *ACCV (1)*, 25–38.

Geiger, A., Ziegler, J., Stiller, C., 2011. Stereoscan: Dense 3d reconstruction in real-time. *2011 IEEE intelligent vehicles symposium (IV)*, Ieee, 963–968.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.

He, S., Li, S., Jiang, S., Jiang, W., 2022. HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 314–330.

He, S., Zhou, R., Li, S., Jiang, S., Jiang, W., 2021. Disparity estimation of high-resolution remote sensing images with dual-scale matching network. *Remote Sensing*, 13(24), 5050.

Hirschmuller, H., 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2), 328–341.

Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote sensing of environment*, 196, 56–75.

Humenberger, M., Engelke, T., Kubinger, W., 2010. A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, IEEE, 77–84.

Ji, S., Liu, J., Lu, M., 2019. CNN-based dense image matching for aerial remote sensing images. *Photogramm. Eng. Remote Sens*, 85, 415–424.

Jiang, S., Jiang, W., Wang, L., 2021. Unmanned aerial vehicle-based photogrammetric 3d mapping: A survey of techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 10(2), 135–171.

Jiang, S., Li, Y., Weng, D., You, K., Chen, W., 2023. 3D reconstruction of spherical images: A review of techniques, applications, and prospects. *arXiv preprint arXiv:2302.04495*.

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.

Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S., 2018. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. *Proceedings of the European Conference on Computer Vision (ECCV)*, 573–590.

Le Saux, B., Yokoya, N., Hansch, R., Brown, M., Hager, G., 2019. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1), 103–105.

Li, J., Huang, X., Feng, Y., Ji, Z., Zhang, S., Wen, D., 2023a. A Hierarchical Deformable Deep Neural Network and an Aerial Image Benchmark Dataset for Surface Multi-View Stereo Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, Q., Huang, H., Yu, W., Jiang, S., 2023b. Optimized Views Photogrammetry: Precision Analysis and a Large-Scale Case Study in Qingdao. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 1144–1159.

Li, S., He, S., Jiang, S., Jiang, W., Zhang, L., 2023c. WHU-Stereo: A Challenging Benchmark for Stereo Matching of High-Resolution Satellite Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, W., He, C., Fang, J., Zheng, J., Fu, H., Yu, L., 2019. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, 11(4), 403.

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.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3061–3070.

Mühlmann, K., Maier, D., Hesser, J., Männer, R., 2002. Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision*, 47, 79–88.

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

Seki, A., Pollefeys, M., 2017. Sgm-nets: Semi-global matching with neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 231–240.

Spangenberg, R., Langner, T., Rojas, R., 2013. Weighted semi-global matching and center-symmetric census transform for robust driver assistance. *Computer Analysis of Images and Patterns: 15th International Conference, CAIP 2013, York, UK, August 27-29, 2013, Proceedings, Part II 15*, Springer, 34–41.

Xu, H., Zhang, J., 2020. Aanet: Adaptive aggregation network for efficient stereo matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968.

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.

Zbontar, J., LeCun, Y., 2015. Computing the stereo matching cost with a convolutional neural network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1592–1599.

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

Zhang, C., Cui, Y., Zhu, Z., Jiang, S., Jiang, W., 2022. Building height extraction from GF-7 satellite images based on roof contour constrained stereo matching. *Remote Sensing*, 14(7), 1566.

Zhang, F., Wah, B. W., 2017. Fundamental principles on learning new features for effective dense matching. *IEEE Transactions on Image Processing*, 27(2), 822–836.

Zhong, Y., Dai, Y., Li, H., 2017. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*.

Zhou, K., Meng, X., Cheng, B. et al., 2020. Review of stereo matching algorithms based on deep learning. *Computational intelligence and neuroscience*, 2020.