EVALUATION OF DEEP LEARNING BASED STEREO MATCHING METHODS: FROM GROUND TO AERIAL IMAGES

J. Liu [1], S. Ji [1,*], C. Zhang [1], Z. Qin [1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China - (liujinwhu; jishunping; whuzhangchi; qzj253qz)@whu.edu.cn

Commission II, WG II/4

KEY WORDS: Dense image matching; Deep learning; Convolutional neural network; Aerial stereos; Transfer learning

ABSTRACT:

Dense stereo matching has been extensively studied in photogrammetry and computer vision. In this paper we evaluate the application of deep learning based stereo methods, which were raised from 2016 and rapidly spread, on aerial stereos other than ground images that are commonly used in computer vision community. Two popular methods are evaluated. One learns matching cost with a convolutional neural network (known as MC-CNN); the other produces a disparity map in an end-to-end manner by utilizing both geometry and context (known as GC-net). First, we evaluate the performance of the deep learning based methods for aerial stereo images by a direct model reuse. The models pre-trained on KITTI 2012, KITTI 2015 and Driving datasets separately, are directly applied to three aerial datasets. We also give the results of direct training on target aerial datasets. Second, the deep learning based methods are compared to the classic stereo matching method, Semi-Global Matching(SGM), and a photogrammetric software, SURE, on the same aerial datasets. Third, transfer learning strategy is introduced to aerial image matching based on the assumption of a few target samples available for model fine tuning. It experimentally proved that the conventional methods and the deep learning based methods performed similarly, and the latter had greater potential to be explored.

1. INTRODUCTION

Dense stereo matching is a classic topic in photogrammetry and computer vision, through which 3D scenes can be further reconstructed. Conventional stereo methods could be grouped into four stages: matching cost calculation, matching cost aggregation, disparity calculation and disparity refinement (Scharstein and Szeliski, 2002). The differences between pixel values or gradients, correlation coefficients and mutual information are typical matching costs. However, these costs are inevitably impacted by texture-less areas, reflective surfaces, thin structures and repetitive patterns (Kendall et al., 2017).

Matching cost aggregation is the strategy to integrate votes (usually measured by the disparity difference between current points and neighbourhood points) from a given neighbourhood and possibly correct the current matching point. SGM (Hirschmüller, 2007) and Graph Cut (Boykov and Jolly, 2001) are two classic stereo methods that employ different aggregation strategy. The latter uses graph model to minimum energy in a 2D neighbourhood region. The former utilizes several 1D cost aggregations to simulate a 2D optimization problem, and greatly improves efficiency. However, both of the solutions assume that every pixel (and disparity) is independent within the neighbourhood. However, it may be not the case as the context and geometric information could be more complicated.

From 2015, the deep learning based methods have been gradually introduced to stereo matching and have shown to be promising. Deep neural convolutional networks (CNN) automatically learn multi-level representations that map the original input to the designated binary or multiple labels (a classification problem), or consecutive vectors (a regression problem). The powerful representation learning ability of CNN has made it gradually replacing the conventional feature handcrafting strategies in detection, classification and stereo applications.

The MC-CNN (Žbontar and Lecun, 2014) is an early attempt to replace the empirical matching cost by multi-layer representations automatically learned by a CNN structure. With proper pre-training for the challenging cases as reflective surface and sharp disparity change, more robust matching cost could be expected. It experimentally proved that MC-CNN obtained better results compared to other matching costs as absolute difference of brightness, census and normalized correlation (Žbontar and Lecun, 2014).

Other CNN networks produce disparity map directly from original stereo pair in an end-to-end manner (Kendall et al., 2017; Pang et al., 2017; T.Brox, 2016). GC-Net (Kendall et al., 2017) learns to incorporate contextual information using 3-D convolutions over a cost volume of cross-disparity feature representations and pack the volume to 2D map to regress disparity values. (Pang et al., 2017) propose a cascade CNN architecture composing of two stages. The first stage utilizes DispNet (T.Brox, 2016) and the second stage rectifies the disparity initialized by the first stage and generates residual signals across multiple scales. (Shaked and Wolf, 2016) Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning] presents an three-step pipeline based on a highway network architecture for the stereo matching problem including computing matching cost, cost aggregation and parallax refinement. These end-to-end methods, especially GC-Net which integrates geometric and contextual information in higher dimension, greatly alleviate the assumptions that pixels of a neighbourhood are independent in a traditional matching cost aggregation.

Basically, in the open-source KITTI 2012 and 2015 Datasets (Geiger, 2015), the deep learning based methods achieve top scores and conventional methods appear uncompetitive. However, deep learning based methods have some challenges. First, the deep learning methods require samples to train their models. Whether a model pretrained on an open dataset could

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-2-593-2018

be directly applied to a target dataset requires further inspection. Second, the KITTI and other Datasets as Driving (T.Brox, 2016) are close-range images, whether the deep learning based methods could well function on aerial dataset should be further checked.

In this paper, we attempt to answer two questions: 1) Does the deep learning based stereo methods have enough generalization ability which guarantees transfer learning from trained models on some open-source dataset to target dataset, and 2) if they could be used in aerial images and outperform traditional methods?

2. METHODOLOGY

2.1 SGM and SURE

SGM (Hirschmüller, 2007) is a classic stereo method that have been widely studied and applied on photogrammetry and computer vision communities. Many variants are developed from SGM, and the SURE software utilizes a multi-view SGM strategy to generate DSM with high accuracy.

The greatest contribution of SGM is the aggregation is achieved by several 1D summing other than 2D summing in neighbourhood like Graph Cut stereo method (Boykov and Jolly, 2001) that results in a very slow processing. SGM utilizes cross-entropy information for matching cost, and shows better than the difference of pixel values.

SURE (Rothermel et al., 2012) firstly generate stereo pairs that are especially convenient for multi-view matching and for multi-view geometry recovering. Then, for each stereo pair in a multi-view group, SGM is applied to obtain the parallax map, separately. At last, the redundant depth estimations across single stereo models are merged through a fusion step. Image pyramid strategy is also utilized to limit the searching area and improve the efficiency.

2.2 MC-CNN

MC-CNN (Žbontar and Lecun, 2014) utilizes a simple Siamese CNN network to extract high-level feature representations from stereo images separately and compare their similarity by a cross product. In Figure 1, image patches are convoluted with convolution kernels and activated with ReLU layer by layer till the last layer with no activation. The last layer features are then packed to 1D and normalized for computing similarity score by dot product.

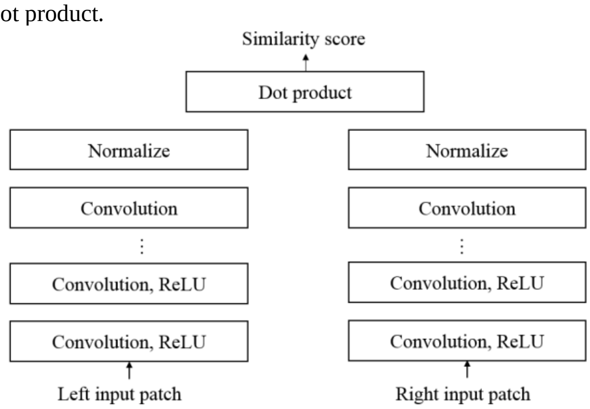

Figure 1. Learning similarity score by MC-CNN (cited from (Žbontar and Lecun, 2014))

The rest process of MC-CNN, i.e., cost aggregation, consistency tests, is similar to SGM.

2.3 GC-Net

GC-Net (Kendall et al., 2017) is an end-to-end strategy that produce disparity maps from inputs of rectified stereo images. First, the stereo images are convoluted by 2D convolution kernels several times to extract feature maps, with shared weights between stereo inputs. The feature maps are then concatenated cross each disparity to form a 3D tensor of *width×height×disparity*. The 3D feature maps are further abstracted by a multiscale 3D convolution and deconvolution. At the last layer, the 3D features, with the same size of original input are flattened to disparity maps by a soft argmax operation. The maps are finally compared to the ground truth by L_1 norm to train the network iteratively.

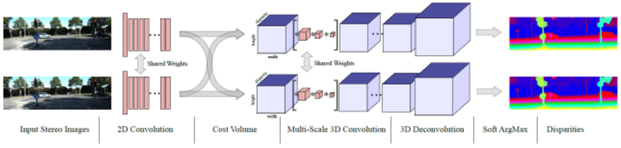

Figure 2. the network structure of GC-Net (cited from (Kendall et al., 2017))

Although there are many novel architectures proposed for stereo matching recently, GC-Net shows its robustness and accuracy, and occupies one of the top scores of the KITTI benchmarks (Geiger, 2015).

2.4 Transfer learning

Transfer learning is a strategy that utilizes the pre-trained model on a source dataset, to apply on a target dataset with a few or without new samples. A case is to predict from the target dataset without parameter tuning. A good result demands for the model is robust and has good generalization ability; otherwise, the sample space of the source and target datasets is expected to be similar. Another case is to leverage the parameters of the pre-trained model as initial state and update them in fine tuning stage with new target samples. In this case, one could freeze the backbone of the network and only train the parameters of the last several layers; or could train all the parameters of the model. The depth of the network structure and the number of target samples usually determine which one to choose.

In our case, to evaluate the generalization ability of a CNN, we firstly directly applied the pre-trained models on source ground/aerial datasets to a target aerial dataset. Then, we tuned all parameters (for MC-CNN and GC-Net are both narrow networks) through new samples.

3. DATASETS

We prepare 5 datasets to thoroughly evaluate the performances of the CNN based methods on aerial stereo images. Two of them are open and close-range datasets: KITTI and Driving datasets. The rest consists of aerial images.

3.1 Close-range datasets

KITTI stereo dataset: The KITTI dataset was produced in 2012 and extended in 2015 (Geiger, 2015). KITTI 2012 dataset contains 194 training images and 195 test images while KITTI 2015 contains 200 training and 200 test images with a size of 1242 × 375 pixels. The epipolar rectified image pairs were acquired by two video cameras mounted on a car. The ground truth, i.e., the depth maps, was acquired by a rotating laser scanner. As similar to the settings of many other studies, we utilize 80% of the images for training and the rest for test.

Driving stereo dataset: The Driving dataset was produced from a virtual street scene from the viewpoint of a virtual driving car, which is similar to the KITTI dataset. It contains more than 4k image pairs with different focal length, scene direction, driving speed with corresponding ground truth maps.

We only select 300 image pairs for experiment within which 80% of the images are used for training and the rest for test. Each image has the size of 960×540 pixels.

3.2 Aerial datasets

Hangzhou stereo dataset: The Hangzhou dataset consists of aerial images with 80% overlap in strip and 60% overlap between strips, acquired from a UAV. After all images are epi-rectified, we cropped them into tiles of 1325×354 pixels, which is suitable for the capacity of a Titan Xp GPU video card. The ground truth was acquired by a laser scanner. After removing a few images with undesirable disparities by manual check, we select 328 image tile pairs as training set and 40 pairs as testing set.

München and Vaihingen dataset: Similar to Hangzhou dataset, the München and Vaihingen dataset were acquired from aerial images with 80% and 60% overlaps respectively. The ground depth maps were acquired from a given DSM, which was generated by the median values of the DSM products from several photogrammetric commercial software. We also cropped the whole image to tiles to suit the capacity of a mainstream video card. Finally, The München dataset consists of 260 stereo pairs with size of 1150×435 pixels while the Vaihingen dataset consists of 730 stereo pairs with size of 955×360 pixels. The ratio between training and test data is also set to 4:1.

Due to the capacity (6G) of our Titan Xp GPU video card, we trained the networks on the three aerial datasets with half pixel resolution.

4. RESULTS

4.1 CNN methods on aerial datasets

We evaluate the deep learning based methods for aerial stereo images by a direct model reuse. The models, as well as all of the parameters, pre-trained on virtual/real street scene datasets, KITTI 2012, KITTI 2015 and Driving datasets, are directly applied to the three aerial datasets, Hangzhou, München and Vaihingen datasets. The MC-CNN and GC-net separately pre-trained on the street scene benchmarks are applied to our aerial datasets. The performance of training on target aerial dataset set are also given for comparison. The results are presented in Table 1 and Table 2.

Training set Test set	**KITTI2012**	**KITTI2015**	**Hangzhou**	**Munchen**	**Vaihingen**
KITTI2012	**0.963**	0.957 (-0.006)	0.941 (-0.022)	0.945 (-0.018)	0.946 (-0.017)
KITTI2015	0.958 (-0.002)	**0.960**	0.951 (-0.009)	0.955 (-0.005)	0.953 (-0.007)
Hangzhou	0.944 (-0.009)	0.942 (-0.011)	**0.953**	0.948 (-0.005)	0.940 (-0.013)
Munchen	0.960 (-0.005)	0.960 (-0.005)	0.960 (-0.005)	**0.965**	0.959 (-0.006)
Vaihingen	0.988 (-0.004)	0.987 (-0.005)	0.987 (-0.005)	0.989 (-0.003)	**0.992**
Driving	0.889	0.888	0.880	0.886	0.872

Table 1. Test results (parallax error < 3pixels considered a correct match) on the different training datasets based on MC-CNN. The numbers in bracket are the difference between the current number and the diagonal number of the current row, indicating the decreasing degree of accuracy when training with extern dataset.

Training set Test set	**Driving**	**Munchen**	**Vaihingen**
Driving	**0.926**	0.895 (-0.031)	0.895 (-0.031)
Munchen	0.969 (-0.015)	**0.984**	0.964 (-0.020)
Vaihingen	0.980 (-0.017)	0.979 (-0.018)	**0.997**
KITTI2015	0.934	0.881	0.942
Hangzhou	0.911	0.940	0.949

Table 2. Test results (parallax error < 3pixels) on the different training datasets based on GC-net. The numbers in bracket are the difference between the current number and the diagonal number of the current row, indicating the decreasing degree of accuracy when training with extern dataset. KITTI2015 and Hangzhou datasets lack of dense disparity map and are only used for test in GC-net.

Table 1 shows the test accuracy of MC-CNN and Table 2 shows that of GC-net. The test accuracy is valued by the percent of pixels whose difference to true disparity is within 3 pixels.

When the training set and test set are from the same dataset, the test accuracy (the bold diagonal elements) can reach 95% (except for Driving dataset). The generalization ability of MC-CNN and GC-net is evaluated by the non-diagonal elements, which are obtained with the model pretrained by different training sets. The red numbers in bracket show the differences between training on target set and using pre-trained model with other datasets. If without new training samples, the accuracy of stereo matching will drop about 0.5~2% (except the virtual Driving dataset) using pre-trained models. It can be concluded that even without target training set, deep learning based stereo methods show high performance and excellent generalization ability. Nevertheless, large training samples of various scenes could be good complementary for a commercial application of deep learning based stereo methods.

4.2 Comparison of CNN and classic methods

The deep learning based methods are compared to SGM, SURE software on the aerial datasets. In Table 3, the results show the deep learning based methods are similar to (or slightly better than) the conventional methods. When the ground is flat and buildings are low, as the case of Vaihingen dataset, all methods including SGM can get a very high accuracy up to 98% and shows no obvious difference of performances between them. On the München dataset, the accuracy of SURE is 93.2% while the accuracy of MC-CNN is 96.5% trained on the target dataset (and 96.0% using model pre-trained on KITTI 2015 directly), and GC-net is 98.4% (96.9% with model transfer from the Driving dataset). However, On the Hangzhou dataset, the accuracy of MC-CNN and GC-net is about 95% whereas the accuracy of SURE is 96.8%. It experimentally proves that the deep learning based methods and conventional methods perform quite equivalent in current stage.

	KITTI2015	Driving	Hangzhou	Munchen	Vaihingen
SGM	0.893	0.713	0.896	0.921	0.987
SURE	-	-	0.968	0.932	0.990
MC-CNN	0.960/0.958	-/0.889	0.953/0.944	0.965/0.960	0.992/0.988
GC-Net	-/0.942	0.926/0.895	-/0.949	0.984/0.969	0.997/0.980

Table 3. The results of SGM and SURE on the aerial datasets (parallax error < 3 pixels).

Figure 1 displays the 3D scenes recovered from disparity maps that were generated by different methods. In Hangzhou dataset, GC-Net performed the best in visual effect and SURE showed some distortion in buildings. In München dataset, MC-CNN, GC-Net and SGM almost perform the same while SURE could produce more details. In Vaihingen dataset, SURE also show more details than the others, however, compared to the reference map, there might be some tiny errors in the flat farmland.

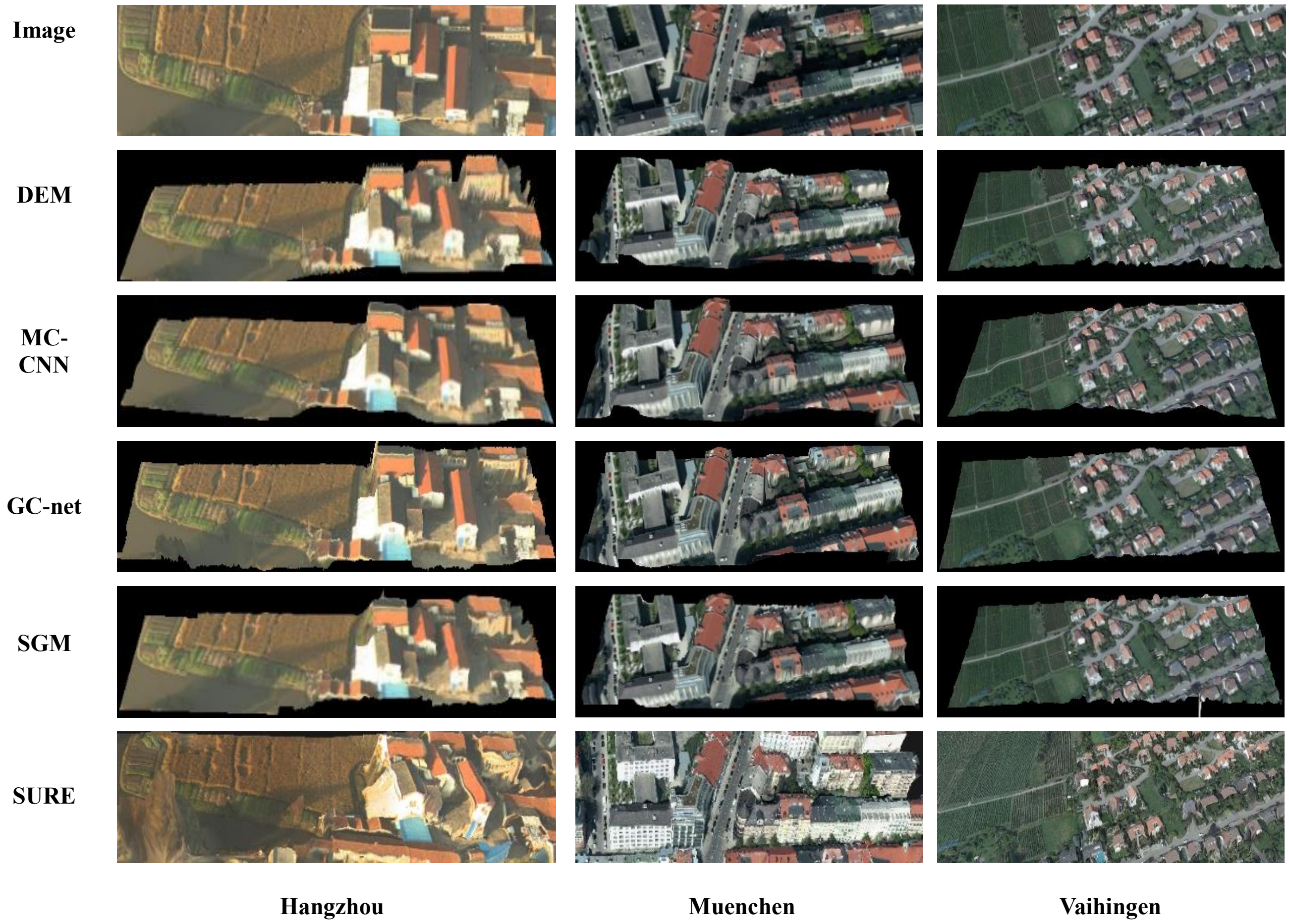

Figure 1. 3D scene recovered from dense disparity maps.

