DEEP LEARNING FOR MONOCULAR DEPTH ESTIMATION FROM UAV IMAGES

L. Madhuanand [1*], F. Nex [1], M. Y. Yang [1]

[1] Dept. of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation(ITC), University of Twente, Enschede, The Netherlands –
l.madhuanand@student.utwente.nl, f.nex@utwente.nl, michael.yang@utwente.nl

Commission II, ICWG I/II

KEY WORDS: Depth, Monocular, Aerial images, Disparity, Deep learning, Image reconstruction, Scene Understanding

ABSTRACT:

Depth is an essential component for various scene understanding tasks and for reconstructing the 3D geometry of the scene. Estimating depth from stereo images requires multiple views of the same scene to be captured which is often not possible when exploring new environments with a UAV. To overcome this monocular depth estimation has been a topic of interest with the recent advancements in computer vision and deep learning techniques. This research has been widely focused on indoor scenes or outdoor scenes captured at ground level. Single image depth estimation from aerial images has been limited due to additional complexities arising from increased camera distance, wider area coverage with lots of occlusions. A new aerial image dataset is prepared specifically for this purpose combining Unmanned Aerial Vehicles (UAV) images covering different regions, features and point of views. The single image depth estimation is based on image reconstruction techniques which uses stereo images for learning to estimate depth from single images. Among the various available models for ground-level single image depth estimation, two models, 1) a Convolutional Neural Network (CNN) and 2) a Generative Adversarial model (GAN) are used to learn depth from aerial images from UAVs. These models generate pixel-wise disparity images which could be converted into depth information. The generated disparity maps from these models are evaluated for its internal quality using various error metrics. The results show higher disparity ranges with smoother images generated by CNN model and sharper images with lesser disparity range generated by GAN model. The produced disparity images are converted to depth information and compared with point clouds obtained using Pix4D. It is found that the CNN model performs better than GAN and produces depth similar to that of Pix4D. This comparison helps in streamlining the efforts to produce depth from a single aerial image.

1. INTRODUCTION

Depth is an important component for understanding 3D geometrical information of objects from a 2D scene. It can be used for improvements in scene understanding tasks like semantic labelling, object recognition, topography reconstruction etc., (Chen et al., 2018). In Photogrammetry, depth is extracted using stereo images that are acquired from different camera positions by visualising the same portion of the scene. The camera calibration parameters along with the parallax of the images will be used to estimate the depth from the stereo pairs (Kang et al., 1999). However, acquiring multiple images covering the same scene with sufficient base may not be possible for complex terrains/environments. This, along with the necessity to extract information from uncalibrated cameras has led to the increased research and development of computer vision techniques for depth estimation. Different techniques that combine computer vision and photogrammetry had been developed for this task. Among that, deep learning has wide range of applications in scene understanding, segmentation, classification and depth estimation tasks (Luo et al., 2018). This successful performance of deep learning techniques in extracting high level features and its applications, makes it a preferable tool for single image depth estimation (Amirkolaee and Arefi, 2019). There are multiple approaches through which single image depth estimation can be achieved. This includes supervised learning where the models are trained with ground truth depths (Eigen et al., 2014; Liu et al., 2016; Laina et al., 2016; Li et al., 2017; Mou and Zhu, 2018; Amirkolaee and Arefi, 2019). The collection of ground truth depths is a time consuming, expensive and difficult process for complex outdoor scenes (Amiri et al., 2019). Another approach involves the use of semantic or other useful information (other than depth) to enhance the depth estimation (Jafari et al., 2017; Ramirez et al., 2018; Amiri et al., 2019; Chen et al., 2019). Although labelled semantic information is easier to obtain than ground truth depth, it is still an added complexity. To overcome the inherent difficulties in these methods, an alternative is given by estimating the depth without ground truth depths and semantic information. This involves using stereo views to learn depths by reconstruction of images during the training stage in a self-supervised manner and then using the trained model to find depth from single images (Godard et al., 2017; Repala and Dubey, 2018; Pilzer et al., 2018; Aleotti et al., 2018). Though these methods have proven to decrease the ambiguity in depth estimation from a single image, they have been applied only on indoor or outdoor scenes taken at ground level. The single image depth estimation from aerial images have been very sparse due to its increased viewpoint complexity and lower resolution images. The increased usage of Unmanned Aerial Vehicle (UAVs) and its widespread availability has made the collection of high-resolution aerial images affordable. This has led to the prevalent use of UAV platforms especially for 3D modelling or 3D digital elevation models (Nex and Remondino, 2014). However, as discussed earlier, it is not always possible to extract stereo pairs and hence the techniques for single image depth estimation at ground level should be extended to aerial images taken from UAVs. Also, such methods could be useful for certain monitoring tasks which does not require acquisition of classical photogrammetric image block and can use the depth estimated from single images with a reduced quality. Besides, there can also

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-V-2-2020-451-2020 | © Authors 2020. CC BY 4.0 License.

be other applications like object detection and tracking where there is a need to know the distance of the objects in complex environments.

