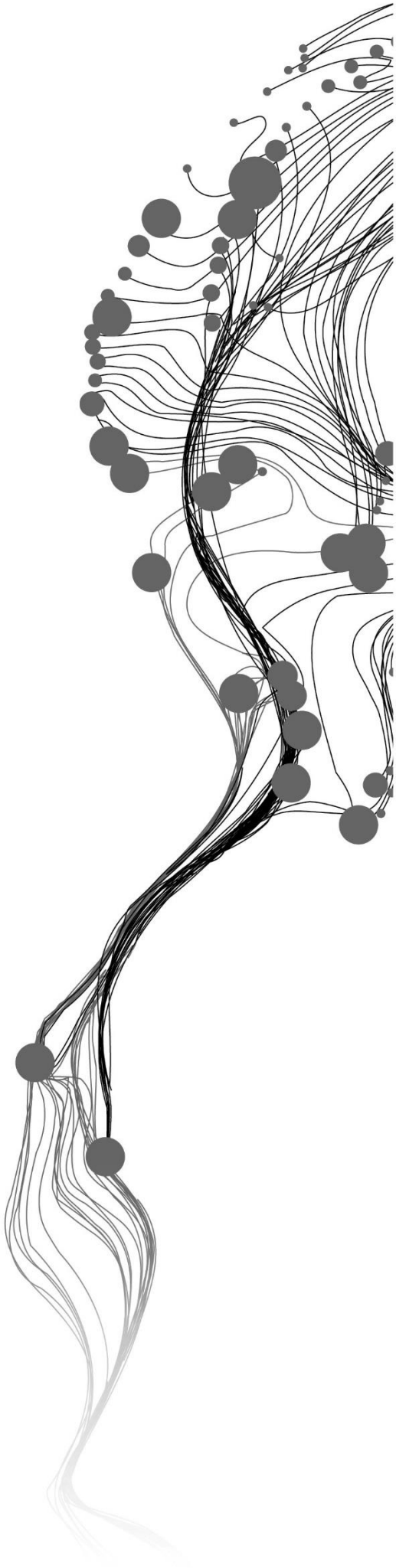
Monocular Depth Estimation of UAV Images using Deep Learning

LOGAMBAL MADHUANAND

June, 2020

SUPERVISORS:
DR. FRANCESCO NEX
DR. MICHAEL YANG



Monocular Depth Estimation of UAV Images using Deep Learning

LOGAMBAL MADHUANAND
Enschede, The Netherlands, June, 2020

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
DR. FRANCESCO NEX
DR. MICHAEL YANG

THESIS ASSESSMENT BOARD:
PROF. DR. ir. M.G.VOSSELMAN (CHAIR)
DR. F. REMONDINO (EXTERNAL EXAMINER, BRUNO KESSLER FOUNDATION, ITALY)

ABSTRACT

UAVs have become an important photogrammetric measurement platform due to its affordability, easy accessibility and its widespread applications in various fields. The aerial images captured by UAVs are suitable for small and large scale texture mapping, 3D modelling, object detection tasks etc. UAV images are especially used for 3D reconstruction which has applications in forestry, archaeological excavations, mining sites, building modelling in urban areas, surveying etc. Depth in an image, defined as the distance of the object from the viewpoint, is the primary information required for the 3D reconstruction task. Depth can be obtained from active sensors or through passive techniques like image-based modelling that are much cheaper. The general approach in image-based modelling is to take multiple images with an overlapped field of view which can be processed to create a 3D model using methods like structure from motion. However, acquiring multiple images covering the same scene with sufficient base may not always be possible for complex terrains/environments due to occlusions. Single image depth estimation (SIDE) can not only overcome these limitations but also have various applications of its own. Estimating depth from a single image has traditionally been a tricky problem to solve analytically. However with recent advancements in computer vision techniques and deep learning, single image depth estimation has attracted a lot of attention. Most studies that estimate depth from a single image has been done with indoor or outdoor images taken at ground level. Using similar techniques to find single image depth from UAV images has applications in object detection, tracking, semantic segmentation, digital terrain model, obstacle or sensor mapping etc. It can also be used to reconstruct a 3D scene with limited images acquired beforehand. The problem is generally approached through supervised techniques that use pixel-wise ground truth depth information, semi-supervised techniques that use some information that is easier to obtain than depth like semantics or self-supervised techniques which doesn't require any extra information other than the images. As the collection of ground truth depths is not always feasible and since the depths produced from self-supervised approach have proven to be comparable to that of the supervised approaches, self-supervised approach is preferable. Thus, this study aims to estimate depth from single UAV images in a self-supervised manner.

For a deep learning model to learn in a self-supervised manner, a large number of images are required. A training dataset with UAV images is prepared by taking images from three different regions. The preparation of dataset involves undistortion and rectification to produce stereopairs. Image patches of smaller size are extracted from the images to accommodate in deep learning models. Around 22000 stereo image patches are produced for training the deep learning model. The main objective is to find a suitable deep learning model for SIDE. Two models, CNN and GAN are chosen due to their proven success in single image depth estimation for indoor images. The network architectures are modified based on the specifications of the UAV images dataset. Both models take as input one image from the stereopair, generates a disparity and then warp it with the other image in the stereopair to reproduce the original image. CNN model is based on VGG architecture consisting of image loss, the difference between original and reconstructed image, for backpropagation. While GAN model consists of, generator and discriminator structure to handle the image reconstruction task. Both models are found to be capable of producing disparity images. The results from both the models are inter-compared qualitatively as well as quantitatively with reference depths from SURE. The disparity output from CNN model showed closer approximation to SURE depths while GAN model produced disparities with fine details reproducing edges of roofs etc. However, GAN model has high noises and spikes in ground surfaces which needed improvement. To improve the quality of the SIDE models, a third model - InfoGAN is suggested where additional mutual information through an added network is used to improve the model performance. The disparity from stereopairs and gradient information is used as mutual information in this study. The InfoGAN model with disparity information shows improved results that are closer to CNN. The right

mutual information provided through extended networks can improve the model performance even further.

Keywords: Single Image Depth, 3D reconstruction, Deep learning, UAV images, CNN, GAN, InfoGAN.

ACKNOWLEDGEMENTS

I would like to thank each and every one who had supported me in completing this thesis work in a fulfilling manner.

I would like to express my sincere thanks to my first supervisor, Dr. Francesco Nex for his expert guidance, motivation, relentless support and kindness shown to me at every stage of my research work. Without his patient guidance, I would not have been able to do this work. Dr. Nex has been a constant source of motivation and I am truly grateful.

I would also like to thank my second supervisor, Dr. Michael Yang. Dr. Yang has been extremely supportive and very prompt whenever I went to him with a query. His frequent emails with publications related to my work have expanded my knowledge and inspired me to aim higher.

I am indebted to my chair Prof. Dr. ir. M.G. Vosselman for his critical evaluation and suggestions for the betterment of the quality of my research.

Thanks to drs. J.P.G. Wan Bakx for his support not only in my research but also in my academic life at ITC in general.

I am also thankful to Sofia Tilon (PhD Student) who has been a consistent source of support whenever needed.

I would like to thank all the teaching faculty and staff at ITC who make ITC such a wonderful place to study and get inspired.

And finally, thanks to my family and friends who have always been there for me through times good and bad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. UNMANNED AERIAL VEHICLES (UAVs) IMAGES

UAVs are alternative photogrammetric measurement platforms, which has wide applications in close range, aerial and terrestrial photogrammetry for exploring the environment (Eisenbeib, 2009). The platform can be mounted with sensors to capture RGB or multispectral images, videos and also LIDAR devices for capturing 3D information as point clouds. UAVs are suitable for both small scale and large scale applications. The widespread availability of UAVs and easier access has led to the increased usage of UAVs for capturing data. Also, due to its low operating cost compared to other manned photogrammetric sources, UAVs have made the collection of high-resolution aerial images more affordable. This has led to the extensive use of UAV images especially for texture mapping, 3D modelling or 3D digital elevation models (Nex & Remondino, 2014). They can be flown in complex terrains and inaccessible areas with faster data acquisition and real-time processing. An image-based 3D modelling using UAVs involves flight planning, ground control point collection, image acquisition, camera calibration and 3D data extraction and reconstruction (F Remondino, Barazzetti, Nex, Scaioni, & Sarazzi, 2011). The initial step is to plan the flight and data acquisition procedures for the area of interest, deciding the ground sampling distance, camera parameters etc. The camera calibration and image orientation are important parameters for 3D reconstruction. They can be calculated either in-flight for low accuracy applications or can be obtained through post-processing after the flight. To create a 3D model using UAV images, multiple images of the same scene with sufficient overlap is acquired. From these images and orientation parameters, a 3D model can be generated through image matching techniques (Szeliski, 2010). 3D models are becoming very popular due to its photo-realistic denotation of the object. It has applications in various fields where the accurate portrayal of 3D model is needed, for example, forestry, archaeological excavation sites, geological mining sites, building modelling in urban areas, surveying etc (Nex & Remondino, 2014).

1.2. 3D MODELLING AND DEPTH INFORMATION

A 3D model can be generated from a depth image and can be used interchangeably for many applications. Depth in an image is defined as the distance from the viewpoint to the surface of scene objects with respect to the viewing angle. Depth is an important component in 3D visualisation to perceive the offset of images and to understand the geometrical patterns in a scene. Depth can enhance the performance of various tasks like semantic labelling, 3D-reconstruction, human body pose estimation in robotics and unmanned vehicle control (Amirkolaee & Arefi, 2019). Also, its societal importance can be seen from its importance in increasing the reliability of other scene understanding tasks like semantic segmentation, object recognition, topography reconstruction etc., (R. Chen, Mahmood, Yuille, & Durr, 2018). The depth or 3D information from an image can be estimated through active or passive techniques (S. Chen, Tang, & Kan, 2019). Active methods include measuring depths using dedicated instruments and sensors to obtain good accuracy. Although there are many depth sensors, like Microsoft Kinect, LIDARs and other laser sensors, they are sometimes affected by illumination, acquisition ranges, noisy images and high-cost factors (Liu, Shen, Lin, & Reid, 2016). On the other hand, passive techniques (image-based modelling) like a stereo, multi-view stereo, shape from motion, shape from shading, depth from focus etc., (Huang, Zhao, Geng, & Xu, 2019) rely on multiple view images or images with different lighting condition of the same scene to extract shape information. Due to its cheaper costs and faster generation compared to depth

sensors, depth extraction from images is highly preferable. They use either mathematical models or shape information for 3D reconstruction. Generally, in photogrammetry, depth is extracted from stereo images that are acquired using different camera positions for visualising the same portion of the scene. The camera calibration parameters along with the parallax from the stereo images are used for estimating the depth from images (Kang, Webb, Zitnick, & Kanade, 1999). The multiple images acquired from the same scene is matched through various feature detection and matching algorithms making this a robust approach (Repala & Dubey, 2018). However, acquiring multiple images covering the same scene with sufficient base may not always be possible for complex terrains/environments due to occlusions causing lack of features for matching images. For example, in urban regions with tall buildings, it is difficult to capture the required scene from multiple directions due to occlusions and inaccessibility. For evaluating damages in structures from available pre-damage images, single image 3D reconstruction is preferable (El-Hakim, 2001). Also in regions where rapid response is needed with low-accuracy requirements, single image depth estimation could be handy. This led to developments towards alternative approaches for estimating depth from monocular images which is still an ill-posed ambiguous problem (Eigen, Puhrsch, & Fergus, 2014).

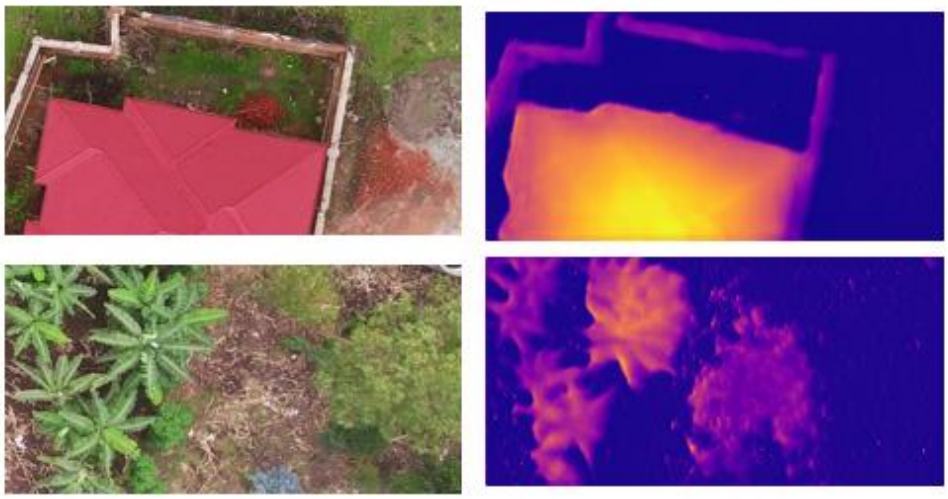


Figure 1: a) Single aerial image b) Disparity image- the colour variations denote the distance of the object from point of view.

Computer vision techniques are utilized in most fields for object recognition and image classification tasks with proven success. The advancement of automated algorithms in computer vision has made the extraction of information from scene geometry possible without the pre-knowledge of camera calibration parameters. The successful performance and recent advancements in deep learning techniques for extracting high-level features makes it a preferred tool for single image depth estimation (Amirkolaee & Arefi, 2019).

1.3. DEPTH EXTRACTION FROM SINGLE IMAGE

Depth can be perceived using cues like shading, gradients, texture variations and object focus etc., to reconstruct the geometrical information from the images (Saxena, Chung, & Ng, 2007). Amongst them, edges are an important source for extraction and differentiation of different objects in a scene (Hu, Zhang, & Okatani, 2019). The depth extraction from single images has been achieved using either supervised, self-supervised or semi-supervised techniques. This started with Eigen et al., (2014) where the depth was predicted by a supervised training approach which uses pixel-wise ground truth depth labelled images for

the training. Some studies have trained neural networks to deal with the estimation of depth from aerial images by training them using Digital Surface Models (DSM) (Amirkolaee & Arefi, 2019). The main challenge in supervised approaches is to obtain large training set with ground truth label or with corresponding DSM (Repala & Dubey, 2018). It is labour intensive and extremely time-consuming to match the image with its corresponding depth image at the same scale. In the semi-supervised approach, the training images are labelled with semantic or any other useful information that may aid in simplifying the computation of depth by guiding with more details about the semantics or other aspects of the scene (Zama Ramirez, Poggi, Tosi, Mattoccia, & Di Stefano, 2019). Although labelled semantic images are easier to obtain than ground truth depths, it is still an added complexity. The unsupervised or self-supervised technique involves computing the depth without the use of ground truth depth or any extra information other than an aerial image. The unsupervised depth estimation problem as proposed by Godard, Mac Aodha, & Brostow, (2017) is approached using rectified stereo image pairs for training the network, with known camera parameters to generate a disparity image through the pixel-wise correspondences. These stereo images acts as extra information for the model to learn disparity without directly training it with ground truth depth. The depth map can then be synthesised from the predicted disparity maps using the baseline and camera constant from the binocular stereo approach. Though these methods have proven to decrease the ambiguity in depth estimation from a single image, they have been applied widely only on indoor scenes like NYU-Depth2 (Silberman, Hoiem, Kohli, & Fergus, 2012) or outdoor scenes like KITTI dataset (Geiger, Lenz, & Urtasun, 2012) and not on UAV images.

