DENSE POINT CLOUD EXTRACTION FROM UAV IMAGERY USING PARALLAX ATTENTION

J. R. Bergado[1]*, F. Nex[1]

[1] Department of Observation Science, ITC, University of Twente, The Netherlands - (j.r.bergado, f.nex)@utwente.nl

KEY WORDS: UAV, point cloud, dense image matching, deep learning, parallax attention.

ABSTRACT:

Unmanned Aerial Vehicles have shown to be one of the most disruptive technologies in over the last decades—having an impact on may different applications such as environmental monitoring, disaster management, land administration, and water management. The photogrammetric pipeline is the core building block that enables researchers and practictioners to deliver UAV-related solutions for these applications. Advances in deep learning show promising results that can help improve steps within this pipeline. This study specifically investigates the use of parallax attention mechanism for improving dense point cloud extraction from UAV imagery. We experimented with three different setups of applying this network and have compared it against a semi-global matching based method. The first setup directly applies a pretrained stereo matching network, the second finetunes the pretrained network on the UAV dataset, and the third retrains the network using disparity values derived from a reference DSM of lower resolution. Results show that there could be notable improvements on the accuracy of resulting extracted point cloud when using a parallax attention stereo matching network for the dense image matching step over the conventional semi-global matching method for the case of easier stereo pair with high overlap and lower occlusion. However, there seems to be unclear improvements when dealing with stereo pairs that are highly different compared to which the networks are originally trained on, e.g. longer-baselines resulting to lower overlap and more occlusions. Furthermore, retraining with a disparity values derived from a lower resolution DSM also does not improve the resulting point cloud.

1. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) have shown to undeniably be one of the most disruptive emerging technologies over the last decades—touching many aspects in different research domains, and industrial and commercial applications (Nex et al., 2022). Within the geoinformation and earth observation science domain, the applications of UAV range from environmental monitoring to disaster management, land administration, and water management. Hardware and domain expertise aside, the core building block that enables us, researchers and practitioners, to carry out these UAV-related solutions is the photogrammetric pipeline.

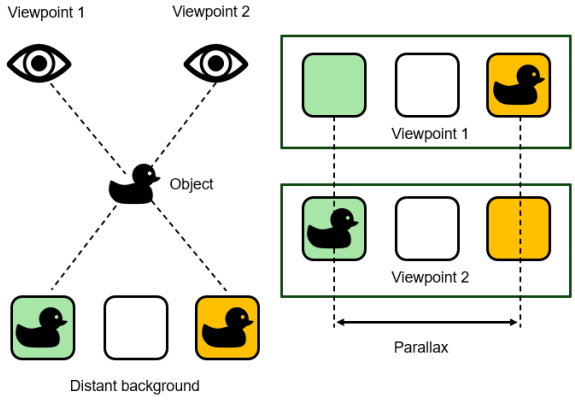

Figure 1. Illustration of the concept of parallax.

The standard UAV photogrammetric pipeline includes: 1) an image orientation step, typically done using a Structure from Motion (SfM) method (Tomasi and Kanade, 1992); followed by 2) a dense image matching step, done using a semi-global (Rothermel et al., 2012) or patch-based method; and finally 3) the generation of immediate photogrammetric data products such as orthophotos, digital surface and terrain models (DSM and DTM) and corresponding quality assessment of these products. These photogrammetric data products can then be used for several downstream mapping tasks, for example, land cover classification which could find a lot of use in different applications listed in the previous paragraph. Recent promising developments explore the use of deep network architectures in one of the steps of the photogrammetric pipeline—from learned feature matching for the image orientation step (Sarlin et al., 2020) to supervised learning of depth map penalties in the dense image matching step (Seki and Pollefeys, 2017).

The concept of parallax is central to deriving 3D information from multiple views of the same scene. It is the apparent displacement of the position of an object viewed across two different viewpoints. Figure 1 shows the same object having two different colors of a distant background when viewed between two different point of views. This apparent displacement is inversely proportional to the distance from the baseline of the two viewpoints, usually referred as the depth, the baseline being the distance connecting the two viewpoints. Hence, nearby objects will have larger parallax than farther ones. Interestingly, this is how our eyes enable us to perceive depth and estimate distances (Steinman et al., 2000).

