Super-resolution of Digital Elevation Models With Deep Learning Solutions

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

KUBADE ASHISH ASHOKRAO
20172075
ashish.kubade@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, INDIA
March 2021

Copyright © Ashish Kubade, 2021
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Super-resolution of Digital Elevation Models with Deep Learning Solutions"** by **Ashish Kubade**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Dr. Avinash Sharma
Date	Adviser: Dr. K. S. Rajan

To the penultimate try

Acknowledgments

Maybe the journey which is about to start has already been started. This journey started in Chennai, where I was motivated to return to academics and explore my dreams. I thank my roommates and my parents for supporting my decision.

After joining IIIT, there were Nitin Nilesh, Abhinav and Nishant Prabhu who were one-stop solution for my endless queries during the rst semester. I feel fortunate to have Dr. Avinash Sharma as my advisor during this journey whom I would rather say played a role of more than a guide, by not just guiding but also making sure to understand hurdles and problems by being in student's shoes. His timely motivations have kept this research going despite some rejections. When I was introduced to Dr. Rajan, who later become my co-advisor, I was a newbie to GIS. Often there was a situation where a newbie and a Subject Matter Expert discussing some experiments. Yet Dr. Rajan helped despite his tight schedule and solved my doubts answering which might be simplest the question he might have answered to his students. I would also like to thank Dr. Hrishikesh Sharma for guiding me during my internship at TCS Innovation Labs.

Sai Sagar who was a peer, a friend and a mentor, had always been there during the tough times of research life. Teasing by Neeraj and Alekh, tips from Yudhik and remote guidance from Abbhinav has made this journey quite easy and joyful. Dr. Dipti Patel, whose help in completing this research task is invaluable. I thank Naman Patel and Nikhilendra for their assistance in this research. I would also like to mention my the circle from my batch especially Ashutosh, Shyam, Vamsi, Murtuza for their help and interactions.

College days will never be complete without a canteen. Prabhashis Baba's followers, the nickname, whom I had always looked for when there were set-backs. Those tea-time discussions will always be missed. I would like to mention my seniors viz. Rashmi K, Tanashree J, Pulkit V, Kishori P, Parijaat for helping me understand how exciting research life could be. Last but not the least, I extend my gratitude to all the staff members of IIIT-H for timely help in administrative and nancial tasks.

Abstract

Terrain, representing features of an earth surface, plays a crucial role in many applications such as simulations, hazard prevention and mitigation planning, route planning, analysis of surface dynamics, computer graphics-based games, entertainment, lms, to name a few. With recent advancements in digital technology, these applications demand the presence of high-resolution details in the terrain. However, currently available public datasets, providing terrain scans in the form of Digital Elevation Models (DEMs) have low resolution compared with the terrain information available in other modalities like aerial images. Publicly available DEM datasets for most parts of the world have a resolution of 30 m whereas the aerial images or satellite images are available at a resolution of 50 cm. The cost involved in capturing of such high-resolution DEMs (HRDEMs) turns out to be a major hurdle for making such high-resolution available in the public domain. This motivates us to provide a software solution for generating high-resolution DEM from the existing low-resolution DEMs (LRDEMs).

In natural image domain, super-resolution has set up higher benchmarks by incorporating deep learning based solutions. Despite such tremendous success in image super-resolution task using deep learning solutions, there are very few works that have used these powerful systems on DEMs to generate HRDEMs. A few of them used additional modalities as aerial images or satellite images, temporal sequence of DEMs etc., to generate high-resolution terrains. However, the applicability of these methods is highly subject to the available input formats. In this research effort, we explore a new direction in DEM super-resolution by using feedback neural networks. Availing the capability of feedback neural networks to rede ne the features learned by shallow layers of the network, we design DSRFB, a DEM super-resolution architecture that generates high-resolution DEM with a super-resolution factor of 8X with minimal input. Our experiments on Pyrenees and Tyrol mountain range datasets show that DSRFB can perform near to the state-of-the-art without using information from any additional modalities like aerial images.

Further, by understanding the limitations of DSRFB, which primarily occur in case of highly degraded low-resolution input. In such cases, the major structures are entirely lost and the reconstruction becomes challenging. In such cases, to avail the elevation cues from alternate sources of information becomes necessary. To utilize such information from other modalities, we inherit the attention mechanism from natural language processing (NLP) domain. We integrate the attention mechanism into the feedback network to present Attentional Feedback Module (AFM). Our proposed network, Attentional Feedback

Network (AFN) with AFM as a backbone, outperforms the state-of-the-art methods with the best margin of 7.2%.

We also emphasize on the reconstruction of the structures across patch boundaries. While generating HRDEM by splitting large DEM tiles into patches, we propose to use overlapped tiles and generate an aggregated response to dilute the artefacts due to structural discontinuities.

To summarize, in this research, we propose two methods DSRFB and AFN to generate a high-resolution DEM from existing low-resolution DEM. While DSRFB achieves near to the state-of-the-art performance, coupling DSRFB with attentional mechanism (i.e., AFN) outperforms state-of-the-art methods.

Contents

Chapter Page

1 Introduction 1
1.1 Motivation: Why super-resolution of DEM? 1
1.2 Challenges 3
1.3 Approach towards the solution 5
1.4 Contributions 7
1.5 Thesis Delineation 8

2 Background 10
2.1 Preliminaries 10
2.2 Tools and Libraries 12
2.3 Available DEM Datasets 12
2.4 Performance Metrics 13
2.5 Related Work 14
2.5.1 Super-resolution of DEM aka Terrain Ampli cation 15
2.5.2 Super-resolution of Images 16

3 Feedback Neural Network based Super-resolution of DEM 21
3.1 Introduction 21
3.2 Contributions 21
3.3 Method 22
3.4 Experiments and Results 25
3.4.1 Datasets 25
3.4.2 Implementation Details 25
3.4.3 Comparison with state-of-the-art method 25
3.4.4 Study of Parameter Choices 27
3.5 Discussion 27
3.6 Conclusion 30

4 Attentional Feedback Networks for DEM Super-resolution 32
4.1 Introduction 32
4.2 Contributions 33
4.3 Method 34
4.3.1 Proposed Attentional Feedback Network Architecture 35
4.4 Experimental Setup 38
4.4.1 Datasets 38

4.4.2 Implementation Details 38
4.5 Results and Discussions 39
4.5.1 Ablation Studies 40
4.6 Intermediate Experiments 44
4.7 Extensions with Loss Functions 46
4.7.1 Logarithmic L1 Loss 46
4.7.2 Loss Function with Gradient Constraint 47
4.8 Conclusion 47

5 Conclusions and Future Work 48

Bibliography 51

List of Figures

Figure Page

1.1 Applications of DEM in GIS: Figure (a) shows landslide hazards Image Source: Flicker. Figure (b) shows terrain application in ood modelling. Figure (c) shows virtual maps of mountain ranges near Kedarnath, India, generated by Google Earth 2
1.2 Other Major Applications of Terrains: Figure (a) shows Bezmiechowa air eld 3D Digital Surface Model captured with Pteryx UAV. Image Source: Kbosak Wiki. Figure (b) shows 3D model of parks in Athence genereated via Google Earth. Source: Google Earth . . 3
1.3 Challenges: Various terrain structures. 4
1.4 Challenges: Terrain showing self-similarity property 4
1.5 Hillshades view de ning the features present in low and high-resolution DEMs. As it can be seen most of the linear structure are vanished out in 15m resolution and recovery of which becomes a challenging task. 5
1.6 Challenges: While some techniques can use information from additional sources like aerial images, these methods can perform poor as aerial images of the terrains covered with snow or dense vegetation could not provide accurate elevation cues. 6
1.7 Low and high-resolution DEMs and corresponding aerial image 8
1.8 An overview of the contributions of this thesis. The modules shown in the green boxes viz. DSRFB and AFN are the key contributions of this thesis. 9

2.1 Various terrain structures. Figure (a) shows aerial view of the terrain. Figure (b) shows corresponding DEM. Figures (c) and (d) show the hillshade maps of the same terrain where the light direction has been chosen to 45^0 and 315^0 respectively. 11
2.2 Indian Project Dataset: The surface level features present in image (a) are barely visible in image (b) and almost absent in image (c). 13
2.3 Architecture of FCN Terrains [11] 17
2.4 SRFBN's recursive setup. Image source: [39] 19
2.5 Features learnt with feedforward and feedback neural networks. Image source: [39]. In the feedforward setup, the network forms a hierarchy of information through the layers where the earlier layers extracting edges and deeper layers suppressing the smoothed structures. On the other hand, with more negative values at the early iteration itself, feedback network tries to suppress the smoothed features at earlier iterations itself and later allows to generate more accurate features. 20

3.1 Outline of the proposed DSRFB network that transform input low resolution DEM to high resolution DEM. The hashed red path denotes the ow of feedback information. Blue circles depict the recovered high delity features. 22

3.2 Network unrolled in time steps. Each data sample goes through T iterations. The output of feedback module at current iteration F^t_{out} will be passed as feedback for the next iteration. At the end of T iterations, system generates T super-resolved versions of input DEM 23
3.3 Layers and skip connections in Feedback Module. 24
3.4 Qualitative results (best visualized in colour and zoomed-in). Blue circles show areas with interesting details. Ellipse in black shows the undesired articulation at the boundary areas for other methods but is well reconstructed in proposed DSRFO. 27
3.5 Qualitative Results: Reconstruction of a patch form Durrenstein region 28
3.6 Qualitative Results: Reconstruction of a patch form Bassiero region 28
3.7 Prediction errors displayed at patches. Red and blue intensities denote prediction below (-ve error) and above (+ve error) the actual terrain, respectively. 29
3.8 Limitations: By focusing on the areas encircled in red, we can see that the features are present in HRDEM are entirely lost in LRDEM. DSRFB nds it dif cult to recover such features which do not have any elevation cues in LRDEM. 30
3.9 Limitations: The features present in this dataset are surface level. We can see that the features are almost barely present in LRDEM and DSRFB tried to enhance them causing false elevation patterns. 31

4.1 Views of the terrain at different resolutions and corresponding aerial image. 33
4.2 Proposed Attentional Feedback Network Architecture 34
4.3 Unrolled Model Structure. The hashed red lines denote the ow of feedback information across time. The green arrows denote the ow of RGB features whereas the blue arrows denote the ow of DEM features. 35
4.4 Feature Extraction Module. We use only two convolution layers for extracting DEM features and rst two layers from VGG-16 network pre-trained on Imagenet dataset . . 36
4.5 Attentional Feedback Module and its components. In Figure (a), the components inside the light red box are part of feedback loop and the hashed red-line denotes the ow of feedback information. 37
4.6 Qualitative Results for a patch from Bassiero region 41
4.7 Qualitative Results for a patch from Forcanada region 41
4.8 Qualitative Results for a patch from Durrenstein region 42
4.9 Qualitative Results for a patch from Monte Magro region 42
4.10 Error Plot: The red and the blue colors denote the elevation values predicted by the respective networks below ground truth elevation (hence +ve numerical error value) and above the ground truth (-ve numerical error value) respectively. 43

List of Tables

Table Page

3.1 PSNR and RMSE for the test regions. While comparing with the methods that use only LRDEM, our networks (DSRFB and DSRFO) outperforms FCND (the variant of FCN with only LRDEM as input) 26
3.2 PSNR and RMSE comparisons between FCN (which uses aerial images) and our proposed DSRFB and DSRFO (which use only LRDEM) 26
3.3 Improvement in performance of DSRFB across iterations T with G set to 12. For each iteration, each column shows PSNR/RMSE for the selected test regions. As the iterations increase, the performance becomes better as can be seen from the number increasing, down the line for each column under PSNR and decreasing for each column under RMSE. 29
3.4 Improvement in performance of DSRFB across iterations T with G set to 16. Compared to Table 3.3, the improvement in the performance of respective iteration shows the effect of increase in the number of residual units 30

4.1 Comparison: RMSE values (in meters) 39
4.2 Comparison: PSNR values (in dB) 39
4.3 Mean SSIM scores for each region computed between ground truth and prediction by various methods 40
4.4 Mean Chamfer Distance (in meters) for each region computed between ground truth and prediction by various methods 40
4.5 Ablation Studies 43
4.6 Percentage drops in PSNR for AFND and FCND with respect to the performance of DSRFB. The drop in PSNR for AFND is 0.5%. 44
4.7 Percentage increase in RMSE for AFND and FCND with respect to the performance of DSRFB. The increase in RMSE for AFND is around 3% which considerably lower than FCND. 44
4.8 Test region wise PSNR and RMSE values for variants of feature fusion techniques . . . 46
4.9 Test region wise PSNR and RMSE values for variants of variants of loss function . . . 47

Chapter 1

Introduction

1.1 Motivation: Why super-resolution of DEM?

Real-world terrain is a complex structure consisting of bare land, high range mountains, river paths, arches, canyons and many more. The terrains and their surface geology are digitally represented using Digital Elevation Models (DEM). The terrain data coupled with Geographical Information Systems[1] (GIS) extract topological information for various applications like modelling water ow or mass movements, analysis of the dynamic behaviour of the earth surface, performing disaster mitigation planning such as ood modelling, landslides, etc (Figure 1.1)[2]. Real-time simulations of terrains are used for fast adaptation and route planning of aerial vehicles such as drones, air-crafts and helicopters to name a few. Realistic terrain rendering also nds its application in ranging simulations, entertainment, gaming etc (Figure 1.2). As the visual detail and depth in many of these applications mentioned above, demand terrain information of high-resolution and delity, capturing or generating such information as accurately as possible is the need of the hour.

As shown in Figure 1.3 the diversity and combinations of the complex topological structures make capture/synthesis and analysis of the terrain a challenging task while considering realism. For instance, computer games with high realistic graphic environments include terrain features for users to experience better realism and allow for detailed exploration. The synthetic or ampli ed terrain can be used as a background for science fantasy lms also, as the synthetic terrain does not exist and ampli ed terrain may be dif cult for the lming process.

Out of many other available formats to represent the elevation of terrain like LiDAR scan, elevation contours etc, DEMs are favoured by the ease of their processing and interpretation as a raster. However, DEMs captured with satellite or other airborne vehicles are still of relatively low resolution (30 m/pixel) [7, 1], etc. On the other hand, the resolution of satellite imagery is relatively higher (50 cm/pixel). Of course, there are a few geographical locations where the DEMs have been captured in high-resolution

[1] We would like to advise the readers new to GIS to refer section 2.1 for getting acquainted with the terminologies used in this thesis.

[2] Same version of this thesis with high resolution images and supplementary material is available in this path.

(a) Landslides Modelling

(b) Flood Model Simulation (c) Terrain Analysis.

Figure 1.1: Applications of DEM in GIS: Figure (a) shows landslide hazards Image Source: Flicker. Figure (b) shows terrain application in ood modelling. Figure (c) shows virtual maps of mountain ranges near Kedarnath, India, generated by Google Earth

(2m/pixel) using customized captured [2, 6, 5] or using airborne LiDAR technology. Though LiDAR technology can provide very high-resolution DEM, they are still currently limited in use due to their high setup cost and processing requirements. