DENSE MATCHING COMPARISON BETWEEN CLASSICAL AND DEEP LEARNING BASED ALGORITHMS FOR REMOTE SENSING DATA

Y. Xia*, P. d'Angelo, J. Tian, P. Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Institute, 82234 Wessling, Germany
(Yuanxin.Xia, Pablo.Angelo, Jiaojiao.Tian, Peter.Reinartz)@dlr.de

Commission II

KEY WORDS: Dense Matching, CNN, GA-Net, SGM, Census, Disparity

ABSTRACT:

Deep learning and convolutional neural networks (CNN) have obtained a great success in image processing, by means of its powerful feature extraction ability to learn specific tasks. Many deep learning based algorithms have been developed for dense image matching, which is a hot topic in the community of computer vision. These methods are tested for close-range or street-view stereo data, however, not well studied with remote sensing datasets, including aerial and satellite data. As more high-quality datasets are collected by recent airborne and spaceborne sensors, it is necessary to compare the performance of these algorithms to classical dense matching algorithms on remote sensing data. In this paper, Guided Aggregation Net (GA-Net), which belongs to the most competitive algorithms on KITTI 2015 benchmark (street-view dataset), is tested and compared with Semi-Global Matching (SGM) on satellite and airborne data. GA-Net is an end-to-end neural network, which starts from an stereo pair and directly outputs a disparity map indicating the scene's depth information. It is based on a differentiable approximation of SGM embedded into a neural network, performing well for ill-posed regions, such as textureless areas, slanted surfaces, etc. The results demonstrate that GA-Net is capable of producing a smoother disparity map with less errors, particularly for across track data acquired at different dates.

1. INTRODUCTION

Dense image matching is a key topic in the community of computer vision for stereo reconstruction. Based on a rectified stereo image pair, 3D scene information is captured via finding the correspondence and calculating the horizontal coordinate difference. The technique is widely applied on object detection and recognition, automatic driving, robot navigation, etc. (Chen et al., 2018) (Hirschmüller, 2011).

Recently, machine learning, deep learning and convolutional neural networks (CNN) (LeCun et al., 1998) are showing great success in the filed of image processing. With appropriate training datasets available, CNN based algorithms are capable of learning specific tasks to achieve very competitive results, by means of the powerful feature extraction ability. In case of dense matching, many algorithms based on deep learning have been developed, which achieve state-of-the-art, such as matching cost based on CNN (MC-CNN), Guided Aggregation Net (GA-Net), Atrous Multiscale Network (AM-Net), etc. (Du et al., 2019) (Zbontar, LeCun, 2016) (Zhang et al., 2019). The majority of them, however, are only tested on close-range or street-view stereo imagery, but not well evaluated with remote sensing datasets, including aerial and satellite data. As more high-quality datasets are available thanks to the recent airborne and spaceborne sensors, it is necessary to compare the well-performed CNN based algorithms to classical dense matching methods on remote sensing data.

Hence, in this paper, GA-Net is selected to be compared with Semi-Global Matching (SGM) (Hirschmuller, 2008) on remote sensing data. SGM is a classical stereo matching algorithm, which achieves a good balance between performance and efficiency. Therefore, it is widely applied on aerial and satellite data for digital surface model (DSM) generation (d'Angelo, Reinartz, 2011). GA-Net, on the other hand, is an end-to-end neural network which starts from an stereo pair and directly outputs a disparity map indicating the scene's depth information. It is one of the most competitive algorithms, on the top of the ranking list from KITTI 2015 benchmark (Menze et al., 2015) (Menze et al., 2018). Besides the outstanding performance, GA-Net is selected because it is based on a differentiable approximation of SGM embedded into a neural network. Thus, SGM is approximated with no parameters to be handcrafted.

In this paper, a satellite dataset from the 2019 IEEE GRSS data fusion contest (Bosch et al., 2019) (Le Saux et al., 2019) and an airborne dataset are used for the experiments. Both visualization and numerical results are provided to compare GA-Net and SGM, in order to study the performance of deep learning based algorithms for remote sensing data.