Applied to the UAV photogrammetric pipeline, parallax facilitates the derivation of 3D geometry of the object scene from a pair of UAV images by identifying pixels corresponding to the same location in the object scene. The images first undergo through a transformation called rectification that uses the

* Corresponding author

parameters derived from the image orientation step to ensure pixels within the same rows are parallel to the image baselines. This transformation narrows down the search for corresponding pixel within one dimension. The difference between the column coordinates of corresponding pixels from the two rectified images is called the disparity. Disparity estimation is the core task in the dense image matching step of the photogrammetric pipeline usually solved by semi-global matching (Rothermel et al., 2012).

Attention mechanism (Bahdanau et al., 2014) was introduced in deep learning to allow networks to capture long-range dependencies which are often necessary to solve natural language processing tasks. This has been extended into self-attention mechanism (Fu et al., 2018), and used in computer vision tasks, to capture long-range correlation among pixels that would otherwise not be captured by the local operations, such as convolution and pooling. A more recent work (Wang et al., 2020) further modified the self-attention mechanism to only find correlation with pixels on the same rows. Applied to rectified stereo pair of images, parallax attention maps can be extracted capturing pixel correspondences in these stereo images, or feature maps derived from them. Figure 2 shows the difference between self-attention and parallax attention mechanisms. These parallax attention maps can be used to derive unsupervised loss function terms that can help improve the popularly used photometric loss (Yin and Shi, 2018).

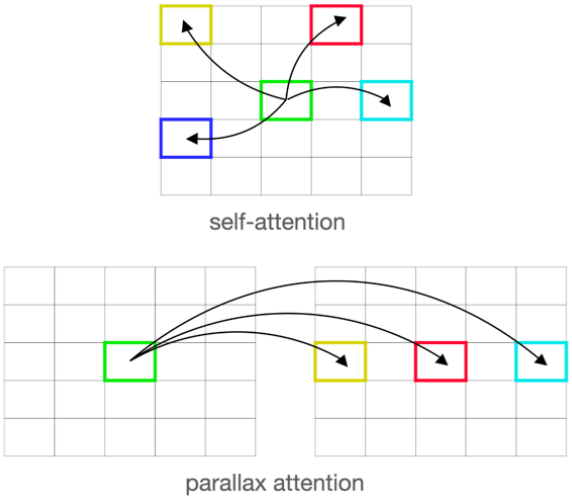

Figure 2. Comparison of self-attention mechanism and parallax attention mechanism.

This paper investigates the use of an unsupervised deep network utilizing the parallax attention mechanism for dense point cloud extraction from UAV imagery. Different setups of applying this network are evaluated and benchmarked against a baseline method based on semi-global matching. The resulting extracted dense point clouds from all the methods are compared quantitatively and qualitatively against each other.

2. DATA AND METHODS

In this study we utilized a stereo matching network with built-in parallax attention mechanism (Wang et al., 2020) to perform the dense image matching step in the UAV dense point cloud extraction pipeline shown in Figure 3. For the experiments, we used a dataset containing very high resolution nadir-looking UAV images capturing a scene of a neighborhood in a town in the central Netherlands called Nunspeet. Since the dataset does not include high resolution 3D data that can be used as a reference for evaluating our different point cloud extraction methods, we processed the dataset with the Pix4D software [1] to have some 3D UAV products, point clouds and DSM, to serve as reference outputs to compare our methods with. We refer to this dataset in the latter parts of this paper as UAV-Nunspeet.

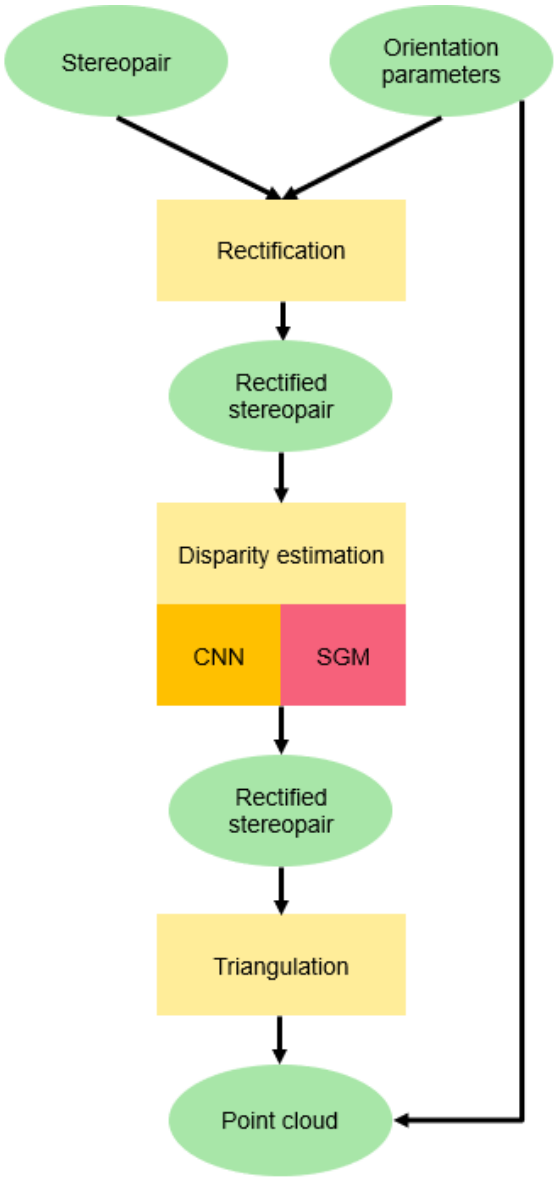

Figure 3. UAV point cloud extraction pipeline.

There are 312 images in total, the raw UAV images have a dimension of 4032 × 3024 and an average ground sampling distance of about 17 mm, while the DSM derived using Pix4D was set to have a spatial resolution of 30 mm. From this 3D product derivation, Pix4D also makes available the camera calibration and orientation parameters, so the point cloud extraction pipeline shown in Figure 3 can skip the camera calibration and image orientation steps necessary to derive these parameters. Figure 4 shows the orthophoto and DSM of the study area produced using Pix4D.

[1] https://www.pix4d.com

To fairly compare different point cloud extraction methods, we perform the same steps in the pipeline shown in Figure 3 except for the disparity estimation step. In the disparity estimation step, we compare different setups of the parallax attention stereo matching network against semi-global matching (SGM) which is one of the most widely used technique, in practice, for dense image matching (Rothermel et al., 2012).

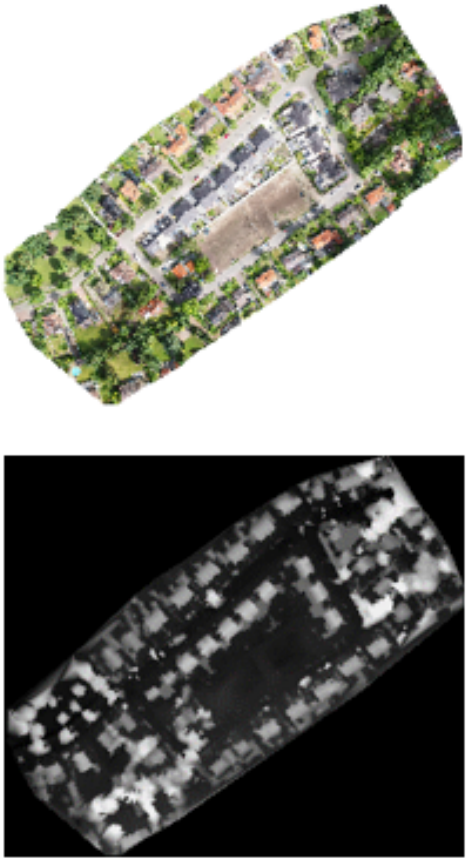

Figure 4. Study area showing orthophoto (top) and DSM (bottom) derived using Pix4D.

Figure 5 shows an overview of the parallax attention stereo matching network used in the point cloud extraction methods compared in this study. The network accepts as an input corresponding subsets of rectified stereo image pairs. Feature maps are then extracted from both the left and right images using an hourglass encoder-decoder architecture similar to the widely-used SegNet (Badrinarayanan et al., 2015), the main difference being convolution with a stride of 2 is used rather than max-pooling to downsample the feature maps and transposed convolutions are used to upsample it back. The multi-scale feature maps are then fed to parallax attention modules at three different scales, each module producing an output (unrefined) disparity map and validity mask at that scale. The resulting disparity map from the largest scale is then fed to another hourglass architecture performing a learned disparity refinement step.

Three different setups of the parallax attention stereo matching network were tested in the experiments. First is applying the pretrained network (Wang et al., 2020) initially trained on SceneFlow dataset (Mayer et al., 2016) and finetuned, in an unsupervised manner, on KITTI dataset (Geiger et al., 2012). We refer to this first setup in the following sections of the paper as $pretrained$ method.

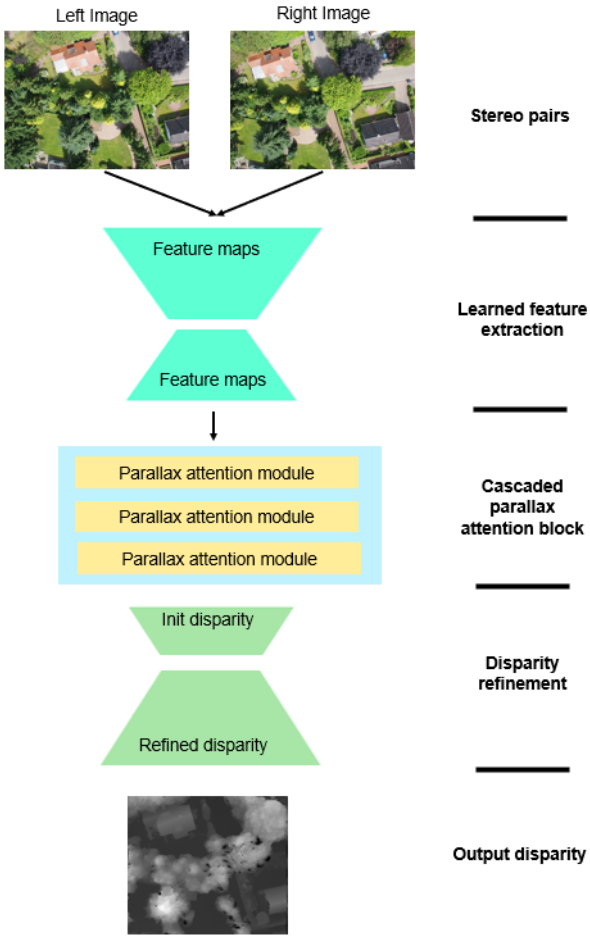

Figure 5. Overview of the parallax attention stereo matching network.

The second setup was using the network in the first setup fine-tuned, again in an unsupervised manner, on the UAV-Nunspeet dataset. For this, the raw images were first rectified as schematically shown in Figure 3. The rectified stereopairs are then divided into 960×540 smaller patches. Since the disparity range can have a relatively high minimum value for stereopairs with longer baselines, the range was shifted—reducing with a constant value, estimated using the mean elevation of the ground control points—effectively setting the disparity corresponding to points with this elevation closer to zero.

The network was trained using the same unsupervised loss as introduced in the original parallax attention stereo matching network (Wang et al., 2020). The loss $\mathcal{L}$ given by:

$$\mathcal{L}_{unsup} = \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_a (0.2\mathcal{L}_a^1 + 0.3\mathcal{L}_a^2 + 0.5\mathcal{L}_a^3) \quad (1a)$$

$$\mathcal{L}_a^s = \mathcal{L}_{a-p}^s + \lambda_{a-s}\mathcal{L}_{a-s}^s + \lambda_{a-c}\mathcal{L}_{a-c}^s \quad (1b)$$

has three components: i) a photometric loss $\mathcal{L}_p$ and ii) a smoothness loss $\mathcal{L}_s$ (Yin and Shi, 2018), as well as iii) parallax attention mechanism loss $\mathcal{L}_a^s$ calculated in three scales $s = (1, 2, 3)$ (Wang et al., 2020). Furthermore, $\mathcal{L}_a^s$ also has three terms: i)

a photometric $\mathcal{L}^s_{a-p}$ and ii) smoothness loss $\mathcal{L}^s_{a-s}$ calculated from the parallax attention maps, and a iii) a cycle loss$\mathcal{L}^s_{a-c}$ with corresponding weight term λ. During finetuning, λ_{a-s}, λ_{a-c}, λs, and λa are set to 1, 1, 0.5, and 0.5. The network was trained setting the initial learning rate to 2×10^{-4} for 15 epochs and decreased 2×10^{-5} for 5 more epochs.

The third setup is the pretrained network retrained, in a supervised manner, on disparity values derived from the DSM extracted using Pix4D. For the following parts of the paper, we call this setup $retrained$. Reference disparity values are calculated by backprojecting each pixel in the DSM to the rectified left and right stereo image pair systems, using the orientation parameters. After these two backprojected points are identified in each of the rectified image, the disparity values will just be the difference in column coordinates of these corresponding points.

This network is retrained using the same 960 $\times$ 540 patches, deriving disparity maps for each patch as described above. A total of 4484 stereo pairs with corresponding derived disparity map was used in training. The network was trained with an initial learning rate of 2×10^{-4} for 40 epochs and decreased 2×10^{-5} for 40 more epochs.

All the networks were optimized using Adam (Kingma and Ba, 2015). During test, i.e. evaluation of the point cloud extraction pipeline, the networks uses an occlusion mask derived from $s = 3$ parallax attention maps. Experiments were run using multiple PC equipped with NVIDIA Titan Xp and NVidia GeForce RTX 2070 GT.

SGM-based point cloud extraction method was implemented using the Pandora framework (Cournet et al., 2020). Multiscale option was used using 3 levels, Census is used as a matching cost, with window size set to 5 pixels, and disparity range limited to -384 to 384.

The Pandora framework provides validity mask serving as a proxy for detecting and excluding occluded pixels that are only visible in one of the image pair but not the other. This validity mask is used to exclude occluded pixels before triangulating the resulting point cloud. Similarly, the parallax attention stereo matching network provides validity maps derived from the learned parallax attention maps but the map with the highest resolution has only 1/4 of the resolution of the original input images. Hence, they are upsampled back to the original input resolution and used to mask pixels before the triangulation step.

3. RESULTS

Method	Pair I		Pair II	
	$\lvert\mu\rvert$	σ	$\lvert\mu\rvert$	σ
SGM	0.072	0.651	0.194	**0.782**
$pretrained$	0.026	0.635	**0.124**	0.867
$finetuned$	**0.021**	**0.617**	0.190	0.799
$retrained$	0.116	0.715	0.347	1.277

Table 1. Error statistics (mean μ and standard deviation σ) of the point cloud generation methods on two test image pairs.

The error statistics of the four different point cloud generation methods assessed on two test stereo image pairs, from four different images, are shown in Table 1. Both the absolute value of the mean $|\mu|$ and standard deviation σ of the resulting point cloud's distance, in meters, from the reference mesh are shown for two test pairs of images. The lower the absolute values for both μ and σ, the better the results. From Table 1, we can see that the two methods, pretrained and finetuned, based on the parallax attention stereo matching network quantitatively performs better than the baseline SGM and the last method utilizing the same stereo matching network, retrained. This implies that we can further improve the point cloud extracted from UAV imagery when using a pretrained or a finetuned, in an unsupervised manner, parallax attention stereo matching network compared to a widely used dense image matching method like SGM.

There is also a noticeable drop in accuracy on the results of all the methods from first image pair to the second image pair. This could be attributed to the fact that the second image pair has a notably lower overlap compared to the first image pair and also has a much larger presence of tall trees, resulting to more occlusions.

The method based on unsupervised finetuning of the parallax attention stereo matching network performed slightly better—about 5 mm on average, which is less than one third of the average ground sampling distance—than the method based on directly applying the same network pretrained on the KITTI dataset on the first, relatively easier, image pair. On the second image pair, however, the pretrained network has significantly lower average distance—about 66 mm on average, which is almost four times the average ground sampling distance—but has higher σ, with at least the same difference as observed in the average distances, compared to both SGM and finetuned. This results show that unsupervised finetuning of the parallax attention stereo matching network can marginally improve the resulting extracted point cloud for easier stereopairs (large overlap and less occlusions) compared to using the same network finetuned on another dataset. However, for more difficult stereopairs (less overlap and more occlusions) that are significantly different compared to the images that the stereo matching network was originally trained on, it is not clear whether there is an improvement in the accuracy of the resulting extracted point cloud using a pretrained or finetuned parallax attention stereo matching network for the disparity estimation step compared to using SGM. This is shown by the fact that both pretrained and finetuned methods have lower absolute values of μ but have higher σ than the SGM.

Retraining the parallax attention stereo matching network, in a supervised manner, using disparity values derived from the reference DSM does not seem to perform well on both the image pairs we tested. This could imply that the quality and sparsity, due to being derived from a lower spatial resolution DSM, of the derived reference disparity values is not enough to further tune the network to improve the quality of the resulting extracted point cloud.

Figure 6 shows the error maps of all the point cloud extraction methods except retrained. Areas where the extracted point cloud is above the reference mesh are shown in red and areas where the extracted point cloud is below the reference mesh are shown in blue. Green areas show locations where the extracted point cloud aligns well with the reference mesh. Majority of the points fall under the green areas, most of the significantly deviating points are in the red areas which correspond to canopies of tall trees with large basal area. The pattern of the deviations looks similar for all the three methods shown except for a notably larger red patch in the southeastern quadrant of the eastern stereopair for SGM.

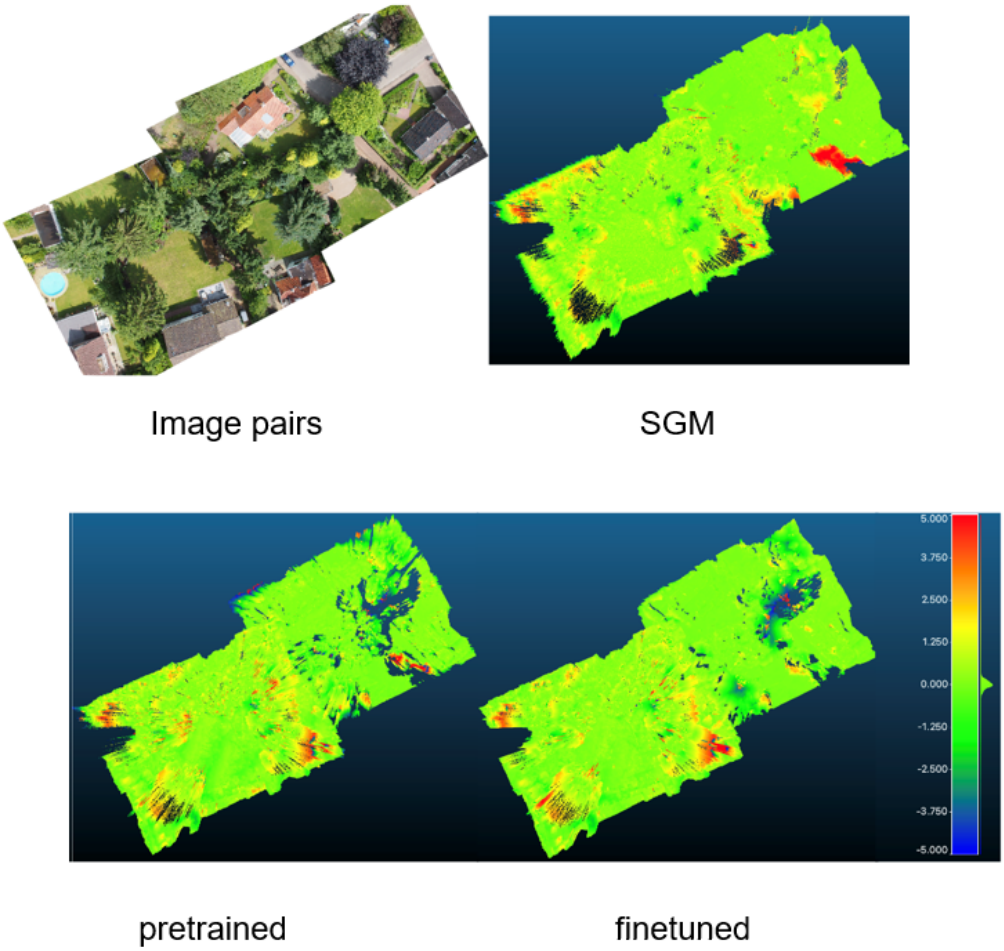

Figure 6. Error maps of the 3D reconstruction methods.

4. CONCLUSION

The photogrammetric pipeline is the core building block that enables researchers and practitioners to deliver UAV-related solutions. Improving the accuracy and efficiency of this pipeline could greatly benefit relevant downstream geoinformatics task that depends on UAV imagery and corresponding 3D data products derived from these images. Several advances in deep learning has promising value in improving the photogrammetric pipeline, specifically the dense image matching step.

This paper shows that there could be notable improvements on the accuracy of resulting extracted point cloud when using a parallax attention stereo matching network for the dense image matching step over the conventional semi-global matching method for the case of easier stereo pair with high overlap and lower occlusion. However, there seems to be unclear improvements when dealing with stereo pairs that are highly different compared to which the networks are originally trained, e.g. longer-baselines resulting to lower overlap and more occlusions. Furthermore, retraining the network in a supervised manner with derived, sparse, and possibly lower quality reference data does not help the quality of the extracted point cloud.

Future works for further improvements will be to use a dataset with high resolution 3D data that can be used directly for training/testing purposes. Integration of semantic information to improve the 3D reconstruction results and coming up with ways to better deal with occlusions.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *CoRR*, abs/1511.00561.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cournet, M., Sarrazin, E., Dumas, L., Michel, J., Guinet, J., Youssefi, D., Defonte, V., Fardet, Q., 2020. GROUND TRUTH GENERATION AND DISPARITY ESTIMATION FOR OPTICAL SATELLITE IMAGERY. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 127–134. https://isprs-archives.copernicus.org/articles/XLIII-B2-2020/127/2020/.

Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H., 2018. Dual Attention Network for Scene Segmentation. *CoRR*, abs/1809.02983. http://arxiv.org/abs/1809.02983.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. 3354-3361.

Kingma, D. P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980.

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. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nex, F., Armenakis, C., Cramer, M., Cucci, D., Gerke, M., Honkavaara, E., Kukko, A., Persello, C., Skaloud, J., 2022. UAV in the advent of the twenties: Where we stand and what is next. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184, 215-242.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. Sure: Photogrammetric surface reconstruction from imagery.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4937-4946.

Seki, A., Pollefeys, M., 2017. SGM-Nets: Semi-Global Matching with Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6640-6649.

Steinman, S., Steinman, B., Garzia, R., 2000. *Foundations of Binocular Vision: A Clinical Perspective*. McGraw-Hill Education.

Tomasi, C., Kanade, T., 1992. Shape and motion from image streams under orthography: a factorization method. *Int. J. Comput. Vis.*, 9(2), 137-154.

Wang, L., Guo, Y., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., 2020. Parallax Attention for Unsupervised Stereo Correspondence Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4), 2108-2125.

Yin, Z., Shi, J., 2018. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 1983–1992.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-1027-2023