The depth estimation uses cues like shading, occlusion, perspective, texture variations and scaling of objects to differentiate and understand the scene (Godard et al., 2017). Also, Hu et al., (2019) suggested the importance of edges for the deep learning models for grasping the geometry of the scene. There are different deep learning models available to estimate depth from monocular indoor images. Amongst them, the models that use stereo pairs to learn depth by reconstruction are comparable in performance with those that use pixel-wise ground truth depth (Godard et al., 2017). Since stereo images are much more accessible than ground truth depths, it is desirable to use models which learn using stereo pairs. Applying such models to aerial images is necessary to understand the required architecture for complex tasks. Aerial images differ from ground-level images, as they may lack information like shading, texture etc., making it more difficult for learning process than ground level images. It's varied perspectives along with its increased distance from the camera point also makes this a challenging task to be addressed. This necessitates the use of deep learning models that can overcome these shortcomings and be able to learn to estimate depth from aerial images. To test that, two deep learning models are used. The first one is a simple Convolution Neural Network (CNN) architecture as proposed by Repala and Dubey, (2018) and the second one is a complex adversarial learning using Generative Adversarial Neural Networks (GAN) as proposed by Aleotti et al., (2018). The importance of choosing the right model can help in increasing the accuracy of depth estimation from aerial images.

Section 2 of this paper explains the UAV images dataset prepared for the training process. The architecture of the deep learning models used in this study is described in section 3 and the evaluation of the test images along with the results and discussions are presented in section 4. The conclusions and the scope for future works are provided in section 5.

2. DATASET

The dataset consists of a collection of different sets of high-resolution aerial images captured by UAVs over various land use/landcover features. The number of images from each region and the number of patches extracted from these images is given in Table 1. All the images have been selected from large image blocks: in particular, adjacent images along the same strip have been used to maximize their overlapping area. The pre-processing involved the generation of undistorted images from the UAV datasets and then image rectification to produce stereo pairs (Monasse et al., 2010). By computing the rectification transformations, the images can further be transformed such that the corresponding points lie along the same rows. The generated stereo pairs are both undistorted and rectified for computing precise depth information. The accuracy of depth estimation are limited by the quality of the stereo images produced for the reconstruction of disparities (Amiri et al., 2019). The errors generated during the rectification and stereo pair generation might get accumulated and carried through the model, which significantly could affect the quality of the generated disparity maps.

Dataset	Average GSD (cm)	Full Images	Image patches
EPFL Quartier Nord, Switzerland	3.5	100	1500
Rwanda, East Africa	3.01	950	17120
Zeche zollern, Germany	2.05	300	4500

Table 1. Number of training images-stereo pairs along with extracted patches

In total 1300 stereo pairs are generated from the available UAV images. The images in the dataset are taken from a mixture of nadir and oblique view, with predominantly nadir view images. The Ground sampling distance (GSD) of the stereo pairs are around 2-3 cm and the average forward overlap is 80%. Due to its high resolution, the size of the aerial images are large. The images are divided into smaller patches of size as per the admissible input size of the CNNs to be used without reducing the resolution. The patches from the same position from the left and right images are matched to produce the patch stereo pairs. The total number of pairs of image patches generated are 22600. From this 22000 patch pairs are used for training while 600 single image patches are used for testing. A sample stereo pair along with the extracted patches is shown in Figure 1.

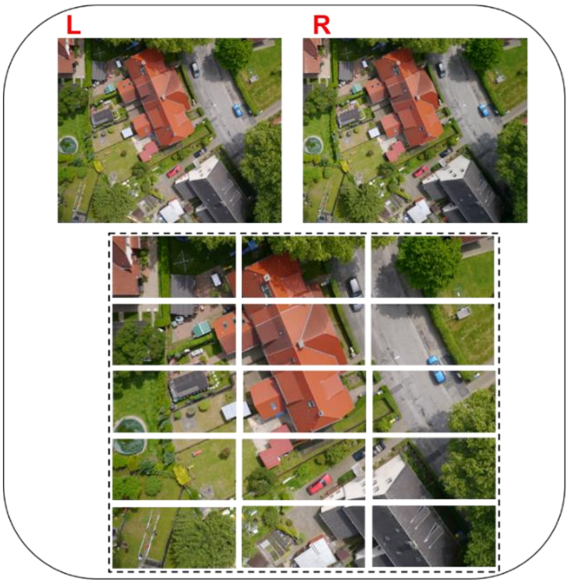

Figure 1. a) Training images-Stereopairs along with extracted patches from left image

3. METHODOLOGY

Depth can be perceived using monocular or stereo cues. Monocular cues include contextual information like object texture variations, gradients, shading etc., (Saxena et al., 2007). Knowing monocular cues alone will not help in obtaining the depth as they deal only with local variations and hence difficult to estimate accurate depth. Stereo cues capture different views of the same object and can be used for 3D reconstruction. The difference between the corresponding points from the left and right pairs forms the disparity map of the image. The disparity is inversely proportional to the object distance from the viewpoint and can be used to calculate the depth variations as shown in equation (1) (Kang et al., 1999).

$$Disparity = Xl - Xr = \frac{Bf}{d} \quad (1)$$

where X_l and X_r denote the corresponding image points, B represents the baseline distance between cameras, f is the camera constant and d is the depth or object distance from the viewpoint. In this study, disparity maps which contain depth information are generated using deep learning techniques.