2. METHODOLOGY

Binocular stereo matching is broadly studied in computer vision, which aims at searching for corresponding pixels from a stereo pair in order to recover the depth information. Four steps are classically designed, including matching cost computation, cost aggregation, disparity calculation, and disparity refinement. Matching cost is firstly computed to perceive the similarity between pixels according to the photo consistency. In our project, a non-parametric measure Census (Zabih, Woodfill, 1994), is used to calculate the matching cost for the following SGM processing. It compares the local intensity structure around the target pixels instead of simply measuring the difference of pixel values. Therefore, the method is robust to im-

*Corresponding author

age radiometric differences and achieves high performance over depth discontinuities (d'Angelo, Reinartz, 2011) (Hirschmuller, Scharstein, 2008). Regarding the cost aggregation step, classical stereo algorithms design a penalty term to regularize and smooth disparities of neighbouring pixels. This requirement, however, is hard to be satisfied when considering the pixel neighborhood in 2D, as the disparity determination for each pixel will affect every other pixel (Bleyer, Breiteneder, 2013). Hence, SGM aggregates the cost along different 1D paths (horizontally, vertically, etc.), which are then summed up to approximate 2D smoothness. Along a path in direction r, an energy function is defined as:

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

in which $L_r(p,d)$ represents the energy for the pixel at location p assuming d as disparity. $C(p,d)$ is the matching cost. A small penalty P_1 is applied for a disparity difference of 1 pixel between p and its previous neighbor $p-r$. For larger differences, a stronger penalty P_2 is utilized. Hence, the disparity is determined according to the minimum energy, and refined by post-processing approaches, such as left-right consistency check, interpolation, etc.

Based on the same theoretical background, modern deep learning algorithms build a trainable network to extract features for matching cost, which is then processed by classical cost aggregation methods, e.g. MC-CNN (Zbontar, LeCun, 2016), or directly aggregate the cost to estimate the disparity within an end-to-end CNN, such as GA-Net, AM-Net (Du et al., 2019), (Zhang et al., 2019), etc. Among them, GA-Net proposes a semi-global guided aggregation layer (SGA) inspired by the pathwise SGM aggregation, and designs a local guided aggregation layer (LGA) to protect thin structures, which has achieved state-of-the-art performances. The SGA layer is designed as:

$$\begin{aligned} L_r(p,d) = C(p,d) + \mathrm{sum}\big(\, & w_1(p,r)\cdot L_r(p-r,d), \\ & w_2(p,r)\cdot L_r(p-r,d-1),\ w_3(p,r)\cdot L_r(p-r,d+1), \\ & w_4(p,r)\cdot \max_i L_r(p-r,i)\,\big). \end{aligned} \tag{2}$$

Compared with equation (1), all the user-defined parameters (P_1, P_2) are replaced by learnable weights w which are adaptive depending on the pixel location and the path. Thus, the algorithm is able to deal with different situations within the scene. The first/external minimum selection in equation (1) is substituted by a weighted sum, which is proved to be effective with no accuracy loss (Springenberg et al., 2014). Furthermore, the second/internal minimum selection in equation (1) is replaced by a maximum, in order to maximize the probabilities at the ground truth disparities rather than minimizing the energy. To avoid the increase of $L_r(p,d)$ along the path, $C(p,d)$ is also included within the weighted sum operation and SGA is finally formulated as:

$$\begin{aligned} L_r(p,d) = \mathrm{sum}\big(\, & w_0(p,r)\cdot C(p,d),\ w_1(p,r)\cdot L_r(p-r,d), \\ & w_2(p,r)\cdot L_r(p-r,d-1),\ w_3(p,r)\cdot L_r(p-r,d+1), \\ & w_4(p,r)\cdot \max_i L_r(p-r,i)\,\big), \\ & \sum\nolimits_{i=0,1,2,3,4} w_i(p,r) = 1. \end{aligned} \tag{3}$$