Most deep learning models use the above mentioned cues for extraction of depth from monocular images. Among the different techniques available for training deep learning models, acquisition of stereopairs is easier and much more accessible than acquiring ground truth depth data. The performance of models that use stereo images for training is comparable with that of those which use supervised training approach with ground truth depths (Pilzer, Xu, Puscas, Ricci, & Sebe, 2018). The general approach of models that use stereopairs is to generate a disparity map with one input image of a stereopair and then warp the generated disparity with the other image in the stereopair to reproduce the input image (Godard et al., 2017). The losses between the original and reproduced input image are backpropagated through the network for learning to reproduce better disparity. The depth is extracted from the disparity through the binocular stereo concept with known baseline and camera constant. These models have proven to be successful in 3D reconstruction from indoor or outdoor images taken at ground level. Applying these models for aerial images taken from UAVs introduces added complexity due to the viewpoint being farther away, different perspectives, scaling issues, lack of certain depth cues etc., Unlike stereo images, depth from a single image not only requires local variations in images but it also needs to understand the global view to effectively integrate the features (Saxena et al., 2007). This necessitates the use of deep learning models that are capable of extracting both local and global variations within a scene.

1.4. APPLICATIONS OF DEPTH FROM SINGLE UAV IMAGE

Estimating depth from single aerial images captured from UAVs can be used to reconstruct 3D information of a scene without the use of multiple images of the same scene. It can be useful in places of natural disaster, where 3D reconstruction of the region is required with already available minimal images. This mechanism can also be used in tasks where regular photogrammetric block acquisition is not possible and in areas where it is acceptable for the 3D reconstruction to be of reduced quality. This will also open up new possibilities of scene explorations from the UAV images. The depth from single images can make the acquisition of Digital Surface Model (DSM) easier and affordable. It can provide height information for various tasks like object detection, tracking, semantic segmentation and Digital Terrain Model (DTM) generation with the limited number of images. Further, it can also be used onboard in UAVs for

augmented simultaneous localization mapping (SLAM) which can help in identifying the rough estimation of the position of the vehicle and obstacles.

1.5. RESEARCH IDENTIFICATION

The wide variety of applications and its importance in various domains makes the estimation of depth single UAV images an important topic of research. However, studies have been sparse due to the increased viewpoint complexity and difficulty in acquiring ground truths. With deep learning models showing promise in self-supervised monocular depth estimation for images take at indoors and at ground level there is an urgent need to apply these techniques to single UAV images. This study uses a self-supervised approach for depth estimation from single UAV images without the use of ground truth depths which hasn't been attempted before.
The scope of this study is to find a suitable model that can estimate depth from single aerial images captured by UAVs without the requirement of ground truth depths, making use of stereopairs for training. A single aerial image along with generated disparity is shown in Figure 1.

1.6. OBJECTIVES

The overall scope of this study is to find a deep learning model that can extract depth from single UAV images with reliable accuracy. The model is to be trained using stereopairs which acts as additional information for the model by replacing the use of the ground truth depth data. Two deep learning models with different architectures are chosen for the study to find a suitable architecture which can be further improved by adding additional features to generate better depth images from monocular scenes. The objectives of this study are:

1) Explore different deep learning models to find a suitable deep learning model for single image depth estimation (SIDE) from UAV images without using ground truth depth data for training.
2) Improve the deep learning model with additional elements to extract depth with reliable accuracy.
3) Assess the model performance and compare the results with the ground truth produced from different sources.

1.7. RESEARCH QUESTIONS

This led to the formulation of the following research questions,

1) What will be the suitable deep learning architecture for estimating depth from single UAV images?
2) What parameters can be included for improving the model performance?
3) How good the models are in relation to commonly used 3D reconstruction tools?

The general information about the importance of this research and the overall objectives are discussed in this chapter. Chapter 2 reviews the different approaches suggested in the literature to handle this problem. The methodology along with workflow and the model descriptions are discussed in Chapter 3. The performance of different models and the improvements are presented in Chapter 4. While Chapter 5 gives the overall conclusions on the appropriate model and suggestions for future work.

2. LITERATURE REVIEW

Depth estimation from images using computer vision techniques are very popular due to its successful performance. It includes the use of stereopairs (Alagoz, 2016), multiple image views of the same scene (Furukawa & Hernández, 2015; Remondino et al., 2013; Szeliski & Zabih, 2000), illumination or texture cues (R. Zhang, Tsai, Cryer, & Shah, 1999) etc. They follow the principle of binocular stereo vision or multi-baseline stereo (Kang et al., 1999) for extracting 3D information from the images.

2.1. DEPTH FROM STEREO IMAGES

To estimate depth, images with an overlapped field of view with different camera position is required. The cameras would be separated by a baseline distance. The images are rectified and projected on to the same plane to form stereopairs. From the image pair, one needs to identify the point or features for performing 3D reconstruction. To achieve this distinct points from one image should be identified and matched with the other image to find the homologous point. Matching the corresponding points from the left and the right image is a difficult task as distinct features or points need to be chosen to avoid confusion with the background scene. To find the corresponding points, sparse matching techniques or dense image matching techniques can be used. The sparse matching technique includes template-based matching, which matches the points through cross-correlation or through least squares (Szeliski, 2010). This is mainly used for orienting the images such that corresponding points lie along the same line. It also includes feature-based matching techniques which matches using key points and key descriptors. The task of key-point identification is done using Harris corner detector (Harris & Stephens, 1988), Förstner operator (Förstner & Gülch, 1987) etc by finding large intensity variations in an image. From key points, the surrounding variations are extracted through key descriptors which can be used for matching the pairs. The key descriptors can be identified through various algorithms, like Scale Invariant Feature Transform (SIFT) (Lowe, 2004) which computes image gradients within a local region surrounding key points. The key points and key descriptors are used to match features from one image to the corresponding feature in other images. Once the corresponding features are identified, matched and used for orienting the images, the 3D depth or point cloud can be obtained. To obtain denser point clouds, dense image matching techniques are used. This includes window-based matching technique which slides a window to calculate the absolute difference between the features, scan line stereo which uses dynamic programming to find the lowest cost path for identifying features, semi-global matching which uses pixel-wise matching and a regularisation term to reduce spurious matches etc. (Szeliski, 2010). Semi-Global Matching (SGM) proposed by Hirschmüller, (2005) has wider adoption in many recent computer vision tasks due to the quality results and its faster performance. SGM is a dense image matching technique, which matches pixel-wise mutual information by matching cost. Instead of using intensity difference alone for matching, SGM uses disparity information to find the corresponding pixels in other images.
The distance between the corresponding points from the left and right image defines the disparity map of the images. This disparity map can be used for 3D reconstruction. The disparity and depth information are related inversely as given in equation (1).

$$Disparity = X_l - X_r = \frac{Bf}{d} \tag{1}$$

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. The obtained disparity map can be used to calculate the depth information from the images through the baseline and camera constant. The concept of binocular stereo is shown in Figure 2.

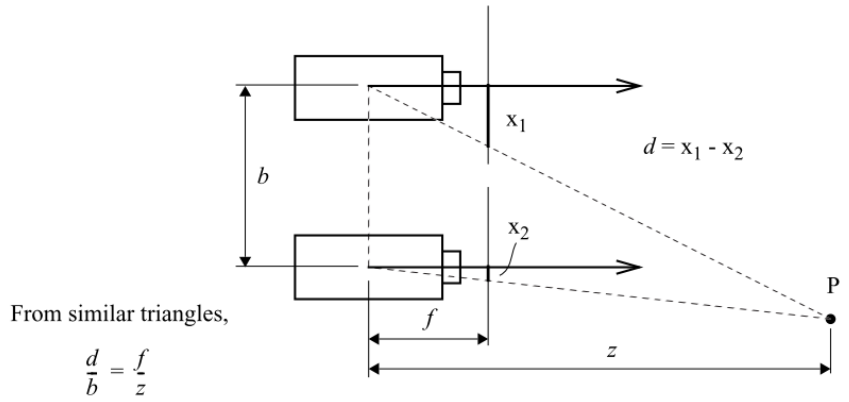


Figure 2: Relationship between different parameters baseline(b), focal length(f), disparity(d), depth(z) and ground point(p). Adapted from " Multibaseline stereo system with active illumination and real-time image acquisition", Kang et al., (1999)., p.3

2.2. IMAGE-BASED APPROACHES FOR SINGLE IMAGE DEPTH

The recovery of 3D information from image-based modelling is done through mathematical models as explained in the previous section or through shape extraction techniques called Shape from X. The shape can be expressed as depth, surface normal, gradient etc. X represents details like shading (Van Den Heuvel, 1998), texture (Kanatani & Chou, 1989), stereo, motion in 2D images (R. Zhang et al., 1999). Most of these techniques employ multiple images and to find corresponding points for matching is complex. Shape from shading developed by Horn in the 1970's is used to compute three-dimensional information from a single image using the brightness difference in the surface. Even though the solutions from shape from shading has proved to be not unique due to ambiguity in parameters of lighting, it acted as a base for many of the future solutions for single image depth estimation (Prados & Faugeras, 2006). It used the change in image intensity to obtain the surface shape and it suffered in areas that do not have uniform colour or texture (Saxena, Chung, & Ng, 2005). The assumption that surfaces are smooth and the difficulty to calculate the surface reflectance properties lead to inaccurate depth information.

Van Den Heuvel, (1998) proposed using the line-photogrammetric method by describing objects in an object model with geometrical constraints like parallels, perpendiculars among lines in objects which represents the edges between planar surfaces for extracting depth from single images. This model is mainly used for areas with man-made surfaces like buildings where the occurrence of such geometrical constraints are higher. El-Hakim (2001), suggested a flexible approach without object model and internal calibration parameters. Different types of constraints like points, surface, topology etc, for solving internal, external camera parameters and also obtaining 3D models are suggested. The shapes of objects are also combined with topological relations like parallels, perpendiculars etc. Similarly, L. Zhang, Dugas-Phocion, Samson, & Seitz, (2001) used a sparse set of user interactions for 3D reconstructions using constraints like surface normal, silhouettes etc. This algorithm yielded better results for objects with distortions forming a constrained optimization problem. Nagai, Ikehara, & Kurematsu, (2007) proposed a novel method for surface reconstruction called shape from knowledge using Hidden Markov Model (HMM). This models the relationship between the RGB image and its corresponding depth information. The approach is influenced by shape from shading mechanism but worked only for facial structures and failed to generalise.

As the research towards single image depth estimation increased, the use of constraints and complementary information also gets modified based on requirements. This added information for depth estimation is handled through three approaches based on the user influence in the model. The achievements in each approach are explained and the approach suitable for our task is selected.

2.3. SUPERVISED APPROACH

The analytical solutions for depth estimation from a single image like shape from X are not as good as that of stereo depth estimation. With recent developments in computer vision and deep learning techniques, there is an increasing possibility of using these techniques to overcome the limitations of analytical methods. This is mainly due to the success of Convolutional Neural Networks in learning depth from colour intensity images. Several studies have been published on depth estimation from a single image using ground truth depths for training deep learning models. Saxena et al., (2007) proposed the use of a global context of the image as local features alone will not be sufficient for single image depth estimation. They used Markov Random Field (MRF) to incorporate the relation between depths at different points within the image. They trained the model with the monocular image of both indoor and outdoor scenes taken at ground level, along with the corresponding ground truth depths. They followed a patch-based model to extract most of the features. But this model had problems with weak unconnected regions without global contextual information. Eigen et al., (2014) suggested the integration of both global and local information by using a multi-scale network for coarse and fine prediction. However, the depth image is inferred directly from the input image compared to other robust techniques and the generated depth image has lower resolution compared to the original input image. The use of deep structured learning for continuous depth values by unifying continuous Conditional Random Field (CRF) and deep Convolutional Neural Network (CNN) framework is implemented by Liu et al., (2016). Li, Yuce, Klein, & Yao, (2017) proposed a two streamed network for predicting depth along with depth gradients which are fused to form a final depth map. This helped them to capture local structures and fine detailing through the two-streamed network. Jafari, Groth, Kirillov, Yang, & Rother, (2017) used cross-modality influence for joint refinement of the depth map and semantic map through monocular neural network architecture. They achieved a beneficial balance between the accuracy of the network and the cross-modality influence. R. Chen et al., (2018) moved a step ahead, by approaching the monocular depth estimation through adversarial learning. They implemented a generator network to learn the global context through patch-based information. The discriminator network distinguishes between the generated depth map and the ground truth depth map. These approaches are mostly implemented on indoor or outdoor datasets taken at ground level.

Julian, Mern, & Tompa, (2017) compared different style transfer methods like pix2pix, cycle GAN, multi-scale deep network for aerial images captured from UAVs. They trained the model using the UAV images along with depth image pairs and refined the feature-based transfer algorithm for this single image depth estimation purpose. Mou & Zhu, (2018) used a fully residual convolutional-deconvolutional network for extracting depth from monocular imagery. They used aerial images along with the corresponding DSM generated through semi-global matching for training the network. The two parts of the networks acts as s a feature extractor and height generator. Amirkolaee & Arefi, (2019) proposed a deep CNN architecture with an encoder-decoder setup for estimating height from aerial images by training them with the corresponding DSM. They extracted the full satellite image into local patches and trained the model with the corresponding depth and finally stitched the depths together. They faced issues for small objects with fewer depth variations like small vegetations, ground surfaces within the scene etc.

Although all these methods proved to be successful, they all require huge amounts of ground truth depth images while training the model. Acquiring UAV images along with their corresponding DSM is

complicated making supervised approach less preferable compared to other approaches even though it produces better accuracies for single image depth estimation.

2.4. SEMI-SUPERVISED APPROACH

The supervised approaches required pixel-wise ground truth depths which is not always practical to acquire. To overcome this, researchers used information other than depths during training. Zama Ramirez et al., (2019) suggested training the network with semantic information which could effectively improve the depth estimation. They had a joint semantic segmentation and depth estimation network architecture, which uses the ground truth semantic labels for training. Even though acquiring semantic information is less complicated than ground truth depth, it is still an added complexity which required manual processing. Amiri, Loo, & Zhang, (2019) approached this semi-supervised task differently. They used both LIDAR depth data and rectified stereo images at the same time during training. They also included a loss term, left-right consistency loss to check the consistency between the generated left and right depth maps. Even though the semi-supervised approach has lesser difficulties in ground truth depth data, yet it has other requirements which make this an equally challenging task. This shifted the interest towards an unsupervised or self-supervised approach which doesn't require laborious ground truth depth construction.