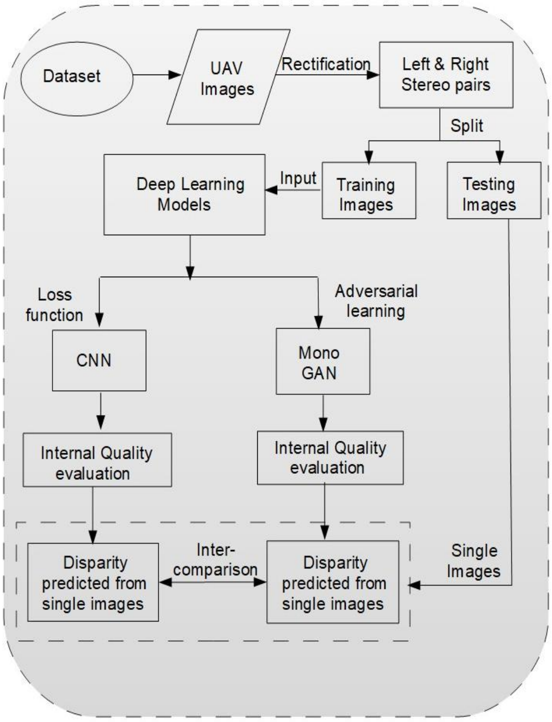

Figure 2. General workflow of the performed tests

To begin with, the UAV images are pre-processed to generate left-right stereopairs. The single image depth estimation problem is treated as an image reconstruction problem, using the encoder-decoder deep CNN model. The model takes only the left image from the stereo pair to produce disparity, which is warped with the right image through bilinear sampling to reconstruct the left image. The difference between the reconstructed left image and the input left image is calculated as a loss. The model backpropagates the loss and learns to produce better disparity from the single left image. This is the general approach of the deep learning models used in this study for learning disparity in an unsupervised manner. Two models - CNN and GAN are trained using the dataset to produce disparity from single image. The detailed methodology for each model is explained in section 3. The internal qualities of the models are evaluated and the disparity images generated from the test images are inter-compared. This helps in understanding the relative performance of different architectures for such ill-posed problems. The process is explained in the flowchart given Figure 2.

3.1 Models Used

In this study, the performance of CNN in estimating depth from a single aerial image is compared with that of GAN. As CNN has been successfully demonstrated for many image reconstruction tasks it is chosen for this study. On the other hand, many studies report that adversarial learning can improve the performance of image generation tasks and hence a GAN model is chosen.

CNNs are deep networks that have been widely applied to various tasks like image classification, detection, semantic segmentation and other computer vision applications (Bhandare et al., 2016). The use of these networks for the depth estimation problem has increased due to its proven success in other domains. This deep learning architecture is tested using the aerial images taken by UAVs. The introduction of GAN by Goodfellow et al., (2016) proved to be an active area of research for such complex problems. The GAN consists of a generator that learns to produce realistic images and discriminator that learns to find the difference with real images. Mehta et al., (2018) introduced structured adversarial training for predicting depth from synthesised stereo image pairs. Many other developments in adversarial learning led to different network modifications like MonoGAN (Aleotti et al., 2018), Cycle GAN (Pilzer et al., 2018), Pix2Pix (Julian et al., 2017) and other adversarial frameworks (Chen et al., 2018). The adversarial learning models marks the current state of the art in many areas where deep learning is being used. Almost all these models are implemented on images taken from a fixed viewing angle on the ground level in contrast to the aerial images captured by UAVs. Also, the distance from the point of view in aerial images is much higher as compared to the datasets in which these models are trained. Due to larger camera distance in aerial view compared to ground level images, absolute values of disparity are much lesser and hence finding local variations in depth are complicated. The models have to be appropriately modified to accommodate the differences brought in by the aerial image dataset. Dual CNN proposed by Repala and Dubey, (2018) and MonoGAN proposed by Aleotti et al., (2018) have been used in this study for inter-comparison as their model produced better accuracy for benchmark KITTI dataset.

3.1.1 Dual CNN: Repala and Dubey, (2018) successfully demonstrated the use of CNNs to estimate depth from single images. They utilised two CNN architectures for left and right stereo images. During the training phase, the left image was given as an input to left CNN (CNN-L) to produce left disparity and the right image was given as an input to right CNN (CNN-R) to produce right disparity as shown in Figure 3. The left and right images are then reconstructed using bilinear sampling with the opposing disparity maps. For instance, the left disparity image, generated from the left CNN was warped with the right image to reconstruct the left image as output and similarly, the right disparity image, generated from right CNN was warped with the left image to produce a right image. The reconstructed left and right images were compared with the original input images to calculate the losses.

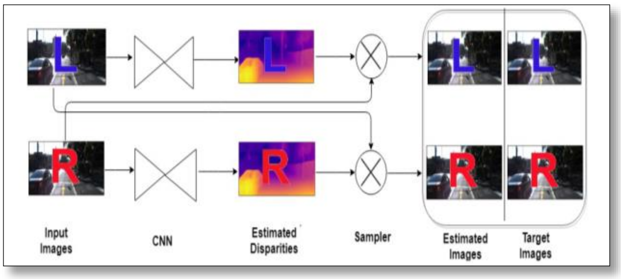

Figure 3. Dual CNN with 6 losses. Adapted from “Dual CNN Models for Unsupervised Monocular Depth Estimation”, by Repala, V. K., & Dubey, S. R. (2018)., p.3.

The three types of losses used for comparison were, matching loss, disparity smoothness loss and left-right consistency loss for both CNN architecture. The matching loss estimates whether generated image was similar to the original image, disparity smoothness loss was to ensure the generated disparity maps to be smooth by calculating the gradients and the left-right consistency loss was to check whether the generated disparity maps were consistent for left and right images. The loss terms was calculated and back-propagated to improve network performance. This forms the main structure of the Dual CNN with 3 pairs of losses (3 for the left image and 3 for the right image).

Repala and Dubey, (2018) reported an RMSE value between the estimated depth maps and ground truth depth maps of 6.162 pixels before post-processing on KITTI (Geiger et al., 2012) driving dataset with the use of 6 losses.

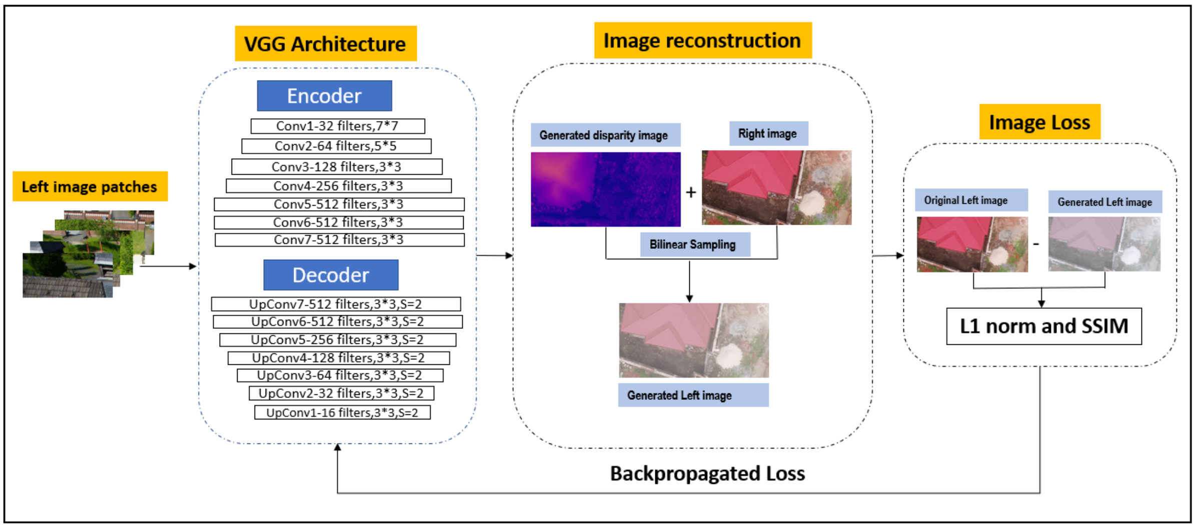

Figure 4. Simple CNN architecture with Image reconstruction loss

Implementation – CNN

Compared to Repala and Dubey, (2018), a single CNN for left image with a VGG based network architecture is utilised as it had less number of computational parameters. The VGG network consists of an encoder and decoder structure for generating the feature maps. The encoder consists of 7 convolutional layers with the increasing number of filters to extract the information. The decoder also consists of 7 convolutional layers with decreasing number of filters to reach the original input size. During training phase, the left image is sent through the encoder for extracting the features through downsampling and back to original input size through the decoder for upsampling which is then warped with the right image to produce a warped left image. In order to improve the reconstruction of the images, a simple loss to calculate the appearance mismatch between the generated left image and the raw left image is included. The model architecture and the image reconstruction process is shown in Figure 4. Repala and Dubey, (2018), utilised three losses for the reconstruction of images, however the left-right consistency loss is not meaningful as there is only left CNN in present study. A single image loss is then used for this study as given in equation (2) which compares the generated left image with the original left image. This image loss is in simpler terms a combination of L1 norm and Structural Similarity Index Metric (SSIM) for left and right images as shown below.

$$Image\ Loss = \frac{1}{N}\sum_{i,j}\alpha\frac{(1-SSIM(I_{i,j}^{\beta},\hat{I}_{i,j}^{\beta})}{2} + (1-\alpha)||I_{i,j}^{\beta} - \hat{I}_{i,j}^{\beta}|| \qquad (2)$$

Where α represents the weight between L1 norm and SSIM, I denotes the original image and the Î represents the warped image, β={l,r} for left and right images and i,j represents pixel position. This specific architecture with a single left CNN and a single loss for back propagation is found to be optimal to reach convergence. Using a single CNN instead of two also makes it more realistic to compare results from this model with that of the GAN model.

To compute the gradients Adam optimizer is used due to its faster convergence compared to stochastic gradient descent. From experimental tests, the number of epochs is fixed as 70 and the learning rate is fixed as 10^{-5} decaying to half that value at the end.

3.1.2 MonoGAN: Aleotti et al., (2018) proposed an architecture consisting of a generator and discriminator network jointly trained through adversarial learning for reconstructing disparity map in a cycle. The generator takes as input the left stereopair and generates a disparity image. This generated disparity image was then warped with the right image through bilinear sampling to synthesize a left image. The discriminator tries to distinguish between the generated left image and the original left image, producing a discriminator loss. The general architecture of the model is shown in Figure 5. The total loss was the sum of the generator loss and discriminator loss denoting the min-max game between the two. Min-max refers to minimising generator loss and maximising discriminator loss simultaneously (Goodfellow et al., 2016).

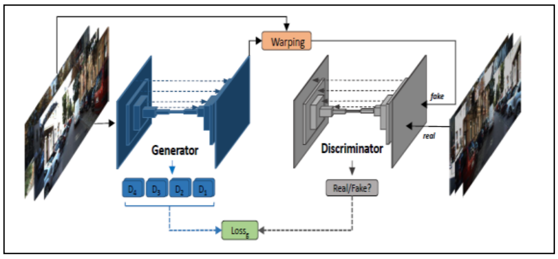

Figure 5. MonoGAN for stereo depth estimation. Adapted from "Generative Adversarial Networks for unsupervised monocular depth prediction", by Aleotti et al., (2018).

The generator will compete with the discriminator to reconstruct better disparity maps and the discriminator will try to increase the probability of distinguishing between the original and generated images. Aleotti et al., (2018) reported the RMSE values between estimated depth map and ground truth data as around 5.998 pixels on KITTI dataset using the monoGAN.