The final output of SGA is set as $L(p,d) = \max_r L_r(p,d)$. Afterwards, LGA is defined as a guided filter to recover thin structures as below:

$$\begin{aligned} L_*(p,d) = \mathrm{sum}\Big(\, & \textstyle\sum_{q\in N_p} w_0(p,q)\cdot L(q,d), \\ & \textstyle\sum_{q\in N_p} w_1(p,q)\cdot L(q,d-1), \\ & \textstyle\sum_{q\in N_p} w_2(p,q)\cdot L(q,d+1)\Big), \\ & \textstyle\sum_{q\in N_p} w_0(p,q)+w_1(p,q)+w_2(p,q) = 1, \end{aligned} \tag{4}$$

in which N_p is a user-defined neighborhood around p. Afterwards, the disparity is finally calculated by summing up the product of each disparity candidate multiplied by the corresponding probability, which is obtained via a softmax operation on the negative of the aggregated cost.

The architecture of the network is shown in Figure 1:

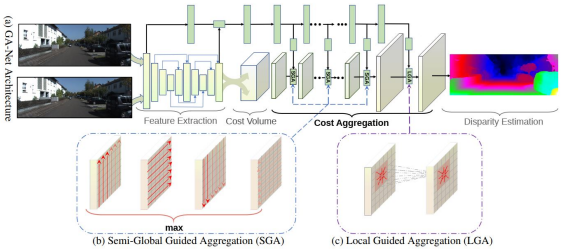

Figure 1. Overview of GA-Net (Zhang et al., 2019)

A stacked hourglass network is applied to extract features in both stereo images, from which a cost volume is created. Afterwards the guided subnetwork produces weights for SGA and LGA layers to aggregate the cost, followed by the final disparity regression.

3. EXPERIMENT

Two experiments are designed to compare SGM and GA-Net applied on remote sensing data, for which a satellite dataset from the 2019 IEEE GRSS data fusion contest and an airborne dataset are tested. Regarding SGM, we use a 7×7 window for Census, with P_1 and P_2 set as 400 and 700, respectively. A left-right consistency check is applied, and areas smaller than 100 pixels are removed to obtain smoother results. As for GA-Net, we directly use a pre-trained model on the Scene Flow dataset (Mayer et al., 2016), which is evaluated on the contest data in our first experiment. The provided ground truth disparity maps are not precise enough to train the network, due to the data inconsistency between the stereo pairs and the collected LiDAR point clouds for ground truth generation. Besides, the robustness of the model is tested when the training data and the data to be processed originate from different domains. In our second experiment, the pre-trained GA-Net model is fine-tuned on an airborne dataset covering a town in Austria, named Eisenerz. 160 stereo pairs are randomly selected, from which 80% of the images are used for finetuning the network and 20% are used for validation. For each stereo pair, a reference disparity map is available which is obtained based on a well calibrated DSM.

3.1 Experiment I

3.1.1 Data The 2019 IEEE GRSS data fusion contest dataset (Bosch et al., 2019) (Le Saux et al., 2019) contains multi-

view WorldView-3 satellite images. Both RGB (used in this paper) and 8-band visible and near infrared (VNIR) multi-spectral images are provided, covering two cities, Jacksonville, Florida and Omaha, Nebraska in United States, with a ground sampling distance of approximately 35 cm. The rectified stereo pairs are adopted for the first experiment, with pairwise ground truth disparity images generated using airborne LiDAR point cloud.

3.1.2 Results A stereo pair is tested for SGM and the pre-trained GA-Net model. The disparity maps obtained by the two algorithms are displayed in Figure 2 (see results of SGM and GA-Net-pre).

From Figure 2, it is found that GA-Net is able to produce a smoother disparity map with less error area. Regarding 3 pixels as the error limit, the accuracy of GA-Net and SGM are 90.58% and 77.10%, respectively. GA-Net achieves good model transferability when heterogeneous data are used for training and validation, which proves the feasibility of deep learning based algorithms in the field of remote sensing.

An interesting case happens to the two tennis fields in the top right quarter of the image. It is found that SGM reconstructs the tennis field on the left much better than the right one. The reason is that, the court on the right side is only visible in the master epipolar image, due to the temporal inconsistency between the stereo pair. SGM, thus, fails to find correspondence, and leads to bad visualization in the resultant disparity map. GA-Net, on the other hand, acquires guidance from the master epipolar image directly through the guidance subnetwork. Hence, the network is able to better recover the tennis field on the right. Similar result is obtained for the baseball field on the lower right of the tennis fields. SGM performs worse than GA-Net, as the field shows different intensity information between the stereo pair.

3.2 Experiment II

3.2.1 Data The EU project DRIVER + was launched in 2014, in order to develop a technical infrastructure for crisis management (Schimak et al., 2020). Regarding an earthquake scenario as a test case, the German Aerospace Center carried out an aerial campaign supported by a 3K camera system which was designed for civil protection missions, in the region of Eisenerz in Austria, from September 12 to 14, 2019. With a left-view, a nadir-view, and a right-view camera, respectively, thousands of geo-referenced aerial photos were captured, which were then rectified to epipolar stereo pairs for dense matching and DSM generation using a classical SGM implementation. During this procedure, heights from multiple overlapping stereo pairs are fused to obtain high quality output. These multi-view heights are used as ground truth for training and validation.

3.2.2 Results 160 stereo pairs collected by the nadir-view camera are randomly selected and rectified for the experiment. Starting from the GA-Net model pre-trained on the Scene Flow dataset, we use 128 epipolar stereo pairs for finetuning the network. The rest (32 pairs) of the images are used for validation. In addition to the finetuned GA-Net, the model pre-trained on the Scene Flow dataset is also used for validation as a baseline. Regarding 3 pixels as the allowed error limit, the validation accuracy for SGM, pre-trained GA-Net (GA-Net-pre) and finetuned GA-Net (GA-Net-fine) are 58.02%, 51.99% and 82.60%, respectively. A stereo pair is selected with the dense matching results visualized for a qualitative evaluation in Figure 3.

For the selected airborne dataset, the pre-trained GA-Net model is not able to surpass SGM according to the overall accuracy. It is found that the pre-trained model achieves very poor results in some mountainous area with pure vegetation, which is barely covered by the training data. However, as shown in Figure 3, GA-Net-pre can provide smoother results than SGM for images containing buildings, which is consistent with the results in Experiment I. Then with the data from the same domain to finetune the network, GA-Net-fine is capable of outperforming SGM by a large margin. The buildings are reconstructed much better by the finetuned model, even for shadow area in which SGM fails (see the buildings in the center of the image).

In addition, we also use the GA-Net model finetuned on the airborne dataset (from Experiment II), to be applied on the satellite data used in Experiment I, as shown in Figure 2 (see results of GA-Net-fine). The reason for this test is to explore the performance of GA-Net, when data from different domain but in similar appearance, e.g. both include buildings taken from overhead, are available for finetuning. The accuracy for this finetuned GA-Net is 84.05%, which is higher than SGM (77.10%) but lower than the pre-trained model (90.58%). Besides as shown in Figure 2, the resultant disparity map of the finetuned model is blurred. One possible reason is that the model is overfitted to the airborne dataset after finetuning.

4. CONCLUSION

Deep learning is proved outstanding to be applied in the filed of computer vision, due to its powerful ability to extract deep features, e.g. CNNs, with multi-scale context information considered. Numerous CNN based algorithms are developed for dense matching, however, close-range and street-view stereo data are mostly tested instead of remote sensing data. In this paper, the well-performed GA-Net is explored for satellite and airborne data, and compared with the classical SGM method. GA-Net is proved to be more robust in textureless area, and able to better handle the inconsistency between the stereo pair, which is particularly appropriate for across track remote sensing data acquired at different dates. The network can achieve similar or better performance than classical SGM (except for some extreme cases), even when the training and validation data originate from different domains. With appropriate data to finetune the network, GA-Net outperforms SGM by a large margin which demonstrates the potential of deep learning for solving remote sensing tasks.

In the future research, probabilistic deep learning is promising to be applied in dense matching, e.g. (Mehltretter, 2020), in order to estimate the uncertainty and detect erroneous depth prediction. Furthermore, self-training should also be studied, to handle remote sensing tasks without appropriate ground truth available.

5. ACKNOWLEDGEMENTS

The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest.

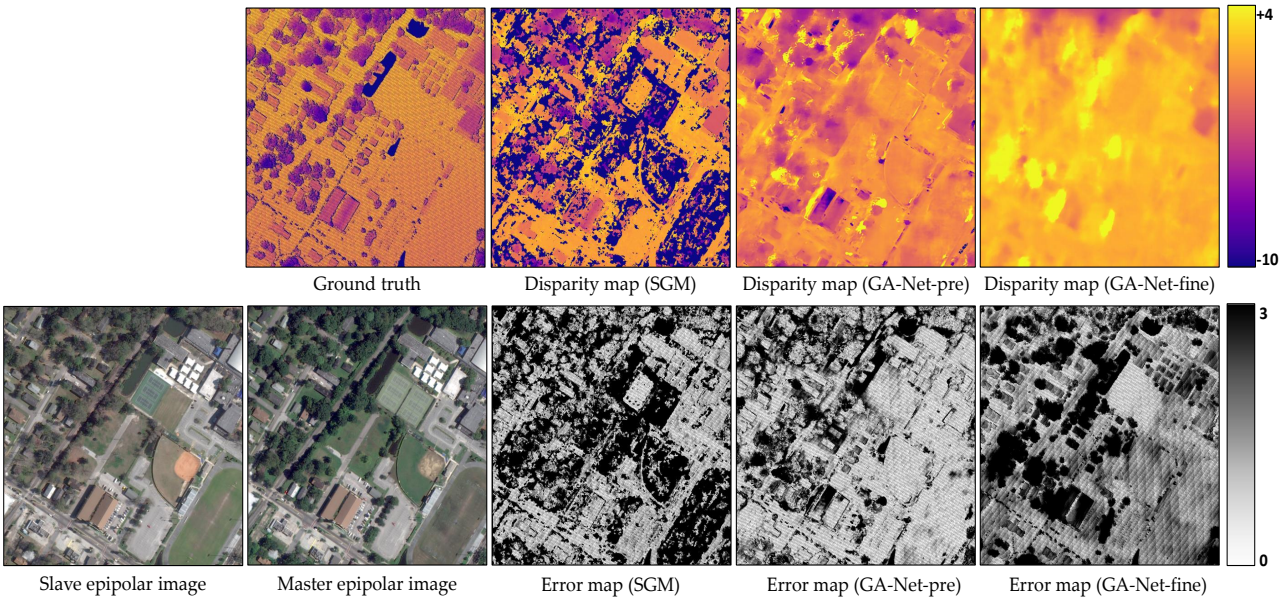

Figure 2. Experiment I: The results of SGM, pre-trained GA-Net and finetuned GA-Net from Experiment II

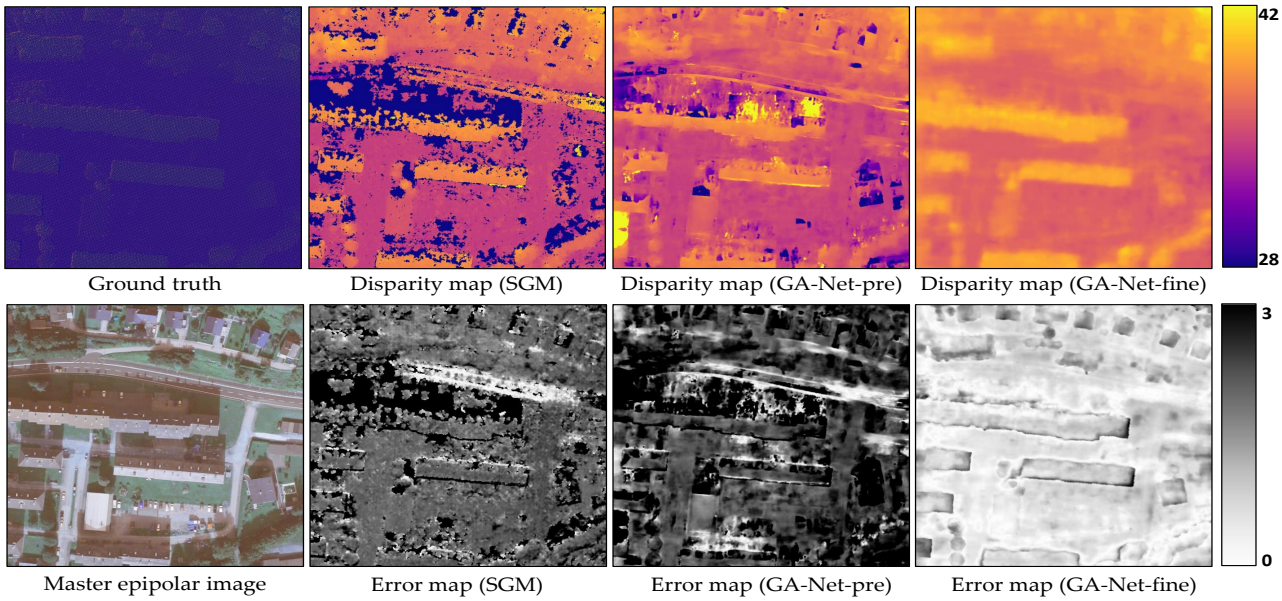

Figure 3. Experiment II: Dense matching results for a validation stereo pair

REFERENCES

Bleyer, M., Breiteneder, C., 2013. *Stereo Matching—State-of-the-Art and Research Challenges*. Springer London, London, 143–179.

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.

Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., Urtasun, R., 2018. 3D object proposals using stereo imagery for accurate object class detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5), 1259-1272.

d'Angelo, P., Reinartz, P., 2011. Semiglobal matching results on the ISPRS stereo matching benchmark. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-4/W19, 79–84. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXVIII-4-W19/79/2011/.

Du, X., El-Khamy, M., Lee, J., 2019. Amnet: Deep atrous multiscale stereo disparity estimation networks. *arXiv preprint arXiv:1904.09099*.

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., 2011. Semi-global matching-motivation, developments and applications. *Photogrammetric Week 11*, 173–184.

Hirschmuller, H., Scharstein, D., 2008. Evaluation of stereo matching costs on images with radiometric differences. *IEEE*

transactions on pattern analysis and machine intelligence, 31(9), 1582–1599.

Le Saux, B., Yokoya, N., Hansch, R., Brown, M., Hager, G., Kim, H., 2019. 2019 IEEE GRSS data fusion contest: semantic 3D reconstruction [Technical Committees]. *IEEE Geosci. Remote Sens. Mag.*, 7(1), 103-105.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. et al., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

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.

Mehltretter, M., 2020. Uncertainty estimation for end-to-end learned dense stereo matching via probabilistic deep learning. *ArXiv*, abs/2002.03663.

Menze, M., Heipke, C., Geiger, A., 2015. Joint 3d estimation of vehicles and scene flow. *ISPRS Workshop on Image Sequence Analysis (ISA)*.

Menze, M., Heipke, C., Geiger, A., 2018. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*.

Schimak, G., Ignjatović, D., Vullings, E., Sammels, M., 2020. Interoperability of solutions in a crisis management environment showcased in trial-austria. I. N. Athanasiadis, S. P. Frysinger, G. Schimak, W. J. Knibbe (eds), *Environmental Software Systems. Data Science in Action*, Springer International Publishing, Cham, 173–187.

Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

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

Zbontar, J., LeCun, Y., 2016. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17, 1–32.

Zhang, F., Prisacariu, V., Yang, R., Torr, P. H., 2019. Ga-net: Guided aggregation net for end-to-end stereo matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 185–194.