2.5. UNSUPERVISED/SELF-SUPERVISED APPROACH

Unsupervised or Self-supervised approaches utilise the multi-view images instead of vast amounts of ground truth depth maps for training the neural networks. The reduced dependencies on laborious ground truth data collection have generated a lot of interest in these approaches. Garg, Vijay Kumar, Carneiro, & Reid, (2016) circumvent the problem faced by supervised learning by utilising stereo images instead of ground truth depth maps. They used the 3D reconstruction concept to generate a disparity image from stereo images and reconstruct the original image through inverse warping. They suggested that this approach can be continuously updated with data and fine-tuned for specific purposes. Although the model performed well, their image formation model is not fully differentiable. Godard et al., (2017) overcome this by including a fully differential training loss term for left-right consistency of the generated disparity image to improve the quality of the generated depth image. Repala & Dubey, (2018) based on the approach of reconstruction of images from disparities, suggested dual CNN with 6 losses for each network to train to generate a corresponding depth map. They utilised two CNN architectures one each for left and right images. The Generative Adversarial Network (GAN) introduced by Goodfellow, Bengio, & Courville, (2016) proved well capable of solving complex computer vision problems. Many developments in adversarial learning led to different network modifications like Conditional GAN (Mirza & Osindero, 2014), Deep Convolutional GAN (Radford, Metz, & Chintala, 2016), Information maximising GAN (X. Chen et al., 2016), Cycle consistent GAN (Zhu, Park, Isola, & Efros, 2017) etc. The adversarial learning models mark the current state of the art in many areas where deep learning is being used. A simple GAN network consists of a generator that learns to produce realistic images and discriminator that learns to find the difference with real images. MonoGAN by Aleotti, Tosi, Poggi, & Mattoccia, (2018) used a combination of generator and discriminator network for the monocular depth estimation. The generator loss is combined with an image loss to improve the disparity image synthesis process. This simple architecture is further modified by different adversarial learning process to achieve the task of depth estimation from a single image. Mehta, Sakurikar, & Narayanan, (2018) used this structured adversarial training to improve the task of image reconstruction for predicting depth images

from the stereo images. The baseline between the stereopairs is varied in a sequential and organised manner within a range, making it crucial for the model to learn. This varying baseline is scaled with the generated disparity which is warped with the left image to produce the right image. To improve the image synthesis process, a complex GAN architecture called Cycle GAN is proposed by Pilzer, Xu, Puscas, Ricci, & Sebe, (2018). The model consisted of a cycle with a combination of generator and discriminator in each half-cycle. The half-cycle uses the right image as an input to the generator for generating a disparity map, warping it with the left image to produce the right image. This is compared by discriminator to identify the false right images from the realistic right images. The produced right image acts as an input for the generator in the next half-cycle to produce a left image. Since this uses a cyclic structure, the model is referred as cycle GAN, where the loss terms include image loss, generator loss, discriminator loss along with a cycle consistency loss term.

These are some of the implementations for solving the monocular depth estimation problem from stereo images. Most of these models used indoor datasets or outdoor datasets taken at ground level. Our approach is also to use the information from stereo views to find an apt model for the aerial image dataset captured by UAVs.

3. METHODOLOGY

This chapter details about the different UAV datasets used to prepare the training dataset. The pre-processing step to prepare stereopairs along with the generated image quality are discussed. The overall workflow, with the detailed description of the different deep learning models chosen and the implementation of the models for our study are also presented. The tools used for reference depth generation are also described.

3.1. PRE-PROCESSING AND PREPARATION OF TRAINING DATASET

The deep learning models require large amounts of data for training in self-supervised approach with stereo images. The dataset consists of high-resolution UAV images captured over different regions the details of which are given in Table 1. This includes many land use/landcover features like buildings, vegetation etc., captured from different perspectives. The UAV images are captured sequentially over a region based on photogrammetric block. The images had around 90% forward overlap and about 70% side overlap for all the selected regions. The images with maximum overlap with adjacent images along the strip are selected to extract the stereo image pairs. The total number of images from each region along with the ground sampling distance is specified in Table 1. In order to make the dataset more representative and avoid overfitting of the model, it is ensured that the dataset consisted of a mixture of UAV images. Figure 3 shows the three photogrammetric blocks that are used for preparing the training dataset.

Table 1: Dataset distribution

Dataset	Average Ground sampling distance (GSD) in cm	Full images	Stereopairs	Image patches
EPFL Quartier Nord, Switzerland	3.05	125	100	1500
Ruhengeri, Rwanda, East Africa	3.01	1115	950	17120
Zeche Zollern, Germany	2.05	375	300	4500

Figure 3: Full Photogrammetric blocks a) EPFL Quartier Nord, Switzerland b) Rwanda, Africa c) Zeche Zollern, Germany

3.1.1. STEREOPAIR GENERATION

The UAV images are pre-processed to remove the distortion and rectify them to generate stereopairs. This processing is required to compute the precise depth information from the stereo images. The images captured by UAVs suffer from radial distortions. The image distortion changes the real geometry of the image. An object looks displaced from the correct position. This will also make it difficult to match the corresponding features specifically near borders of the image. This is corrected by using the camera

calibration parameters which is obtained during initial processing in Pix4D tool. The camera parameters include extrinsic, intrinsic and distortion coefficients. The extrinsic parameters represent the transformation of object point from the world coordinate system to image coordinate system through translation and rotation. While the intrinsic parameters refer to the projection of the object point to the ideal image point in pixel coordinates. The image coordinates are modelled for non-linear image errors like distortion using the equation (2).

$$x_{corrected} = x_{image} + f(x, y)$$
$$y_{corrected} = y_{image} + f(x, y) \quad (2)$$

Where $x_{corrected}, y_{corrected}$ represent the undistorted image coordinates, x and y represent the distorted image coordinates with respect to principal distance and projection center which are added with an additional term to describe distortion and $f(x, y)$ represent the non-linear error function. The undistorted images are then rectified to make the homologous points in the generated stereopairs lie along the same rows (Junger, Hess, Rosenberger, & Notni, 2019). This is performed by comparing images taken from multiple views of the same scene with good overlap, then extracting the features and matching the corresponding points through the support of epipolar constraint (Szeliski, 1999; Remondino et al., 2013; Aicardi, Nex, Gerke, & Lingua, 2016). The epipolar geometry restricts the location of the feature in the second image within a line making it easier to identify the corresponding features as shown in Figure 4.

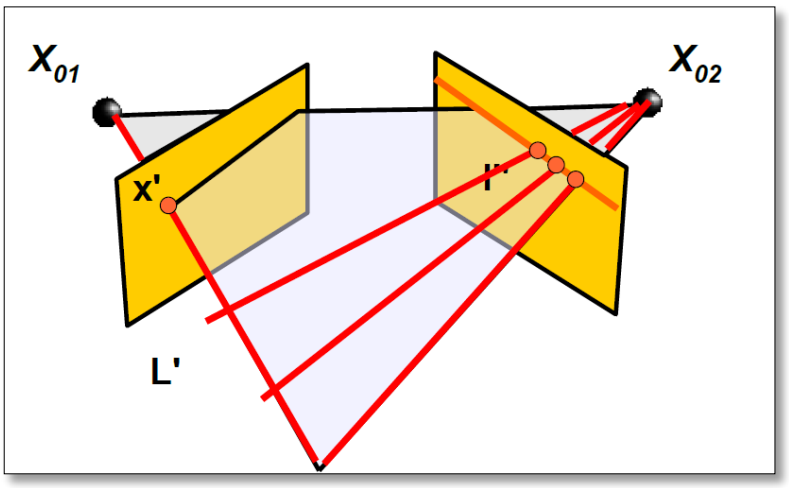


Figure 4: Epipolar constraint for feature identification

In Figure 4, X_{01} and X_{02} denote the overlapped images and x' represent feature in image 1 and x" represent the same feature in image 2. The epipolar constraints restrict the position of the x" along the line I", making it easier for identification. After image matching, the images are projected and oriented onto a common plane where the shift of corresponding pixels of the left and right images are only in x-direction. This process is repeated for all the UAV images in a photogrammetric block. The stereopair generation is automized with MATLAB scripts. The total number of stereopairs generated from the full block of UAV images are given in Table 1 and samples of generated stereopairs are shown in Figure 5.

Figure 5: Examples of rectified Stereopairs a)EPFL Quartier Nord, Switzerland b)Rwanda, Africa c) Zeche Zollern, Germany

The accuracy of the generated stereopairs limits the accuracy of the depth estimation model since the stereopairs are the only information guiding the model during training (Amiri et al., 2019. Errors in stereopair might arise due to improper rectification, wrong matches while feature matching, residual distortions that the camera calibration could not be handled etc. This can cause the generated stereopairs to have homologous features not along the same row. The error in stereopair generation will add up with the errors produced by the model leading to the generation of poor quality disparity or depth images. Hence while generating the stereopairs, a condition is imposed such that the matching error between the corresponding points is not more than 0.2 pixels. This means that the stereopair will not be generated if corresponding features are shifted by more than 0.2 pixels. Also, the generated stereopairs are randomly selected and verified for the pixel values for the homologous points in the left and right images if they lie along the same row as shown in Figure 6. It is found that for the randomly selected pairs, the homologous features lie along the same line with the same row number and different column number. In Figure 6, Pixel positions (Column, Row) are shown for the left and right image. The position of the same feature in both images lies along the same row and different column.

Figure 6: Examples of rectified Stereopairs showing the homologous points (marked in yellow circle) along same row

3.1.2. EXTRACTION OF PATCHES

The generated stereopairs are of high resolution and feeding them directly to the model has computational difficulties. The resizing of the stereopairs leads to loss of details and hence to maintain the resolution along with the information present in the scene, the images are extracted into smaller patches. Each stereo image pair is divided into smaller patches by following the admissible input size of the model. From the total of 1300 stereopair images, 22000 image patches are generated for use in training and 600 image patches for testing. The process is automated using scripts written in MATLAB R2018. The corresponding patches from the left and right images are used to form the stereopairs for training the model. The size of each patch is 512 * 1024. A sample stereopair along with the extracted patches from the left image is shown in Figure 7.

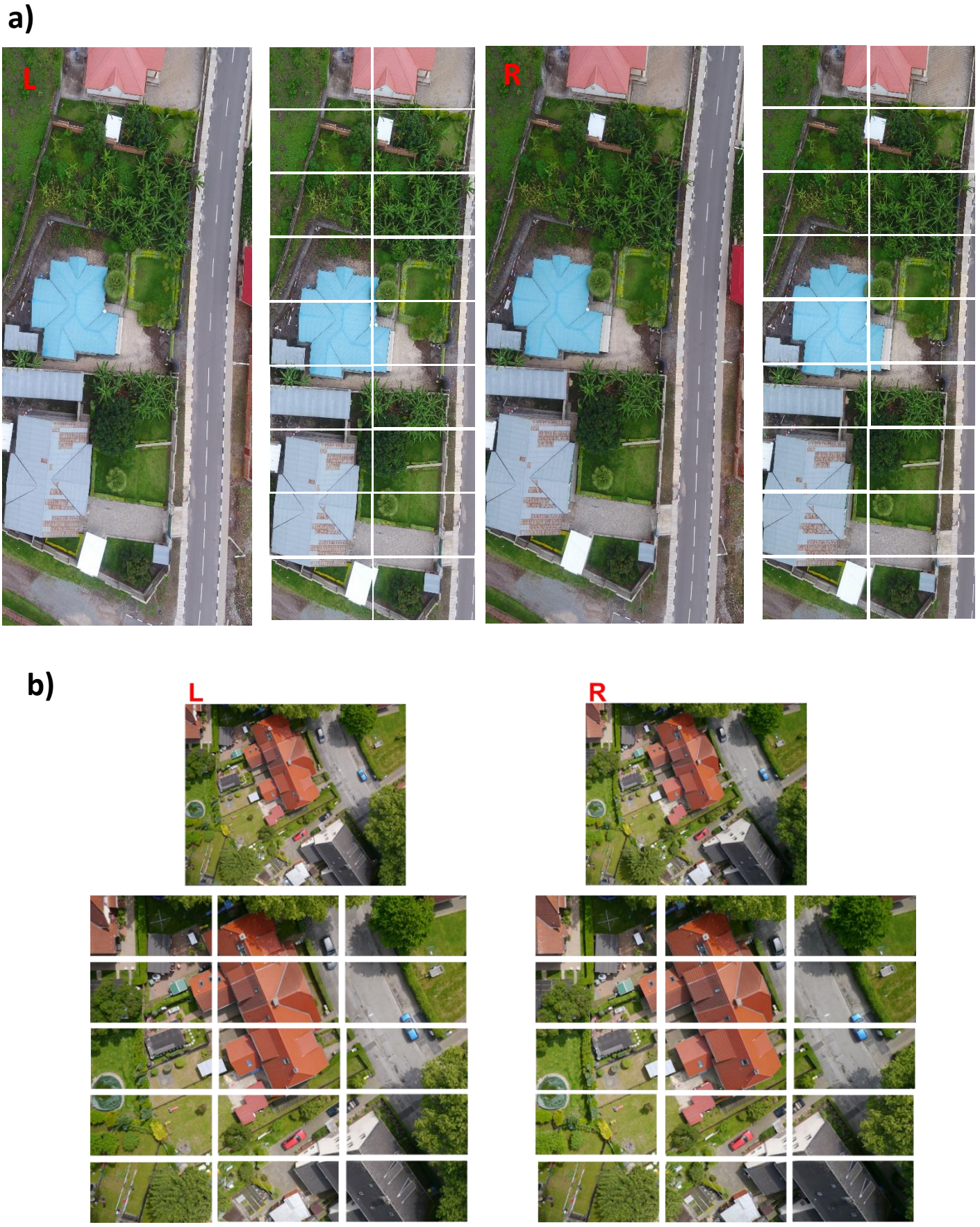


Figure 7: Rectified Stereopairs a) Rwanda, Africa b) Zeche, Germany along with the extracted patches from Left image and Right image

3.2. WORKFLOW

The overall workflow involves preparing a UAV image dataset which is pre-processed to generate left-right stereo patches, training deep learning models and evaluate the accuracy of the tested image patches. The stereo patches are used as a replacement for ground truth depth data for training the models. The single image depth estimation problem is treated as an image reconstruction problem, using the encoder-decoder deep CNN model. The model takes the left image from the stereopair to produce disparity. The model produced disparity is warped with the right image through bilinear sampling to reconstruct the left image. The right image is not directly given as input to the model but used with the generated disparity to produce the left image. The difference between the reconstructed left image and the input left image will be 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 a self-supervised manner. Two models - CNN and GAN are trained using the UAV dataset to produce disparity from a single image.

The internal qualities of the models are evaluated and the disparity images generated from the test images are inter-compared. This will help in understanding the relative difference in the performance of the two architectures for such ill-posed problems. They are further compared with point clouds generated through commercially available photogrammetric tools (Pix4D and SURE). Based on the comparison, fine-tuning tasks are included which involves giving additional information for the models to perform better. To improve the performance of SIDE model, additional information with the help of third network is provided. This forms architecture of third model InfoGAN which further improves the model performance in generating disparity maps. This overall structure is shown in Figure 8.

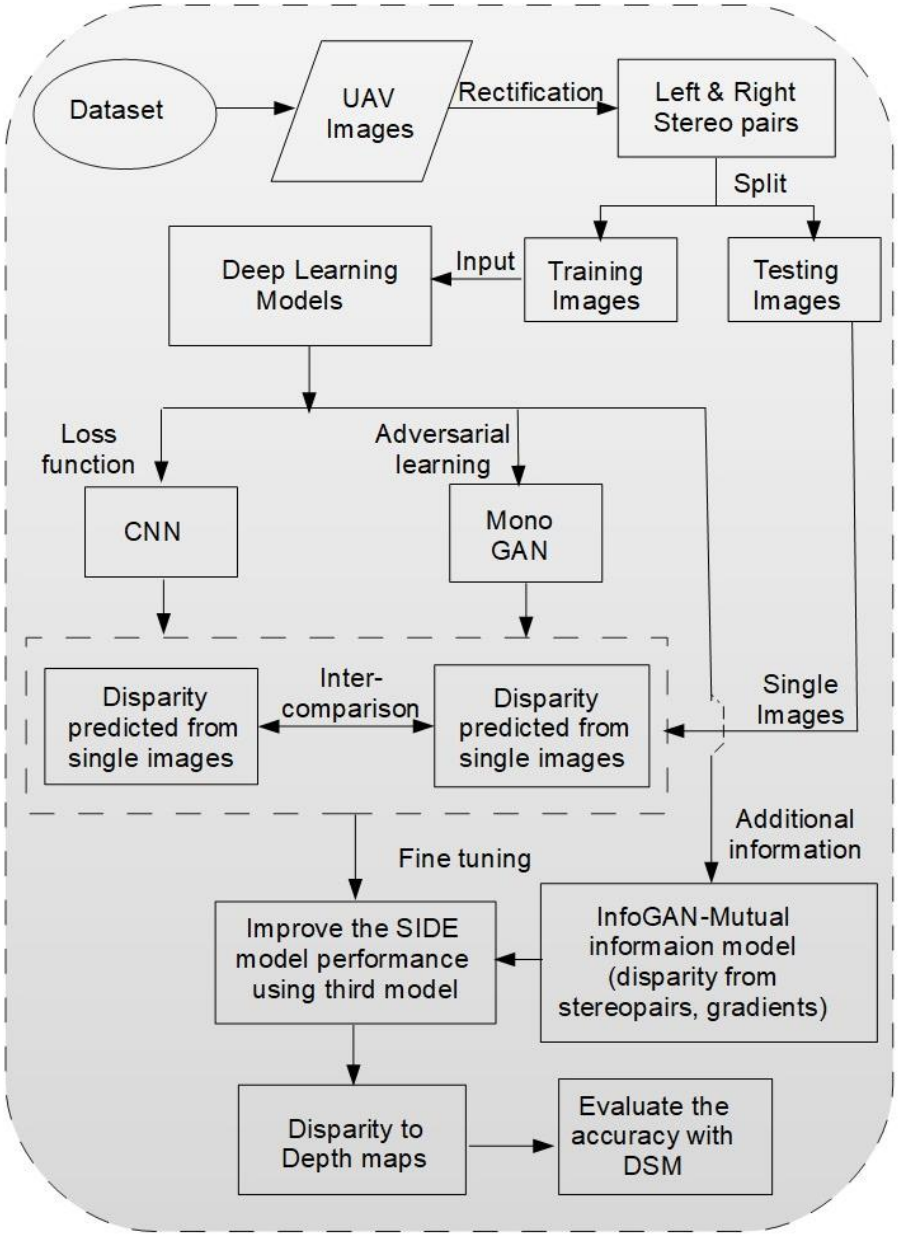


Figure 8: Workflow for single image depth estimation model

3.3. MODELS USED

The extracted patches are trained with deep learning models which differ in network architecture. Among available networks, two models, CNN and GAN are used for training the stereo patches. Various studies have proved CNN to be suitable for image reconstruction tasks which makes it preferable for this depth estimation problem. Similarly, GAN has shown successful performance in image generation tasks. These two deep learning models are chosen to study their ability in single image depth estimation from UAV aerial images.

Most studies have tested these models on images taken at indoor or outdoor scenes taken at ground level like KITTI dataset. Using these models for aerial images introduces more complexity compared to images taken at ground level. In aerial view, the objects are very much far away from the point of view compared

to the ground level images. This makes the absolute disparity ranges to be very small as the depth from aerial view is large. Also, the images at ground level contain more objects and details compared to aerial view which makes the model learn more variations. In aerial perspective, most of the details get faded due to large distance from the camera and also the local variations between similar objects are difficult to identify.

Dual CNN proposed by Repala and Dubey, (2018) and MonoGAN proposed by Aleotti et al., (2018) has been used in this study for inter-comparison as their model produced better accuracy for the benchmark KITTI dataset. The models are modified to take into account the differences in the characteristics of the dataset. Based on the model results, GAN architecture is further modified to form the third model to increase model performance. X. Chen et al., (2016) suggested the use of mutual information (complementary cues) to increase the model performance calling it InfoGAN. There are two mutual information's used in this study for improving the model performance, one is stereopairs to produce disparity and the other is gradient information. The architecture of the improved InfoGAN model with both mutual information is also explained below. The overall network architecture of the four deep learning models and the changes made for accommodating the UAV images are explained below. All models are executed in Python(3.6) using TensorFlow (1.15) platform.

3.3.1. CNN

The network architecture from Repala & Dubey, (2018) - Dual CNN model is shown in Figure 9. They utilised two CNN architectures each for left and right images. During the training phase, the left image is given as an input to left CNN (CNN-L) to produce left disparity and the right image is given as an input to right CNN (CNN-R) to produce right disparity. The left and right images are then reconstructed using bilinear sampling with the obtained disparity maps. For instance, the left disparity image, generated from the left CNN is warped with the right image to reconstruct the left image as output and similarly, the right disparity image, generated from right CNN is warped with the left image to produce a right image. The reconstructed left and right images are compared with the original input images to calculate the losses. The three types of losses used for comparison are, matching loss, disparity smoothness loss and left-right consistency loss for each CNN architecture. The loss terms will be calculated and back-propagated to improve network performance. This is the main structure of the Dual CNN with 6 losses (3 for the left image and 3 for the right image). Also, the dual network with 12 losses is proposed by modifying the left and right CNN to produce two output disparities from each CNN architecture. Repala & Dubey, (2018) trained the model with images from KITTI dataset covering outdoor scenes taken at ground level.

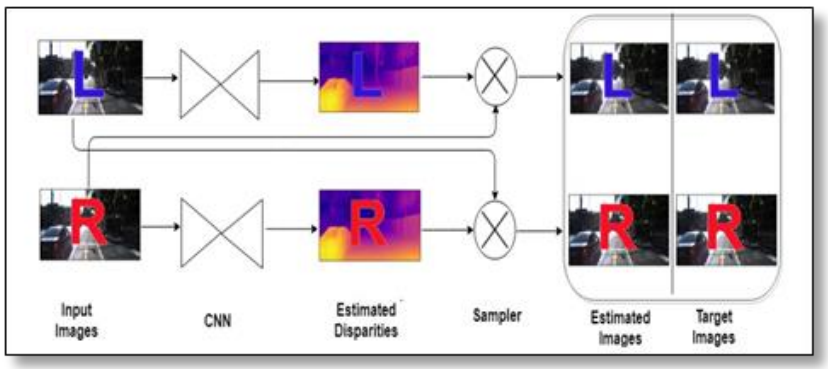


Figure 9. Dual CNN with 6 losses. Adapted from "Dual CNN Models for Unsupervised Monocular Depth Estimation", by Repala & Dubey, (2018)., p.3.

Implementation:

Compared to indoor and outdoor images taken ground level, aerial images cover larger areas with finer details. In order to accommodate this information, the network needs to be tuned for the chosen training dataset. To optimize the computational effort, a VGG based network architecture is chosen. A single CNN for left image is found to converge better for this training dataset than the two CNNs proposed by Repala & Dubey,(2018). This could be due to the complexity of the training images which requires a simpler network for minimisation of loss. The network consists of encoder and decoder structure for downsampling and upsampling respectively. The encoder consists of seven downsampling layers with the increasing number of filters in each layer. The sizes of the filters are of dimensions 7*7, 5*5 and predominantly 3*3 to extract fine details. This reduces the size of the original input image as it passes through each layer. The decoder also consists of seven layers for upsampling the input from the encoder to the original input size. Here, the number of filters used in each is reduced as we move through the layers. The dimensions of the filters are 3*3 for all layers. Further, the output from each decoder layer is concatenated with the convolutional output layer from the encoder from the last. For example, the first convolutional layer from the decoder is concatenated with the last layer of the encoder and the convolutional filters are applied to the concatenated feature. This is carried out to transverse the information through layers without losing features. From the UAV stereo patch dataset, the left patch is given as input to the VGG based encoder-decoder network. To maintain consistency, both left and right disparity are generated from the network using the left image alone. The generated left disparity is warped with right stereo patch through bilinear sampling to reconstruct left stereo patch and the generated right disparity is warped with the left stereo patch to reconstruct right stereo patch. The loss between the generated stereo patch and the original stereo patch is added as image loss. This is backpropagated through the network for improving the disparity generation.

Repala & 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 the present study. A single image loss was then used for this study as given in equation (3) which compares the generated image with the original image. This image loss is in simpler terms a combination of L1 norm and Structural Similarity Index Metric (SSIM) (Wang, Bovik, Sheikh, & P.Simoncelli, 2004) for left and right images as shown below. SSIM is used for measuring the similarity between two images where one image is considered as good quality compared to the other image. It compares factors like luminance, contrast, structure within widows of one image with others.

$$Image\ Loss = \frac{1}{N}\sum_{i,j} \alpha \frac{(1-SSIM(I_{i,j}^{\beta}, \hat{I}_{i,j}^{\beta})}{2} + (1-\alpha)||I_{i,j}^{\beta} - \hat{I}_{i,j}^{\beta}|| \quad (3)$$

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. The model architecture along with the loss component is shown in Figure 10.

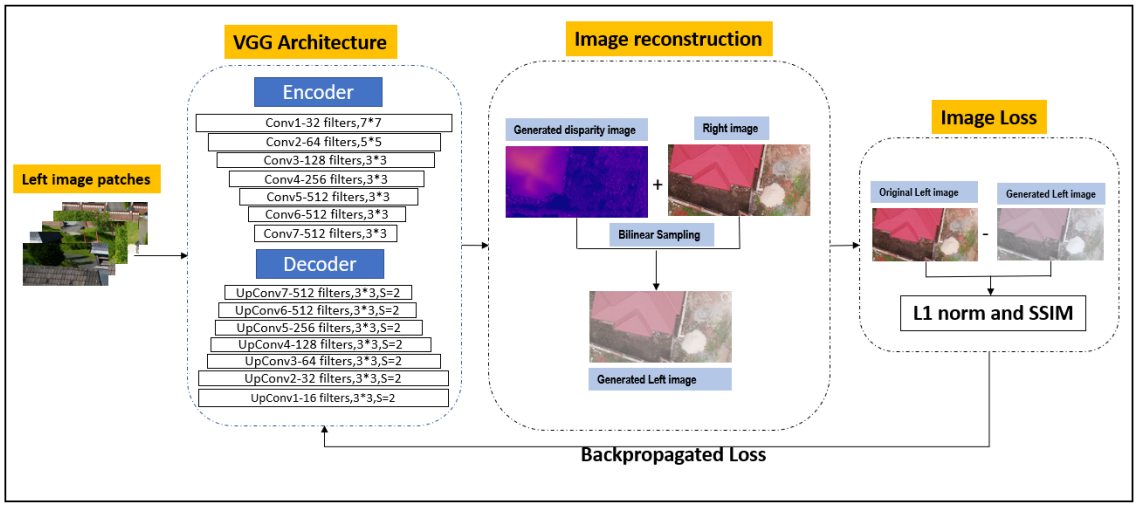


Figure 10. Simple CNN architecture with Image reconstruction loss modelled

This specific architecture with a single left CNN and a single loss for backpropagation 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 final epoch.

3.3.2. GAN

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 is then warped with the right image through bilinear sampling to synthesize a left image. The discriminator will try 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 11. The total loss will be 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). The generator competes with the discriminator to reconstruct better disparity maps and the discriminator tries to increase the probability of distinguishing between the original and generated images.

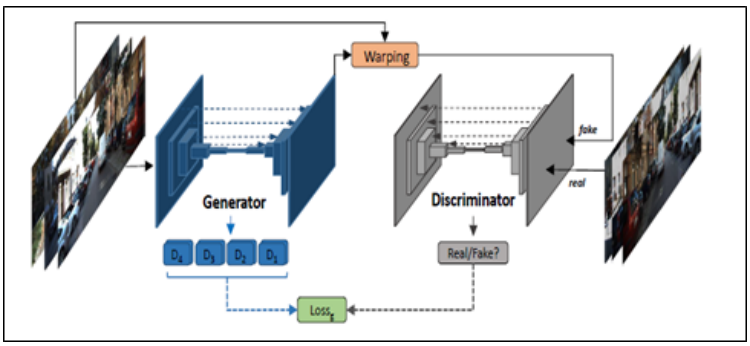


Figure 11. MonoGAN for stereo depth estimation. Adapted from "Generative Adversarial Networks for unsupervised monocular depth prediction", by Aleotti et al., (2018).

Implementation:

VGG based network architecture similar to the CNN architecture proposed in the previous section is used for the generator network for feature map generation. With the generator, the second network for discriminator is included whose task is to distinguish between the real and fake images which is much easier compared to the task of the generator which has to reconstruct images. Hence the discriminator will have a simpler architecture with less number of feature maps generated from each of its layers compared to the generator. The discriminator consists of a set of 5 convolutional layers with decreasing number of filters to reduce 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 are 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 will be a sum of the generator and discriminator loss as shown in equation (6). 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 (4). While the discriminator loss is the probability that the original image and generated image is classified accordingly as given in equation (5). The model architecture is shown in Figure 12 with different losses combined.

$$Generator\ Loss\ =\ \text{Image Loss}\ +\ \alpha_{i,j} * E_{\hat{I}}(log(D(\hat{I})) \tag{4}$$

$$Discriminator\ Loss = -1/2[E_I(log(D(I)))] - 1/2[E_{\hat{I}}(log\left(D(1-\hat{I})\right)] \tag{5}$$

$$Total\ Loss = Generator\ Loss + W_d * Discriminator\ Loss \tag{6}$$

where the Image loss is calculated similar to that given in equation (3), I is the original image and the $\hat{I}$ is the warped image. The Adam optimizer is also chosen here 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 either generator or discriminator. To resolve this, several trials were required to identify the right balance between the generator and the discriminator. The weighted adversarial term ($\alpha_{i,j}$) and the weightage (W_d) between the generator and discriminator loss are hyperparameters that 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.

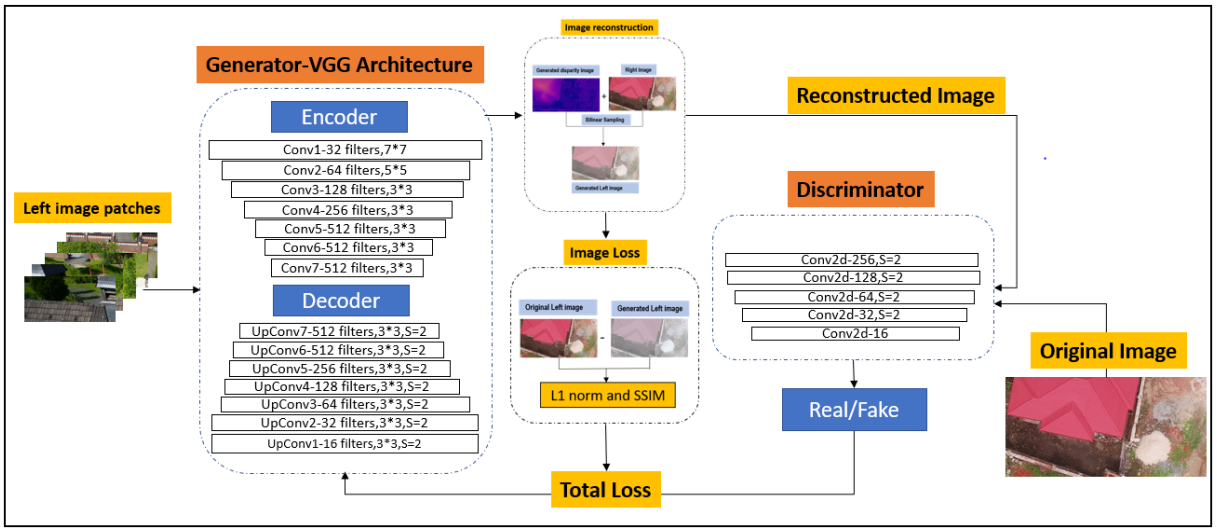


Figure 12. GAN architecture with Generator and Discriminator loss

3.3.3. InfoGAN

X. Chen et al., (2016) suggested that maximising the use of mutual information with simple modifications in GAN network resulted in interpretable representations. This is implemented by adding a regularization term to the GAN equation as shown in equation (7),

$$\min_{G,Q}\max_{D} V_o(G,D)=V(G,D) - \lambda L(G,Q) \quad (7)$$

Where V(G,D) represents the generator and discriminator loss terms, L(G,Q) is the mutual information between the generator output and the latent code. The latent code is the input to the second network which estimates the additional parameter. This information along with the generator output acts as extra information to calculate the required information. The latent code can be generated in an unsupervised manner based on statistical distribution. The added mutual information maximises the learning between generator input and output and brings meaningful information within data.

Implementation:

GAN architecture needs additional information to improve its disparity generation mechanism. Similar to an infoGAN architecture, the mutual information is to be provided through a third network for increasing its performance. This mutual information could be any information like semantics, gradients etc that can be generated easily and also serve as a complementary cue for depth generation. The selection of right information will depend on using different details and assessing the model performance. For this study, the additional information chosen to be included is disparity produced from stereopairs instead of just a single image used to produce disparity. GAN shows the capability of reproducing features from the original image but lack at the task of disparity generation. The network architecture is similar to GAN with a generator, discriminator and a third network for mutual information similar to a generator. The third network is based on VGG architecture consisting of downsampling and upsampling layers, which takes both left and right image as input. Both images are concatenated which passes through filters of decreasing kernel size. Repala & Dubey, (2018) and Pilzer et al., (2018) showed that disparities produced by deep learning models with stereopairs are better than that of models with single images. Hence disparities using stereopairs from the third network is used to better train the model to produce disparities from single images. The loss used is shown in equation (8). The image loss is also used here, as we deal

with two disparity images and we are required to calculate the similarity between the two for improving the other.

$$Stereo\ Image\ Loss = \frac{1}{N}\sum_{i,j} \alpha \frac{(1-SSIM(S_{i,j}^{\beta}, D_{i,j}^{\beta})}{2} + (1-\alpha)||S_{i,j}^{\beta} - D_{i,j}^{\beta}|| \quad (8)$$

Where α represents the weight between L1 norm and SSIM, S denotes the disparity image produced by stereopair and D represents the disparity image produced from single image, β={l,r} for left and right images and i,j represents pixel position. The model architecture along with the loss component is shown in Figure 13.

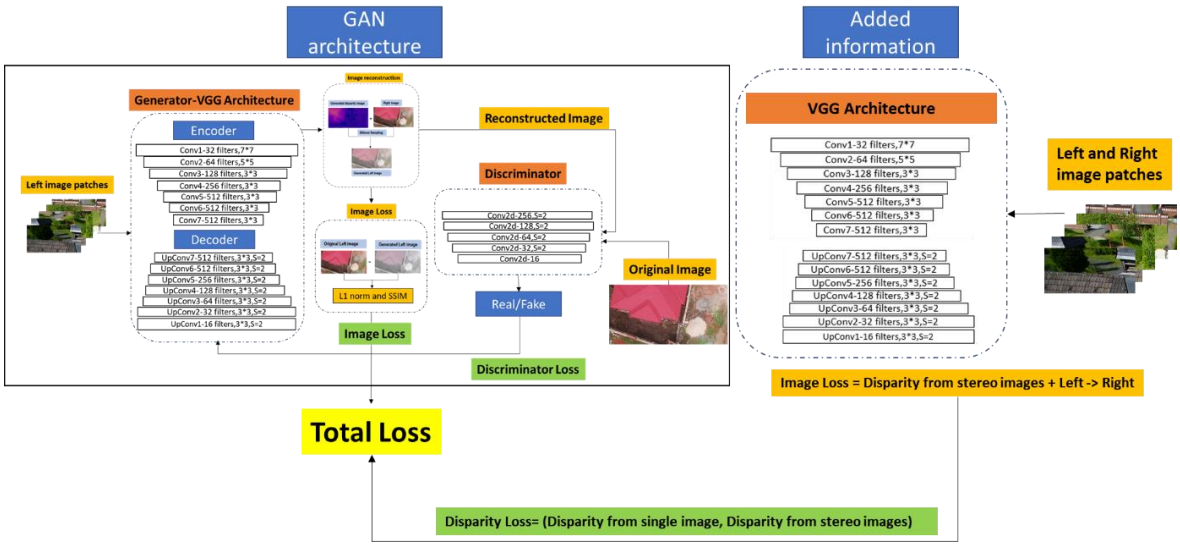


Figure 13. Proposed InfoGAN architecture with third network

The hyperparameters for this model includes the weightage for the third loss for producing the disparities of stereo images and also for the difference between disparities from stereopairs and disparities from single images. The learning rate, batch size are similar to the previous networks. The duration for training the model is around 38~39 hrs due to increased computation for the third network. Also, the network is modified with ResNET50 architecture to assess its performance based on increased complexity. Although the number of parameters increased with architecture, the performance of the model remained the same and didn't improve much.

3.3.4. InfoGAN WITH GRADIENTS

Along with the third network, the use of extra information like gradients is also explored. van Dijk & de Croon, (2019) discussed about various information neural networks sees while computing depth from monocular images. One among that is the edge information for computing depth from single images. In order to obtain edge information which are boundaries of objects, gradients from both horizontal and vertical direction are calculated using derivatives as shown in equation (9) and (10). Vertical edges are calculated from the horizontal gradient and horizontal edges are calculated from vertical gradients by taking differences between columns and rows in images respectively.

$$Horizontal\ edges\ I\_x = (I(x+1, y) - (I(x-1, y))/2 \quad (9)$$

$$Vertical\ edges\ I\_y = (I(x, y+1) - (I(x, y-1))/2 \quad (10)$$

Where I(x,y) represents the pixel position. The gradients in both x and y direction are calculated for the original and the generated left image. This is compared using mean squared loss and SSIM as specified in equation (3).

$$\text{Mean Squared Error } = (G_\beta - G'_\beta)^2 \tag{11}$$

Where G_β represents gradient calculated from generated left image, G'_β represents gradient calculated from the original left image, $\beta = \{horizontal, vertical\}$ edges. Figure 14 shows a training image with its horizontal and vertical gradient. Also, the hyperparameters include the weightage of this new gradient loss along with the total loss of the model.

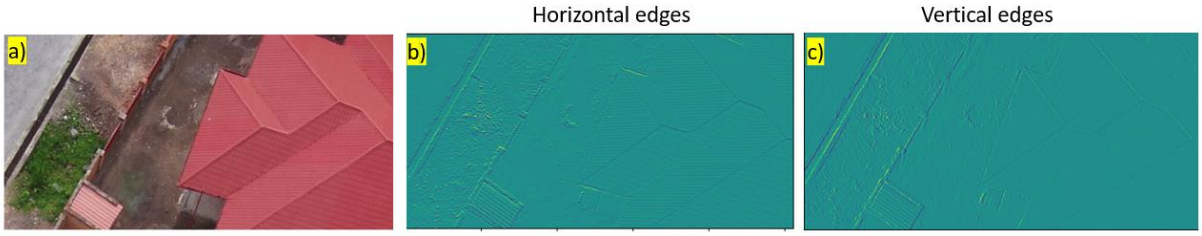


Figure 14. a) Original image b)Vertical gradient c) Horizontal gradient

3.4. GROUND TRUTH REFERENCE

To estimate the accuracy of the depth from the single image depth estimation models, point clouds generated using reliable photogrammetric tools are used. To generate point clouds, multiple UAV images of the same scene are captured and processed using dense image matching techniques. The generated point clouds are used to produce a smooth DSM and are then used for quantitative assessment with the predicted depths from the models. The reference point clouds are produced from two reliable sources PIX4D ("Pix4D (version 4.4.12)," 2020) and SURE ("SURE (version 4.1)," 2020).

PIX 4D mapper (4.4.12) :

PIX 4D is proprietary software which uses images captured from various devices to generate point clouds, DSM and orthomosaics. The processing includes three steps. The initial step is to determine the camera position and orientations for aligning images. The full image resolution is used for initial processing. The next step is to generate point clouds for which also the original resolution of images is used. The final step is to generate DSM and orthomosaics from the point clouds of optimal density. The outputs are saved as *.las format. These are further scaled and extracted into patches similar to the test images used in this study for comparison. This software mostly acts as "black box" in specifying about the processing going within the models (Niederheiser et al., 2016).

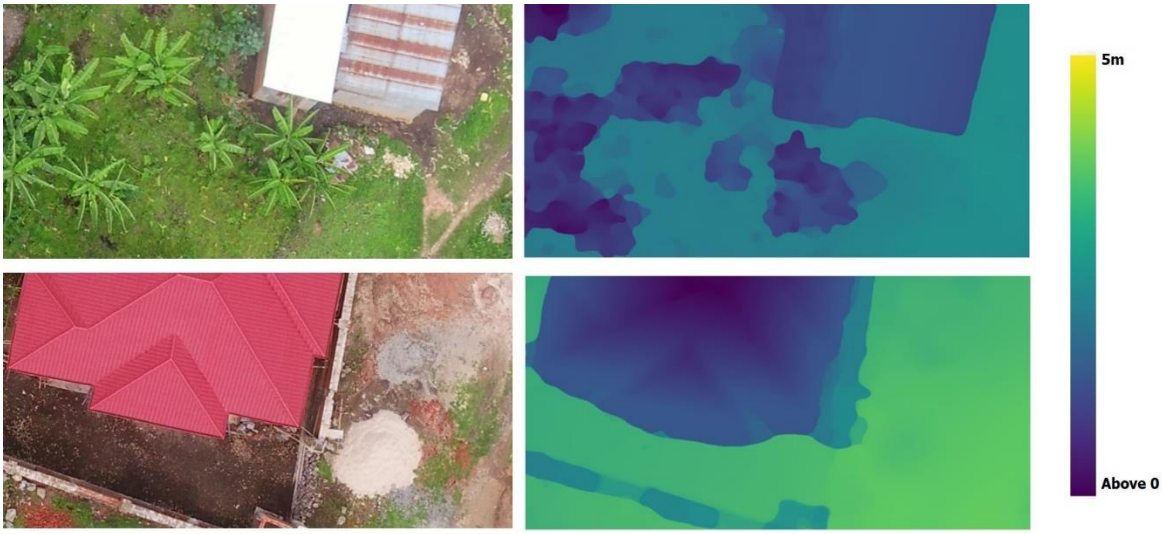


Figure 15. Generated DSM - PIX 4D

SURE (4.1) :

Rothermel, Wenzel, Fritsch, & Haala, (2012) used the multi-view stereo method for dense image matching. The camera orientations are loaded from PIX 4D. From the oriented images, surface reconstruction is carried out. It takes a single image as a base image and compares it with other proximate images for potential overlap to form stereo models. For each stereopair epipolar images are generated. It uses semi-global matching (SGM)(Hirschmüller, 2005) for matching stereopairs and to calculate disparities dynamically. The dense matching algorithm finds corresponding pixels of the same object in stereopairs which involves disparity generation by minimizing the global cost function as shown in equation (12).

$$E(D) = \sum C(x_b, D(x_b)) + \sum P_1 T[||D(x_b) - D(x_N)|| = 1] + \sum P_2 T[||D(x_b) - D(x_N)|| > 1] \quad (12)$$

Where the global cost function is calculated from the disparity estimation from base image pixels x_b. E is composed of data term $C(x_b, D(x_b))$ for calculating pixel-wise dissimilarity measure of a pixel with disparity and the two terms for claiming smooth surfaces where P_1 and P_2 are penalty parameters. It is used to calculate disparity maps pixel-wise by measuring the similarity of each pixel in one stereo image with the other stereo image. The SGM approach is modified to include dynamic disparity search ranges. Then depth is estimated within the stereo models redundantly. The redundant information is used to increase accuracy and eliminate noises. Finally, the depth information is obtained from single stereo models or from disparities merged together from the stereo models with the same base image through triangulation. The point density is higher compared to PIX 4D and also the time for generation is reduced significantly.

Figure 16. Sample Ground truth Test images – SURE

Both models are run in a system configuration of 8Gb memory capacity. The dataset with around 300 images takes around 14~16 hrs in PIX4D and 4~6 hrs in SURE for DSM generation. DSM generated from both PIX 4D and SURE software's are in the similar range as shown in Figure 15 and 16. Yet the DSM from SURE is much sharper and edges are clearly identifiable as compared to DSM from PIX 4D. Hence for further comparison of the reference ground truth depth with deep learning models results, DSM from SURE is used. For the reference depth, 50 images are selected from the dataset and depths generated from SURE are verified manually. The DSM generated are with reference to the WGS84 coordinate system. This is converted to relative depths to ease the comparison with the results from deep learning models. The conversion of the disparity images produced by models to depth are explained in Chapter 4.

4. RESULTS AND DISCUSSIONS

The disparities generated from all models (CNN, GAN, InfoGAN) are compared with each other for qualitative understanding and also with ground truths for quantitative estimation. CNN model took 25 hours and GAN model took around 29 hours for training. The improved GAN model, InfoGAN and InfoGAN with gradients framework took around 39 hours and 43 hours for training respectively. In order to assess the performance of the models, 600 single images are tested to generate disparities which are further converted to depths. The testing of 600 single images for CNN model takes 30 seconds for generating 600 disparities with 0.05s/image. While for GAN, InfoGAN and InfoGAN with gradients models, the testing takes around 35~40 seconds for generating 600 disparities. All the models are implemented and tested in a system with a single Nvidia Titan Xp GPU of 16GB memory. This chapter describes the results of each model separately. Then the disparities from both the models are intercompared and also they are converted to depth and compared with reference depth from SURE. Based on the results from GAN, InfoGAN model is suggested and the results of InfoGAN with disparity information and gradient information are discussed in the next section. In the last section, all models are intercompared and with reference depths. This chapter is concluded with discussion section.

4.1. CNN

The trained CNN model is used to generate disparities from single images and some of the generated disparities are shown in Figure 17. From the qualitative perspective, it is observed that the generated disparity images are smooth with large disparity range variation. The model tries to reproduce the original image by warping the disparity and other stereopair. The image loss as given in equation (3) is used in this model as backpropagation loss for producing incrementally better images of closer approximation with the original image. With every epoch in the training where the model reproduces better left image, the disparity map gets better. As the disparity is inversely related to depth, the lower the disparity values, the farther the object is to the camera and vice versa. In Figure 17, the disparity results can be interpreted as yellow regions (top of the roof) of high disparity are closer to the camera compared to blue regions (ground) of low disparity. The smooth transition from yellow to blue regions is also visible. The local variations in the ground surface are difficult to differentiate. The small shift in the roof, vegetations are shown in disparity maps yet the generated disparities are smoother, with fine edge information over the roofs is lost.

Figure 17. Generated single image disparities - CNN

4.2. GAN

The trained GAN model with fixed parameters are used to generate disparities from single images and some examples are shown in Figure 18. From the qualitative perspective, it is observed that the generated disparity images reproduce fine details like edges similar to original images. From Figure 18, it could be seen that the colour variations are slightly shifted towards blue compared to CNN. On further inspection of the disparity maps, it is seen that there are very low disparity values (noises and spikes) in some regions. This noise can be more clearly seen in depth image shown in Figure 20. This kind of noise is almost always seen only in ground regions and not in the roof. This leads one to believe that the GAN model's ability to produce disparities with distinguishable features like roof which has prominent details is much better compared to ground regions with no distinguishable features. The model shows the difficulty in learning this ground information compared to the features with structures. The GAN model uses both generator and discriminator loss as specified in equation (4)-(6) for learning. The combination of the two losses, helps the model learn to reproduce the disparity images but the noises in the ground does not seem to affect the losses as much as to influence training. The losses keep on reducing throughout the learning process and saturates before the noises could be eliminated as other parts of the image like the roof has been learnt well.

Figure 18. Generated single image disparities - GAN

4.3. INTER-COMPARISON BETWEEN CNN AND GAN MODELS

The generated disparities from the two models are intercompared as shown in Figure 19. From qualitative observation, it could be seen that the disparity produced from CNN shows colour variations from yellow (high disparity) to blue (low disparity) while GAN disparity images are shifted towards the blue side of the spectrum. This difference in the colour variations can be attributed to more noises and spikes in GAN model in ground areas. Also, disparity images generated by CNN are smoother with lesser noises compared to GAN. While the disparities produced by GAN model shows fine details like roof edges, more sharply than CNN. However, the regions with shadows and low vegetations are difficult for both models to discriminate.

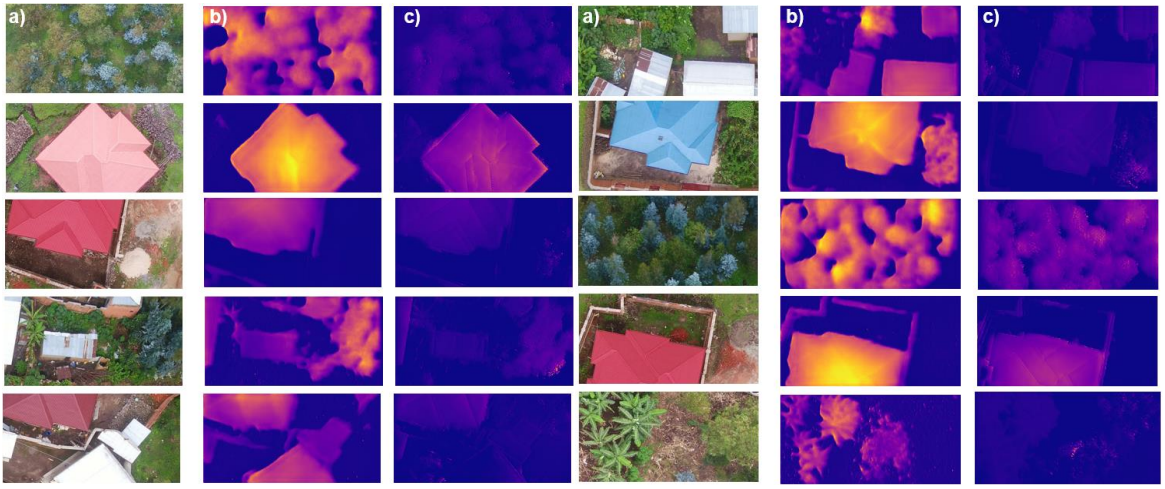


Figure 19. I- Generated single image disparities from CNN and GAN- a) original image -Rwanda, Africa b) CNN result c) GAN result

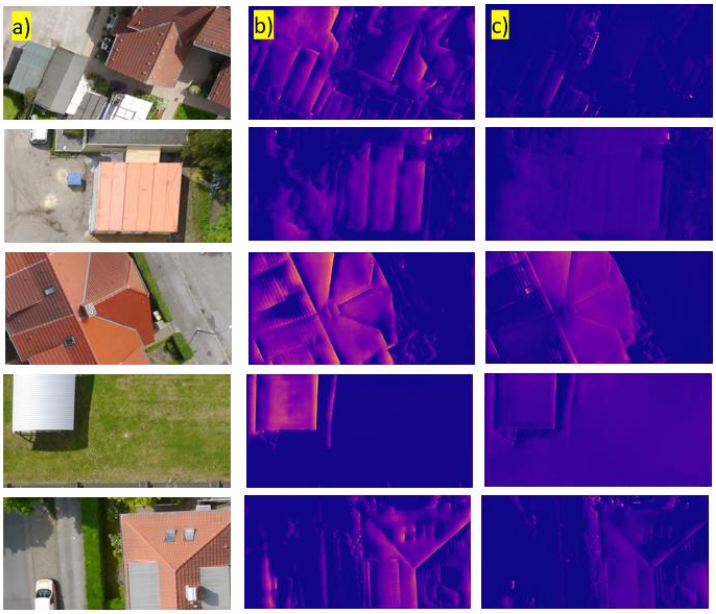


Figure 19. II- Generated single image disparities from CNN and GAN- a) original image -Zeche, Germany b) CNN result c) GAN result

The disparities produced from the two models are converted to depths using equation (1), where depth is related to disparity inversely. The dataset includes images with different focal length and baseline parameters and also involves a series of pre-processing steps that may affect these parameters. Hence a scale factor is used to convert disparities into depths as give equation (13).

$$Depth = \frac{Scale\ factor}{disparity} \tag{13}$$

To find the value of scale factor reference depths from SURE is used. In order to compare SURE depths with the depths from SIDE models, relative depths are required. To convert the depth from SURE into relative variations for comparison, the least height within each patch is subtracted from the entire patch thereby shifting the datum to the lowest point within the field of view. From SURE 20 reference images are randomly chosen and compared with corresponding disparity maps to find the scale factor. The

optimal scale factor is the one which produces the least value in all of the error metrics. The obtained depths from SIDE models are relative depths that can be converted into absolute depths by using ground control points.

To assess the performance of both the models quantitatively, various evaluation metrics are used. The produced depths from models along with reference depths are shown in Figure 20. For producing the metrics, 50 reference ground truth images taken from SURE are tested with both models. The evaluation of the accuracy of the generated depth is done using error metrics adopted by several previous works of similar nature. D'(x) represents the depth generated during testing by models and D(x) represents the ground truth depth produced from SURE. The error metrics includes Absolute Relative difference (Abs Rel) given in equation (14), Squared Relative difference (Sq Rel) given in equation (15), Root Mean Square Error (RMSE) given in equation (16), RMSE log (17) and d1-all given in equation (18) as reported from (Amirkolaee & Arefi, (2019);Repala & Dubey, (2018);Aleotti, Tosi, Poggi, & Mattoccia, (2019).

$$Abs\ Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D(x_i)-D\prime(x_i)|}{D(x_i)} \tag{14}$$

$$Sq.Rel = \frac{1}{N}\sum_{i=1}^{N}\frac{|D(x_i)-D\prime(x_i)|^2}{D(x_i)} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D(x_i) - D'(x_i))^2} \tag{16}$$

$$RMSE\ log = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(D(x_i)) - log(D'(x_i)))^2} \tag{17}$$

$$D1 - all = \frac{1}{n}\sum bad\ pixels * 100 \tag{18}$$

Where, bad pixels are those which satisfies the condition, $|D(x_i) - D'(x_i)| >= 3\ and\ (|D(x_i) - D'(x_i)|)/D(x_i) >= 0.05$. The lower the value of these metrics, the better the quality of generated depth maps. The units of depths as well as evaluation metrics are in meters.

Table 2. Metrics on the external accuracy between the depth image from the models (CNN , GAN)and the reference depth (in meters)

Method	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all
CNN	0.693	1.772	1.852	1.898	19.78
GAN	0.775	2.131	2.165	2.009	25.80

In Figure 20-I the relative depths are within a range of 0-5m in the reference data from SURE. The depths obtained from the models are of a range of 0-6m with some noises or spikes reaching a value more than 8m. Based on the comparison of metrics with ground truth as shown in Table 2, it is observed that CNN produces depth that is much closer to reference depths than GAN. GAN is found to produce depth

images with noises and spikes near ground regions due to its difficulty in interpreting indistinguishable features.

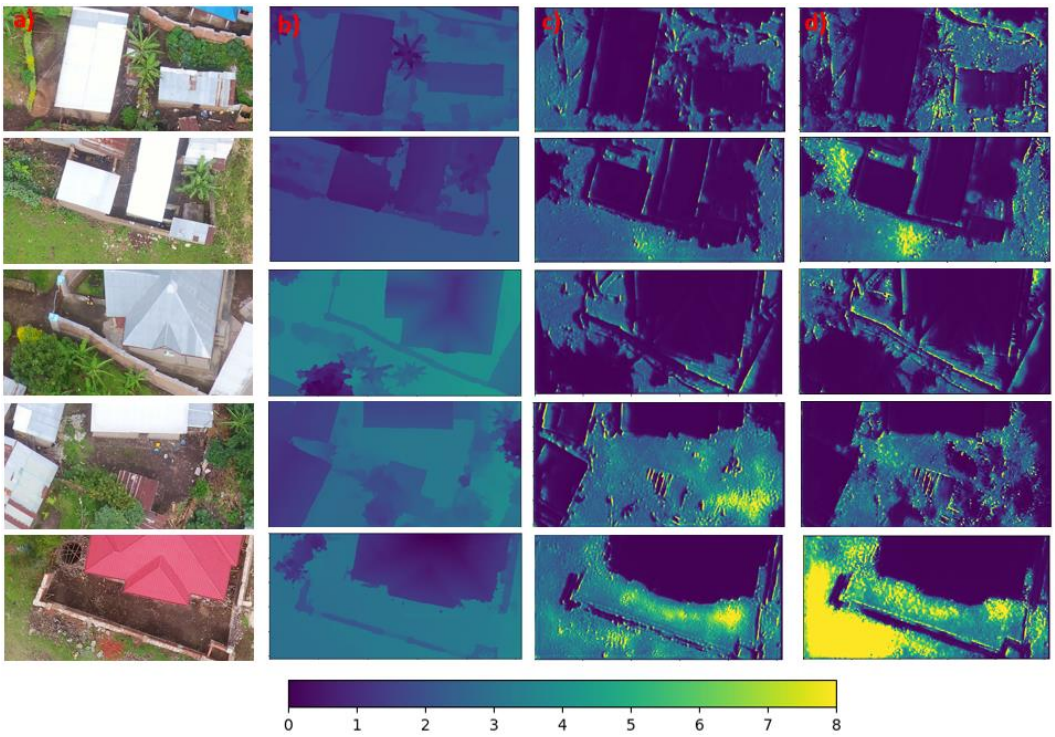


Figure 20. I-Produced single image depth (in meters) -a) Original image-Rwanda, Africa b) Reference depth from SURE c) CNN depth d) GAN depth

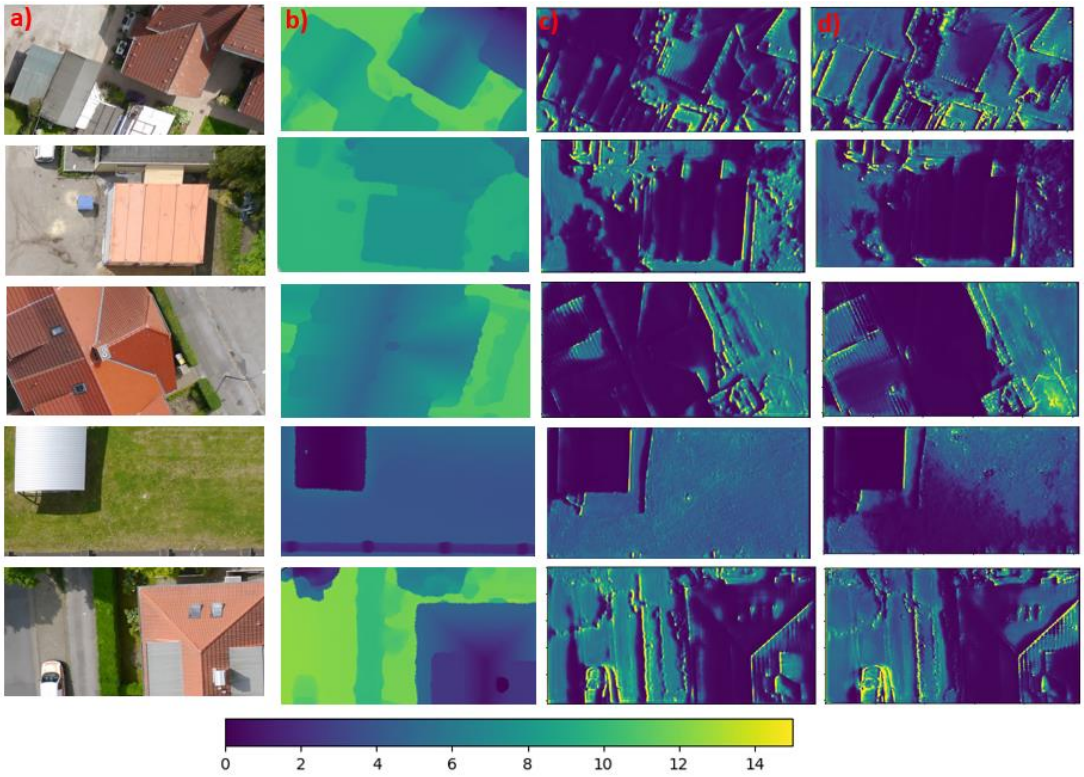


Figure 20. II-Produced single image depth (in meters) -a) Original image-Zeche, Germany b) Reference depth from SURE c) CNN depth d) GAN depth

In Figure 20 II, the relative depths are within a range of 0-12m in the reference data from SURE. The depths obtained from the models are of a range of 0-14m with some noises or spikes more than 14m. To further understand the difference between the generated depths from the model and the reference depth from SURE, the relative absolute pixel-wise difference for each image is required.
The difference between the reference depth and the model depth for CNN and GAN is visualised in Figure 21. It is observed from Figure 21 that the difference between reference depth and model depth does not exceed more than 2m for both CNN and GAN. The colours here denote the depth difference, where yellow denotes underprediction and blue denotes overprediction of depth by the SIDE model in comparison with SURE. The difference map for GAN model shows that more ground regions are over predicted and assume higher depth values than they are which is consistent with the noise creation in GAN that is explained in previous sections. In most of the images, the values stick closer to 0.5-1m for both CNN and GAN model.

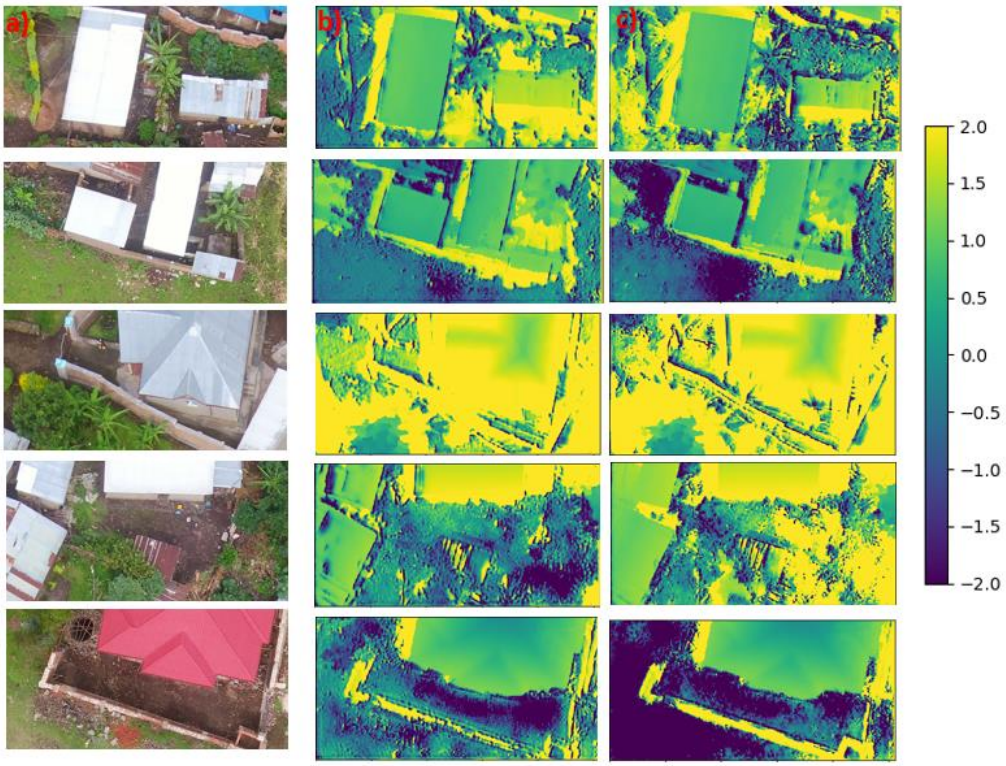


Figure 21. Absolute difference between reference depth and model depth(in meters) -a) Original image-Rwanda, Africa b) CNN depth c) GAN depth

From the metrics, it is observed that CNN performs better in depth generation compared to GAN. Even though CNN produces better depths from single images GAN produces fine details and has scope for further improvement. To improve the performance of GAN the modification as specified in InfoGAN section is made to the network architecture and the results are explained below.

4.4. InfoGAN

As explained in chapter 3, InfoGAN uses mutual information as complementary cues for disparity generation. This includes information like disparities from stereo images, gradients to guide the network

towards disparity generation with single images. Disparities generated by InfoGAN with additional information from stereo images and with gradient information are shown in Figure 22. The disparity from InfoGAN shows improvement in the generation of disparity compared to the simple GAN model. The extra information provided by stereo images is found to generate significantly better quality disparities and also reduce noise. However, adding gradient information along with the extra information from stereo images didn't show any further improvement. The provided gradient information tends to smooth the image in areas other than edges thereby further reducing its capability to differentiate ground surfaces.

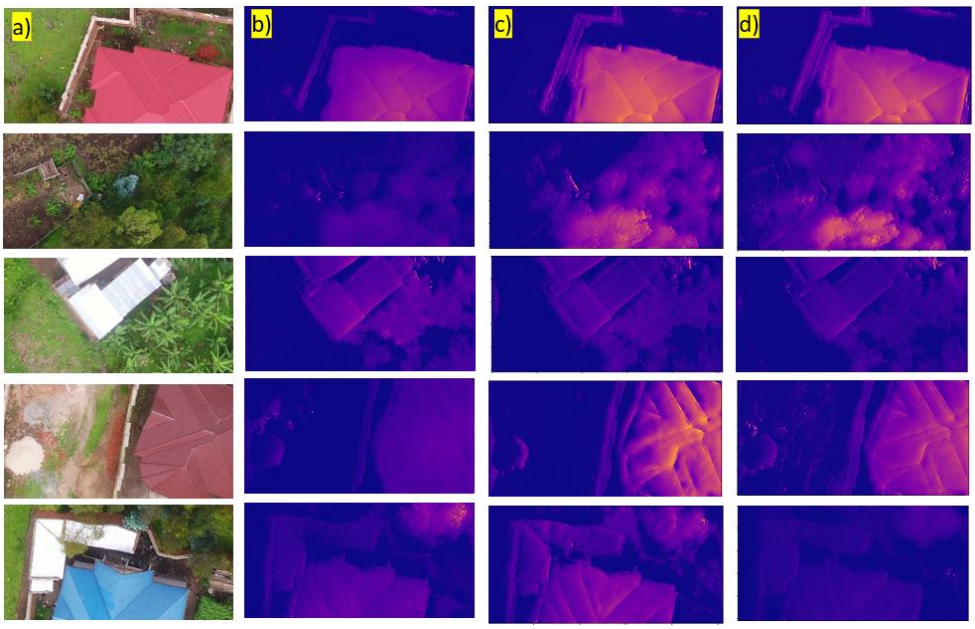


Figure 22. I-Model disparity results a) Original image-Rwanda, Africa b)GAN c)InfoGAN d)InfoGAN with gradients

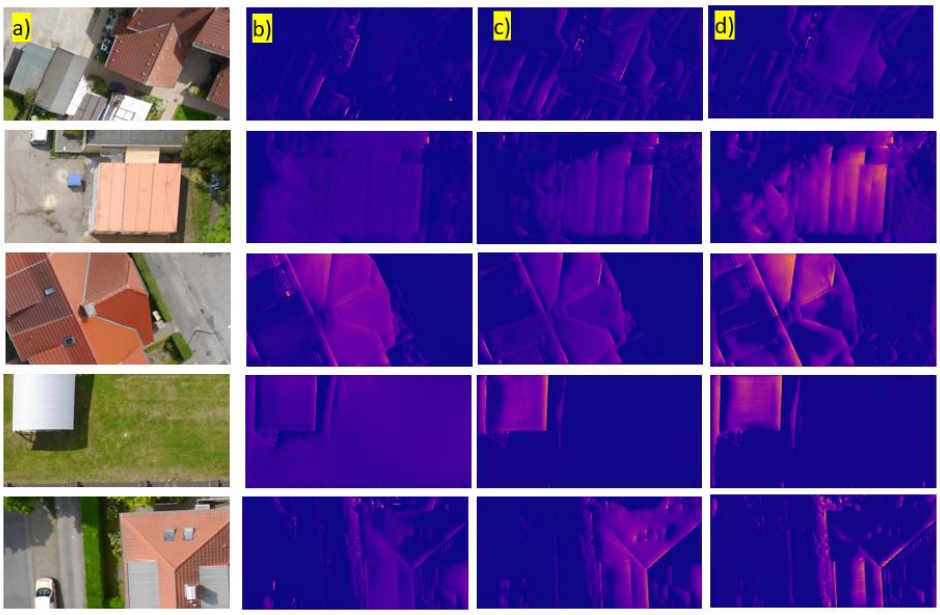


Figure 22. II-Model disparity results a) Original image-Zeche, Germany b)GAN c)InfoGAN d)InfoGAN with gradients

From Figure 22, it could be observed that the disparity generation for both InfoGAN and InfoGAN with gradients has improved in comparison to simple GAN model. The colour variations appear almost similar to the disparity generated from CNN. The noises generated in GAN near the ground surface has also reduced which can be seen in Figure 23. The improvement from GAN is quantified using the same metrics given in equation (14-18). The disparities are converted to depths for comparison with the ground truth reference from SURE.

4.5. INTER-COMPARISON BETWEEN ALL MODELS

The evaluation metrics are shown in Table 3. From Table 3 it can be seen that CNN produces better results on all metrics. GAN results show significant divergence in metrics compared to CNN. However, InfoGAN shows better performance than GAN in all metrics and also shows better performance than CNN in Absolute Relative difference and Squared Relative difference. This means that InfoGAN not only shows improved performance than GAN but also gives results that are comparable with CNN. This validates the approach of adding mutual information to GAN architecture to improve the performance of the model in single image depth estimation tasks. However, including gradient information, the results from InfoGAN did not show much improvement. This shows that while the framework of InfoGAN i.e using mutual information through a third network to increase the network performance can indeed work desirably, finding the right complementary cue is critical for improving the model performance. This study used disparity from stereopairs and gradients as mutual information, while disparity from stereopairs showed an improvement, gradients didn't show much improvement in model performance. Different mutual information through multiple networks can be added for improving the performance of InfoGAN will be scope for future studies. The depths along with the reference data are shown in Figure 23.

Table 3. Metrics on the external accuracy between the depth image from the models (InfoGAN) and the reference depth (in meters)

Method	Abs Rel	Sq Rel	RMSE	RMSE log	D1-all
CNN	0.693	1.772	1.852	1.898	19.78
GAN	0.775	2.131	2.165	2.009	25.80
InfoGAN	0.666	1.60	1.87	1.91	20.43
With_Gradients	0.73	1.97	2.05	2.02	24.3

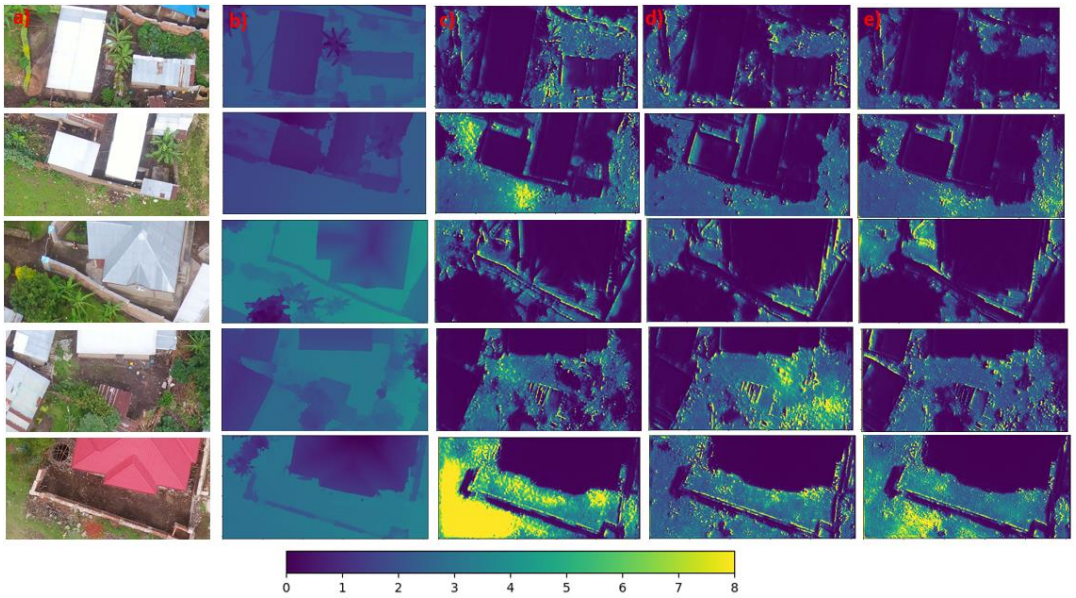


Figure 23. Produced single image depth -a) Original image b) Reference depth from SURE c)GAN depth d) InfoGAN depth e)InfoGAN with gradients depth

Also, to find further relationship between different model results and reference depth, their average mean and standard deviation are calculated.

Table 4. Metrics on the external accuracy between the depth image for all models and the reference depth (in meters)

Model	Average Mean	Standard Deviation
SURE	2.98	0.72
CNN	2.74	0.84
GAN	3.21	1.22
InfoGAN	2.70	0.87
InfoGAN with gradients	3.01	1.09

The values of mean and standard deviation for all models and the corresponding reference depth are shown in Table 4. The mean and standard deviation of CNN and InfoGAN are more closer to SURE than other models.

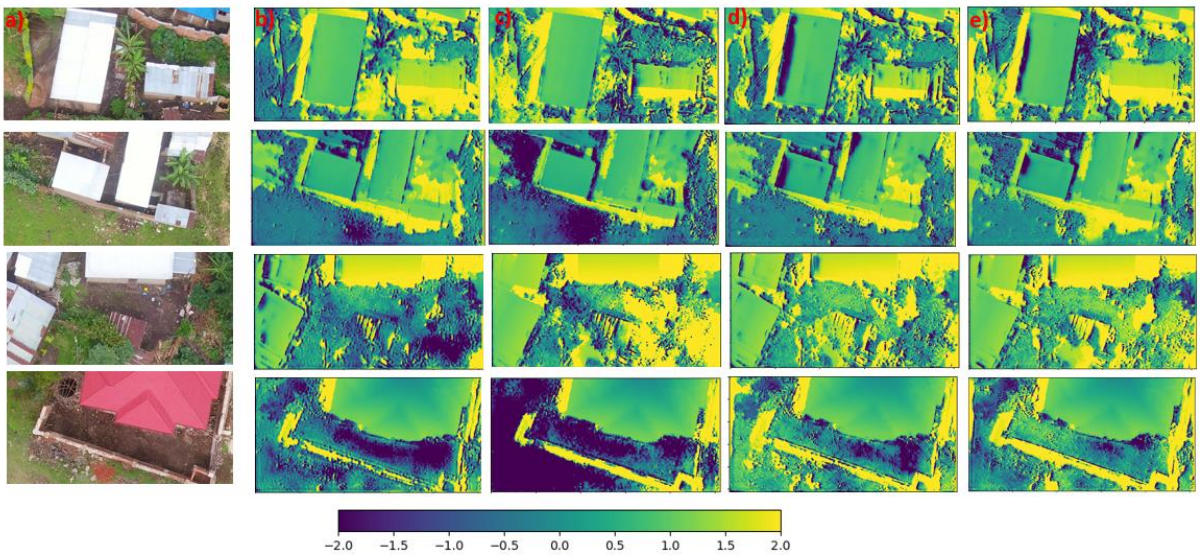


Figure 24. Absolute difference between reference depth and model depth(in meters) -a) Original image-Rwanda, Africa b) CNN depth c) GAN depth d) InfoGAN depth e) InfoGAN with gradients

From Figure 24, the absolute difference between various models and the reference depth from SURE can be seen. It could be observed that the overprediction by GAN model in ground regions, has shown a significant reduction in InfoGAN model. This also shows that the use of mutual information in the form of disparity from stereo images has proven to improve the single image depth estimation model.

4.6. DISCUSSIONS

All the models discussed in this study are tuned for learning rates, number of epochs and weightages of different loss terms to produce these results. Both CNN and GAN models are able to produce realistic disparities from single UAV images. Whereas the disparity produced by CNN are smoothed out, GAN produces disparity maps that reproduces fine edges. It can be clearly seen that the GAN model is better at reproducing sharp edges and distinguishable features than CNN. However, there are a significant amount of noises in the GAN model at regions where there are a lack of distinguishable features, especially at ground level. The GAN model with a composition of generator and discriminator structure uses image loss and discriminator loss. The difference in performance between GAN and CNN may be attributed due to the discriminator network in GAN model which focusses on finding the probability of the generated image being the original image. While this architecture has a beneficial impact in reproducing sharp edges as in the original image, it could not reduce the noises at the ground level. The loss terms reach stability (for different weights) before the noises in the ground surface are removed as distinct features like roofs occupy a significant portion of the patch. In contrast, the simple architecture of CNN with just image loss seems to be better in producing single image disparities.

The framework of InfoGAN which is to provide mutual information (depth complementary cues) through a third network to better the model performance shows promising results. The use of disparity derived from stereo images as mutual information significantly improved results from that of GAN. Based on the evaluation metrics it can be seen that the results are comparable to that of CNN. An attempt to further improve the performance of InfoGAN by providing gradients as mutual information did not yield the desired improvement and rather degraded the performance. This may be due to the nature of gradient information which is to smooth out regions surrounding the edges. This means that while the framework

of InfoGAN can indeed improve the performance of single image depth estimation, finding the appropriate mutual information to act as the complementary cue is critical.

The depth produced from all SIDE models is compared with depth produced from SURE. An absolute relative difference of 0.6 to 0.9m and RMSE of 1.8 to 2m is achieved. Extracting depth from a single UAV image is challenging due to the limited information available. The problem can be compounded due to the errors introduced in various stages of pre and post-processing. Various complementary depth cues can be used as mutual information through additional networks within the framework of InfoGAN to increase the model performance even further.

The use of scale factor instead of baseline and focal length may not be the optimal solution for the conversion of disparity to depth. However, due to the inaccessibility of georeferenced images (due to the prevailing Covid-19 pandemic) methods other than the iterative scaling used here cannot be tested for conversion to depth. This can be taken up once the access is gained to these images so an optimal solution for the conversion from disparity to depth can be found.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. CONCLUSIONS

The primary objective of this study is to find a deep learning model that can estimate depth from single UAV images in a self-supervised manner. A dataset with UAV images covering different regions is prepared for training the deep learning models. The dataset preparation involves undistortion and rectification of the images to form stereopairs. From the dataset, 22000 stereo patches are extracted for training. Two deep learning models, CNN and GAN are trained with the UAV dataset. Both models have different architecture yet work on a similar training mechanism. It involves generating disparity from one image of the stereopair then warping the disparity with the other stereopair to reconstruct the original image and then backpropagating the losses between the original and reconstructed image. Both the models converge and can produce disparity maps from single images. The disparities are converted to depth using the inverse relationship and a scale factor. On comparison, GAN can reproduce distinguishable feature but it could not learn to reduce noise at ground level. So, to improve the GAN model, an InfoGAN framework is adapted where additional information in the form of depth through a third network is added to the architecture. This study uses disparity from stereo images and gradient information as mutual information. To assess the performance quantitatively, reference DSMs are generated from SURE for comparison with the models. It is seen that CNN and InfoGAN perform better than other models based on the evaluation metrics. Within the InfoGAN framework, the use of disparity from stereo images as mutual information increased the model performance, the use of gradient information, in addition, did not increase the performance any further. Hence finding the right depth cue is important for such task and could be critical in deciding the performance of the network. The following key conclusions are arrived from this study.

- CNN model is better than GAN in terms of producing disparities with a simple architecture.
- While GAN model shows promise in regions with distinguishable features within the scene it produces a significant amount of noises in other regions.
- The framework of InfoGAN has not only improved the performance from GAN but also has the potential for further improvement through the use of additional network and different depth cues.

This leads us to believe that CNN with a simpler structure is well capable of SIDE and InfoGAN framework with right mutual information is having a lot of capacity for improving the performance of single UAV image depth estimation task.

The results demonstrate the effectiveness of the proposed SIDE model with UAV images. The model can be used in places where we need the depth information with the limited images captured beforehand and also with images that do not have camera calibration parameters. With the accuracy achieved in this study, the proposed SIDE models can be used in areas where a rough estimate of the terrain is needed prior to the actual planning of the main survey. In situations where homologous features between images could not be matched, the SIDE models could be an alternate solution. With SIDE it becomes possible to estimate depth on the fly opening up applications in visual odometry, SLAM etc. The SIDE models (static) can also be used as a base for video-based depth estimation. This is one of the first studies to do self-supervised SIDE of UAV images and with the proposed InfoGAN framework there is a significant

potential for increasing the accuracy of SIDE. SIDE models with better accuracy can be used in building DSM, 3D reconstruction, object detection and other scene understanding tasks.

5.2. RECOMMENDATIONS

The main recommendations from this research will be:

- Different depth cues can be tried out to find the right mutual information for improving the performance of SIDE within the framework of InfoGAN. This can improve model performance and increase the accuracy of the proposed models.
- Depth information is also being used in applications like semantic labelling and object detection. Depth extracted from SIDE models can be used for these applications and their effectiveness in comparison to other methods of depth extraction for the same applications can be compared.
- One of the limitations of deep learning techniques is its transferability over different geographical regions. The model that learnt using images from certain regions may not be able to perform well on images from totally different regions. Transfer learning which is a process of using the knowledge learnt in one problem to solve the similar related problem is a possible way of overcoming this issue. The ability of the proposed SIDE models under transfer learning with different datasets should be tested.
- Ways of replacing the scale factor with analytical solutions involving baseline and focal length can be looked into. With access to georeferenced images different methods, the use of scale factor or the use of baseline and focal length, to convert disparity to depth can be compared to the find the optimal solution.

LIST OF REFERENCES

Aicardi, I., Nex, F., Gerke, M., & Lingua, A. M. (2016). An image-based approach for the Co-registration of multi-temporal UAV image datasets. *Remote Sensing*, *8*(9), 1–20. https://doi.org/10.3390/rs8090779

Alagoz, B. B. (2016). A Note on Depth Estimation from Stereo Imaging Systems. *Anatolian Science*, *1*(1), 8–13.

Aleotti, F., Tosi, F., Poggi, M., & Mattoccia, S. (2018). Generative Adversarial Networks for Unsupervised Monocular Depth Prediction. In *ECCV* (pp. 337–354). https://doi.org/10.1007/978-3-030-11009-3_20

Aleotti, F., Tosi, F., Poggi, M., & Mattoccia, S. (2019). Generative adversarial networks for unsupervised monocular depth prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11129 LNCS*, 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. Retrieved from http://arxiv.org/abs/1905.07542

Amirkolaee, H. A., & Arefi, H. (2019). Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 50–66. https://doi.org/10.1016/j.isprsjprs.2019.01.013

Chen, R., Mahmood, F., Yuille, A., & Durr, N. J. (2018). Rethinking Monocular Depth Estimation with Adversarial Training, 10. Retrieved from http://arxiv.org/abs/1808.07528

Chen, S., Tang, M., & Kan, J. (2019). Predicting depth from single RGB images with pyramidal three-streamed networks. *Sensors (Switzerland)*, *19*(3), 1–12. https://doi.org/10.3390/s19030667

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 2180–2188.

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, 1–9. Retrieved from http://arxiv.org/abs/1406.2283

Eisenbeib, H. (2009). *UAV Photogrammetry*. *ETH ZURICH*. https://doi.org/10.3929/ethz-a-010782581

El-Hakim, S. F. (2001). A flexible approach to 3D reconstruction from single images. *Acm Siggraph*, *1*(July).

Förstner, W., & Gülch, E. (1987). A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. *ISPRS Intercommission Workshop*.

Furukawa, Y., & Hernández, C. (2015). *Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision* (Vol. 9). https://doi.org/10.1561/0600000052

Garg, R., Vijay Kumar, B. G., Carneiro, G., & Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9912 LNCS*, 740–756. https://doi.org/10.1007/978-3-319-46484-8_45

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. Retrieved from www.deeplearning.org

Harris, C., & Stephens, M. (1988). A Combined Corner and Edge Detector, 23.1-23.6. https://doi.org/10.5244/c.2.23

Hirschmüller, H. (2005). Accurate and Efficient Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(2), 328–341. https://doi.org/10.1109/TPAMI.2007.1166

Hu, J., Zhang, Y., & Okatani, T. (2019). Visualization of Convolutional Neural Networks for Monocular Depth Estimation, 3869–3878. Retrieved from http://arxiv.org/abs/1904.03380

Huang, T., Zhao, S., Geng, L., & Xu, Q. (2019). Unsupervised monocular depth estimation based on residual neural network of coarse–refined feature extractions for drone. *Electronics*, *8*(10).

https://doi.org/10.3390/electronics8101179
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. *Tech. ReP.*., 1–7.
Junger, C., Hess, A., Rosenberger, M., & Notni, G. (2019). FPGA-based lens undistortion and image rectification for stereo vision applications. In *Proceedings of SPIE* (p. 11). https://doi.org/10.1117/12.2530692
Kanatani, K. ichi, & Chou, T. C. (1989). Shape from texture: General principle. *Artificial Intelligence*, *38*(1), 1–48. https://doi.org/10.1016/0004-3702(89)90066-0
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
Li, J., Yuce, C., Klein, R., & Yao, A. (2017). A two-streamed network for estimating fine-scaled depth maps from single RGB images. *Computer Vision and Image Understanding*, *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 Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 1–16. https://doi.org/10.1109/TPAMI.2015.2505283
Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
Mehta, I., Sakurikar, P., & Narayanan, P. J. (2018). Structured adversarial training for unsupervised monocular depth estimation. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, 314–323. https://doi.org/10.1109/3DV.2018.00044
Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets, 1–7. Retrieved from http://arxiv.org/abs/1411.1784
Mou, L., & Zhu, X. X. (2018). IM2HEIGHT: Height Estimation from Single Monocular Imagery via Fully Residual Convolutional-Deconvolutional Network. *CoRR*, 1–13. Retrieved from http://arxiv.org/abs/1802.10249
Nagai, T., Ikehara, M., & Kurematsu, A. (2007). HMM-based surface reconstruction from single images. *Systems and Computers in Japan*, *38*(11), 80–89. https://doi.org/10.1002/scj.10685
Nex, F., & Remondino, F. (2014). UAV for 3D mapping applications: A review. *Applied Geomatics*, *6*(1), 1–15. https://doi.org/10.1007/s12518-013-0120-x
Niederheiser, R., Mokroš, M., Lange, J., Petschko, H., Prasicek, G., & Elberink, S. O. (2016). DERIVING 3D POINT CLOUDS FROM TERRESTRIAL PHOTOGRAPHS - (Vol. XLI-B5, pp. 12–19). https://doi.org/10.5194/isprsarchives-XLI-B5-685-2016
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
Pix4D (version 4.4.12). (2020). Retrieved April 18, 2020, from https://www.pix4d.com/
Prados, E., & Faugeras, O. (2006). Shape from shading. In *Handbook of Mathematical Models in Computer Vision* (pp. 375–388). https://doi.org/10.1007/0-387-28831-7_23
Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 1–16.
Remondino, F, Barazzetti, L., Nex, F., Scaioni, M., & Sarazzi, D. (2011). UAV Photogrammetry for Mapping and 3D Modeling - Current Status and Future Perspectives, *XXXVIII*(September), 14–16. https://doi.org/10.5194/isprsarchives-XXXVIII-1-C22-25-2011
Remondino, Fabio, Spera, M. G., Nocerino, E., Menna, F., Nex, F., & Gonizzi-Barsanti, S. (2013). Dense image matching: Comparisons and analyses. In *Proceedings of the DigitalHeritage 2013 - Federating the 19th Int'l VSMM, 10th Eurographics GCH, and 2nd UNESCO Memory of the World Conferences, Plus Special Sessions fromCAA, Arqueologica 2.0 et al.* (Vol. 1, pp. 47–54). https://doi.org/10.1109/DigitalHeritage.2013.6743712
Repala, V. K., & Dubey, S. R. (2018). Dual CNN Models for Unsupervised Monocular Depth Estimation, 9. Retrieved from http://arxiv.org/abs/1804.06324
Rothermel, M., Wenzel, K., Fritsch, D., & Haala, N. (2012). SURE : Photogrammetric Surface

Reconstruction From Imagery, (May 2015).
Saxena, A., Chung, S. H., & Ng, A. Y. (2005). Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, 1161–1168.
Saxena, A., Chung, S. H., & Ng, A. Y. (2007). 3-D depth reconstruction from a single still image. *International Journal of Computer Vision*, *76*(1), 53–69. https://doi.org/10.1007/s11263-007-0071-y
Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In *ECCV-2* (Vol. 7576 LNCS, pp. 746–760). https://doi.org/10.1007/978-3-642-33715-4_54
SURE (version 4.1). (2020). Retrieved April 18, 2020, from https://www.nframes.com/
Szeliski, R. (1999). Prediction error as a quality metric for motion and stereo. In *Proceedings of the IEEE International Conference on Computer Vision* (Vol. 2, pp. 781–788). https://doi.org/10.1109/iccv.1999.790301
Szeliski, R. (2010). Chapter 6: Feature-based alignment. In *Computer vision: Algorithms and Applications* (pp. 311–320). Springer-Verlag Berlin, Heidelberg ©2010. https://doi.org/10.4324/9780429042522-10
Szeliski, R., & Zabih, R. (2000). *An experimental comparison of stereo algorithms. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1883). https://doi.org/10.1007/3-540-44480-7_1
Van Den Heuvel, F. A. (1998). 3D reconstruction from a single image using geometric constraints. *ISPRS Journal of Photogrammetry and Remote Sensing*, *53*(6), 354–368. https://doi.org/10.1016/S0924-2716(98)00019-7
van Dijk, T., & de Croon, G. C. H. E. (2019). How do neural networks see depth in single images? Retrieved from http://arxiv.org/abs/1905.07005
Wang, Z., Bovik, A. C., Sheikh, H. R., & P.Simoncelli, E. (2004). Image Quality Assessment:From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1139/cjce-2016-0381
Zama Ramirez, P., Poggi, M., Tosi, F., Mattoccia, S., & Di Stefano, L. (2019). Geometry Meets Semantics for Semi-supervised Monocular Depth Estimation (Vol. 11363 LNCS, pp. 298–313). https://doi.org/10.1007/978-3-030-20893-6_19
Zhang, L., Dugas-Phocion, G., Samson, J. S., & Seitz, S. M. (2001). Single view modeling of free-form scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*. https://doi.org/10.1109/cvpr.2001.990638
Zhang, R., Tsai, P. S., Cryer, J. E., & Shah, M. (1999). Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(8), 690–706. https://doi.org/10.1109/34.784284
Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 2242–2251. https://doi.org/10.1109/ICCV.2017.244