Although now deep learning based methods perform almost the same level as conventional methods, it should be addressed that deep learning based methods only leverage stereo information up-to-now and is extremely faster than conventional methods if pre-trained, while SURE utilizes multi-view geometry. It is expected the deep learning based methods could benefit largely from the introduction of the multi-view geometry constraints.

4.3 Transfer learning on aerial datasets

Transfer learning strategy is introduced to aerial image matching based on the assumption of only a few new samples available for training. We divide the datasets into a small training set and a large testing set. The pre-trained models are used as base network. All parameters are tunable with pre-trained parameters as initial values.

we tested the Hangzhou dataset using the models pretrained on

KITTI2015. We tune all parameters in the MC-CNN model and gradually increase the number of training samples from 25 stereo pairs to 300 pairs. Table 4 shows the changes of test accuracy on different sizes of training samples for direct training (DT) only on the available training set (with initial random weights) and a fine-tuning strategy with transfer learning (TL) from KITTI 2015. The accuracy of TL is 94.89% on 25 training samples, compared to 94.39% for DT. As the size of training set increases, the gain of TL slows down up to 0.1%.

Table 5 is the test accuracy on different size of training samples on München dataset based on GC-Net. DT means directly training on the target dataset with random initial weights, while TL means transfer learning with pre-trained parameters on the Driving dataset. The accuracy of TL on 25 training samples is 96.5%, compared to 78.3% of DT. As the size of training set increases, the gain of TL slows down up to 0.61% when 250 pairs are used for training.

Dataset size	25 pairs		50 pairs		100 pairs		200 pairs		300 pairs	
Method	DT	TL	DT	TL	DT	TL	DT	TL	DT	TL
Accuracy	0.9439	0.9489	0.9448	0.9485	0.9467	0.9481	0.9514	0.9526	0.9526	0.9537
Improvement	0.50%		0.37%		0.14%		0.12%		0.11%	

Table 4. The test accuracy on different size of training samples. DT means directly training on the dataset with random initial weights, while TL means transfer learning.

Dataset size	25 pairs		50 pairs		100 pairs		200 pairs		250 pairs	
Method	DT	TL	DT	TL	DT	TL	DT	TL	DT	TL
Accuracy	0.7832	0.9650	0.9024	0.9476	0.9288	0.9612	0.9593	0.9775	0.9723	0.9784
Improvement	18.1%		4.52%		3.24%		1.82%		0.61%	

Table 5. The test accuracy on different size of training samples based on GC-net. DT means directly training on the dataset with random initial weights, while TL means transfer learning.

5. CONCLUSIONS

The paper evaluates the performance of two deep learning based stereo methods, MC-CNN and GC-Net on three aerial datasets. It was experimentally proved that the two methods both can generate high accurate disparity maps both in the case of training models on target dataset and in the case of using pre-trained models on other open-source datasets. Compared to SGM and SURE, we could conclude that conventional methods and deep learning based methods perform almost the same level up-to-now whereas the latter has better potential.

ACKNOWLEDGEMENTS (OPTIONAL)

This work was supported by the National Natural Science Foundation of China (41471288).

REFERENCES

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, *47*(1-3), 7-42.

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, vol. abs/1703.04309*.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, *30*(2), pp. 328-341.

Boykov, Y. Y., & Jolly, M. P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In: *IEEE International Conference on Computer Vision*, Vol. 1, pp. 105-112.

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

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. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4040-4048.

Pang, J., Sun, W., Ren, J. S., Yang, C., & Yan, Q., 2017. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: *International Conf. on Computer Vision-Workshop on Geometry Meets Deep Learning (ICCVW 2017)*, Vol. 3, No. 9.

Shaked, A., & Wolf, L., 2017. Improved stereo matching with constant highway networks and reflective confidence learning. *CoRR, vol. abs/1701.00165*.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3061-3070.
http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo

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. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4040-4048.
https://lmb.informatik.unifreiburg.de/resources/datasets/SceneFlowDatasets.en.html

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N. 2012. SURE: Photogrammetric surface reconstruction from imagery. In: *Proceedings LC3D Workshop, Berlin*, Vol. 8, pp. 29.

https://doi.org/10.5194/isprs-archives-XLII-2-593-2018 |