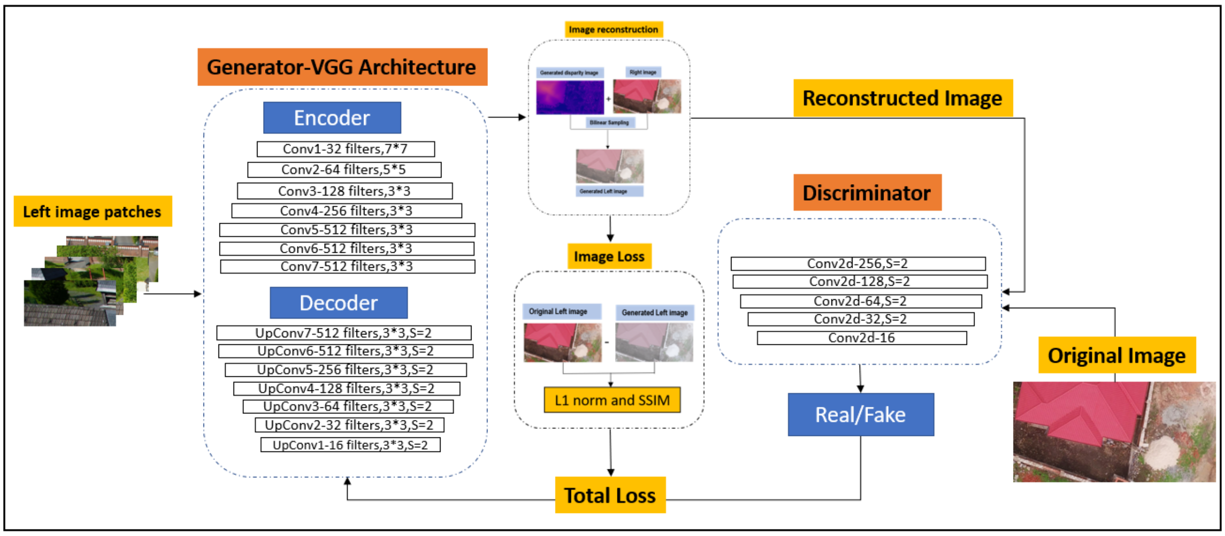

Figure 6. GAN architecture with Generator and Discriminator loss

Implementation – GAN

VGG based network architecture similar to the one described in section 3.1.1.1 was used for the generator network for feature map generation. The task of the discriminator is to distinguish between the real and fake images which is much easier compared to the generator which has to reconstruct images. Hence the discriminator has a simpler architecture with less number of feature maps generated from each of its layers. The discriminator consists of a set of 5 convolutional layers which reduces the size of the input image by a factor of 2. Both the generator and discriminator are trained simultaneously. The generated left image and the original left image is compared by the discriminator. The higher probability of identifying the generated images from the original images by the discriminator makes the generator to increase its performance in generating more realistic images. The total loss in this structure is the sum of the generator and discriminator loss as shown in equation (5). The generator loss is the combination of images loss and the probability of identifying the generated image as fake by discriminator as given in equation (3). While the discriminator loss is the probability that the original image and generated image is classified accordingly as given in equation (4).

$$Generator\ Loss\ =\ \text{Image Loss}\ +\ \alpha_{i,j} * E_{\hat{I}}(log(D(\hat{I}))) \tag{3}$$

$$Discriminator\ Loss = -1/2[E_I(log(D(I)))] - 1/2[E_{\hat{I}}(log\left(D(1-\hat{I})\right))] \tag{4}$$

$$Total\ Loss = Generator\ Loss + W_d * Discriminator\ Loss \tag{5}$$

where the Image loss is calculated similar to that given in equation (2), I is the original image and the Î is the warped image. The Adam optimizer is used for optimisation with a decaying learning rate of 10^{-5} due to its adaptive learning rate and momentum. In order to converge, both generator and discriminator models should achieve an optimal balance. In initial runs, the model suffered from collapses due to faster convergence of generator or discriminator. To resolve this several trials are required to identify the right balance between the generator and discriminator. The weighted adversarial term $(\alpha_{i,j})$ and the weightage (W_d) between the generator and discriminator loss are hyperparameters which are tuned to achieve the best results. The discriminator loss attained saturation much faster than generator loss and hence the ratio at which the weights are updated are more frequent in generator than the discriminator.

4. RESULTS AND DISCUSSIONS

4.1 Results

The pixel-wise disparity map is generated using both the models for test images. The disparity can be converted to depth maps using equation (1). The time taken for training with 22000 images for both models is around 20~23 hours with a single Nvidia GPU memory of 16GB. The model is tested with images had objects like rooftops, walls, vegetation and the plain ground surface.

4.1.1 Internal quality assessment: To assess the performance of the model in reproducing what it has learnt during training, the models are tested as an initial assessment with images from training dataset. The disparity learnt by the model during the training stage at the last epoch is compared with the disparity generated during testing for the same image. The disparity images for both the models generated during training and testing are shown in Figure 7, 8.

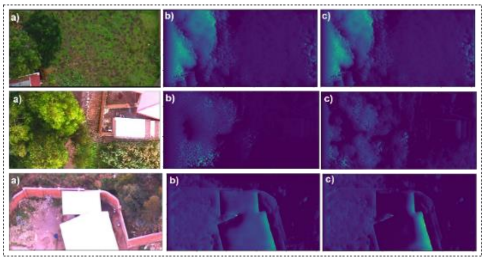

Figure 7. a) Input image b)Disparity from Training c) Disparity from Testing – for CNN model

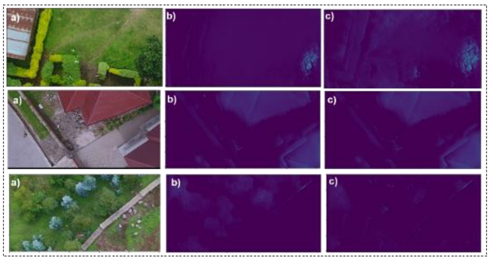

Figure 8. a) Input image b) Disparity from Training c) Disparity from Testing – for GAN model

The evaluation of the internal accuracy is done based on calculating several metrics between the disparity generated during training T(x) after fixing the model parameters and disparity generated during testing D(x). This includes Absolute Relative difference (Abs Rel) given in equation (6), Squared Relative difference (Sq Rel) given in equation (7), Root Mean Square Error (RMSE) given in equation (8), RMSE log and d1-all given in equation (9).

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|T(x_i)-D(x_i)|}{T(x_i)} \tag{6}$$

$$\text{Sq Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|T(x_i)-D(x_i)|^2}{T(x_i)} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T(x_i)-D(x_i))^2} \tag{8}$$

$$\text{D1-all} = \frac{1}{n}\sum bad\ pixels * 100 \tag{9}$$

Method	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all
CNN	5.309	0.1614	0.016	1.468	0
GAN	1.675	0.0314	0.009	1.042	0

Table 2. Metrics on the internal accuracy between the disparity image during training and testing for the two models (in pixels).

Also, the mean and standard deviations of the generated disparity during training and the generated disparity during testing is calculated to show the quality of the testing compared to its learning.

Method	Mean		Standard. Deviation	
	Trained disparity	Tested disparity	Trained disparity	Tested disparity
CNN	0.0101	0.0091	0.0135	0.0112
GAN	0.0078	0.0063	0.0105	0.0089

Table 3. Metrics on the disparity image during training and testing for the two models (in pixels).

It is observed from the evaluation metrics from Table 2 and Table 3 for both the models, that the trained and tested disparity generation is very similar in terms of metrics like mean, standard deviation etc. This shows the quality of image generation of both the models in the training and testing phase. The trained models are tested with single images to produce disparity images and the results are shown in Figure 9 - 12. The disparity results from CNN and GAN are different in terms of range variations. This could be due to the difference in learning process of both the models. CNN focuses only on image loss to improve the disparity generation while for GAN the task is to produce more realistic images as that of the original input image. This might have influenced GAN in reproducing the edges, texture variations similar to that of original image and may have limited its potential in producing more disparity range variations.

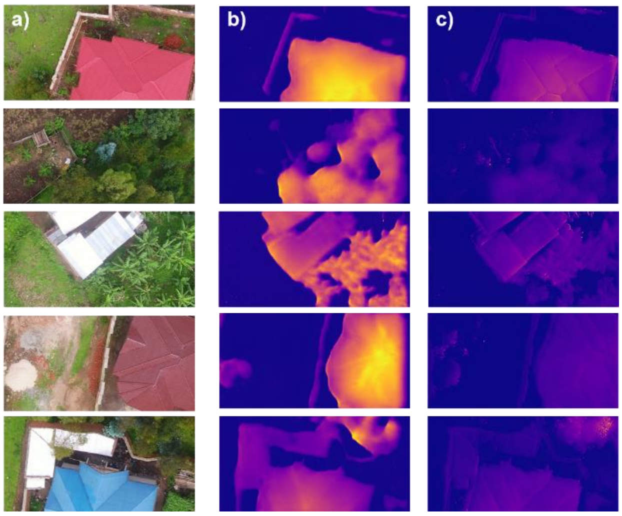

Figure 9. a) Single image b)Disparity from CNN c) Disparity from GAN

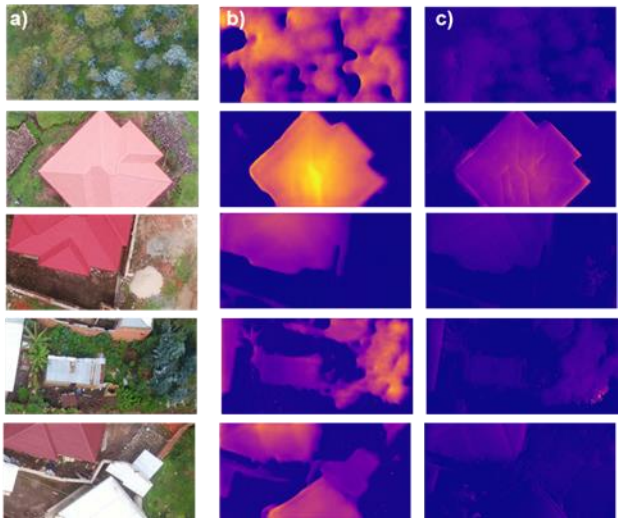

Figure 10. a) Single image b)Disparity from CNN c) Disparity from GAN

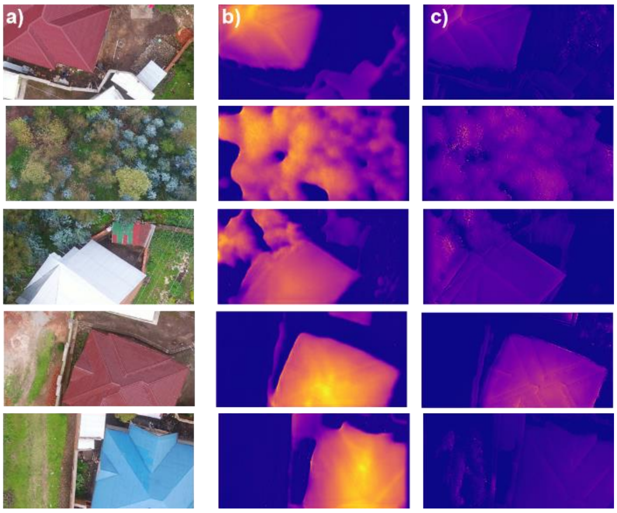

Figure 11. a) Single image b)Disparity from CNN c) Disparity from GAN

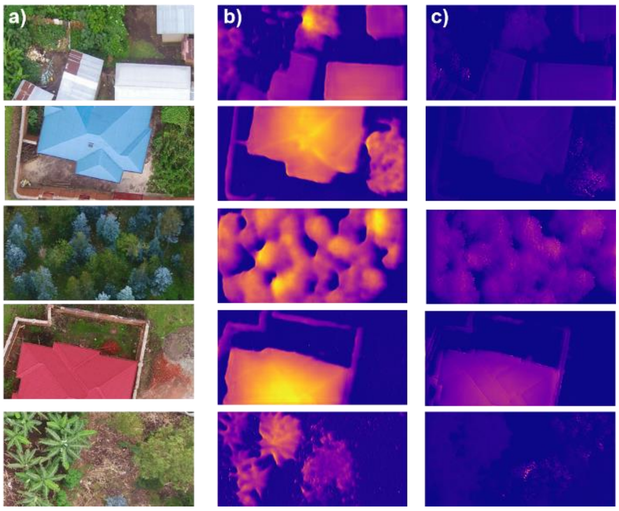

Figure 12. a) Single image b)Disparity from CNN c) Disparity from GAN

4.1.2 External quality assessment: To assess the quality of the generated depth images of both models, the results from these models are qualitatively compared with point clouds obtained from a commonly used Photogrammetric tool (Pix4D). The point clouds obtained from Pix4D are in real world coordinate system (WGS 1984, UTM 32N) with mean sea level as datum. On the other hand, the disparity generated from both the models are converted to depth using the equation (1), where the values of baseline and focal length are determined from Pix4D. It is to be noted that, the depth values in point clouds from Pix4D are determined with respect to the mean sea level while the depth values from the deep learning models are relative variations of depth within the field of view. As an example, in Figure 13, the depth values ranges from 224m to 232m for a chosen scene of point clouds and the predicted depth values from the models ranges between 0.2 m to 15m. To make these results comparable, the depths in Pix4D are converted to relative depths by shifting the datum from mean sea level to the lowest point in the field of view (by subtracting the ground level elevation value from all the points in point cloud).

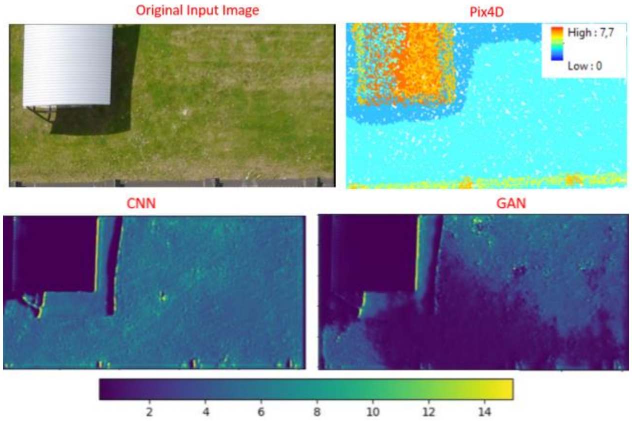

Figure 13. External quality assessment-Depth map

Method	Average Mean	Standard. Deviation
Pix4D	4	1.55
CNN	4.58	2.30
GAN	2.28	2.022

Table 4. Metrics for the depth estimated from the Pix4D and two models (in meters).

The point clouds have a new range from 0 to 7.7m relatively where the lower value represents the bottom most part and higher value represents the top most part. The depth from the model ranges from 0.2m to 15m, however where the lower value represents the topmost and higher value represents the bottom most point in field of view. The mean and standard deviations of all models are shown in Table 4 that are calculated based on relative depth variations. It can be seen from Figure 13 that the range of depth value produced by the models are mostly within 8m with some spikes of high values (represented by yellow). These spikes can be attributed to systematic errors (shadows). While comparing with Pix4D the performance of CNN is better than GAN which shows lots of artefacts. This behaviour of CNN might be attributed to its simple network architecture and loss parameter which makes it to learn better disparities.

4.2 Discussion

As can be seen from the generated disparity images, both CNN and GAN are capable of producing disparity maps. This can be seen from the original image (Figure 9 - 12), that both models have reproduced well, easily identifiable features and/or features that are closer to camera. It can be seen from Figure 8 that, while CNN smoothens out most of the sharp edges, GAN reproduces these edges. The range of disparity values produced by CNN is consistently higher than the range of values produced by GAN. After scaling and comparing the model results with Pix4D it is found that CNN produce better results whereas GAN underpredicts depth. It can be seen form Figure 13, that GAN produces lots of artefacts and blobs in the ground area. This might indicate the importance of image loss of CNN over discriminator loss of GAN in the learning process to produce disparity images.

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

The present study extends the single image depth estimation techniques at ground level to aerial images. UAV aerial images from different regions consisting of various features are collected. The images are pre-processed to produce rectified stereopairs and divided into patch pairs to create an aerial image dataset. This dataset is used for training two deep learning models - CNN and GAN, to generate disparity from single images. The internal quality of the disparity generation is evaluated using error metrics and is found that both the models are of good internal quality. Both the models are tested with single aerial images and is found to produce realistic disparity images. It is found that CNN produces a higher range of disparity values compared to GAN. For external quality assessment, the results from both models are compared with the point clouds generated by Pix4D. The model generated disparities are converted to depth using baseline and focal length. The Pix4D point cloud depth values are converted to relative depths by shifting the datum. On comparison it is observed that CNN produces realistic depth values than GAN.

To understand the results better, it will be compared with ground truth depth data to produce a comprehensive quantitative assessment. The models will be further improved by adding additional information that can help in achieving results that are closer to ground truth depth. Modification of the network architecture to include the baseline information such that depth can be learnt directly by the model is being taken up.

REFERENCES

Aleotti, F., Tosi, F., Poggi, M., Mattoccia, S., 2018. Generative Adversarial Networks for Unsupervised Monocular Depth Prediction, in: Lecture Notes in Computer Science. pp. 337–354. https://doi.org/10.1007/978-3-030-11009-3_20

Amiri, A.J., Loo, S.Y., Zhang, H., 2019. Semi-Supervised Monocular Depth Estimation with Left-Right Consistency Using Deep Neural Network.

Amirkolaee, H.A., Arefi, H., 2019. Height estimation from single aerial images using a deep convolutional encoder-decoder network. ISPRS J. Photogramm. Remote Sens. 50–66. https://doi.org/10.1016/j.isprsjprs.2019.01.013

Bhandare, A., Bhide, M., Gokhale, P., Chandavarkar, R., 2016. Applications of Convolutional Neural Networks. Int. J. Comput. Sci. Inf. Technol. 7, 2206–2215.

Chen, P.-Y., Liu, A.H., Liu, Y.-C., Wang, Y.-C.F., 2019. Towards Scene Understanding: Unsupervised Monocular Depth Estimation With Semantic-Aware Representation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2624–2632.

Chen, R., Mahmood, F., Yuille, A., Durr, N.J., 2018. Rethinking Monocular Depth Estimation with Adversarial Training 10.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network 1–9.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2012.6248074

Godard, C., Mac Aodha, O., Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. pp. 6602–6611. https://doi.org/10.1109/CVPR.2017.699

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Hu, J., Zhang, Y., Okatani, T., 2019. Visualization of Convolutional Neural Networks for Monocular Depth Estimation 3869–3878.

Jafari, O.H., Groth, O., Kirillov, A., Yang, M.Y., Rother, C., 2017. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation, in: Proceedings - IEEE International Conference on Robotics and Automation. https://doi.org/10.1109/ICRA.2017.7989537

Julian, K., Mern, J., Tompa, R., 2017. UAV Depth Perception from Visual , Images using a Deep Convolutional Neural Network. pp. 1–7.

Kang, S.B., Webb, J.A., Zitnick, C.L., Kanade, T., 1999. Multibaseline stereo system with active illumination and real-time image acquisition, in: IEEE International Conference on Computer Vision. pp. 88–93. https://doi.org/10.1109/iccv.1995.466802

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks, in: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016. pp. 239–248. https://doi.org/10.1109/3DV.2016.32

Li, J., Yuce, C., Klein, R., Yao, A., 2017. A two-streamed network for estimating fine-scaled depth maps from single RGB images. Comput. Vis. Image Underst. 186, 25–36. https://doi.org/10.1016/j.cviu.2019.06.002

Liu, F., Shen, C., Lin, G., Reid, I., 2016. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. IEEE Trans. Pattern Anal. Mach. Intell. 38, 1–16. https://doi.org/10.1109/TPAMI.2015.2505283

Luo, Y., Jiao, H., Qi, L., Dong, J., Zhang, S., Yu, H., 2018. Augmenting depth estimation from deep convolutional neural network using multi-spectral photometric stereo, in: 2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017 - . pp. 1–6. https://doi.org/10.1109/UIC-ATC.2017.8397464

Mehta, I., Sakurikar, P., Narayanan, P.J., 2018. Structured adversarial training for unsupervised monocular depth estimation. Proc. - 2018 Int. Conf. 3D Vision, 3DV 2018 314–323. https://doi.org/10.1109/3DV.2018.00044

Monasse, P., Morel, J.M., Tang, Z., 2010. Three-step image rectification. Br. Mach. Vis. Conf. BMVC 2010 - Proc. 1–10. https://doi.org/10.5244/C.24.89

Mou, L., Zhu, X.X., 2018. IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network. CoRR 1–13.

Nex, F., Remondino, F., 2014. UAV for 3D mapping applications: A review. Appl. Geomatics 6, 1–15. https://doi.org/10.1007/s12518-013-0120-x

Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N., 2018. Unsupervised Aersarial Depth Estimation using Cycled Generative Networks, in: Proceedings - 2018 International Conference on 3D Vision, 3DV 2018. pp. 587–595. https://doi.org/10.1109/3DV.2018.00073

Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S., Stefano, L. Di, Oct, C. V, 2018. Geometry meets semantics for semi-supervised monocular depth estimation, in: 14th Asian Conference on Computer Vision. p. 16.

Repala, V.K., Dubey, S.R., 2018. Dual CNN Models for Unsupervised Monocular Depth Estimation 9.

Saxena, A., Chung, S.H., Ng, A.Y., 2007. 3-D depth reconstruction from a single still image. Int. J. Comput. Vis. 76, 53–69. https://doi.org/10.1007/s11263-007-0071-y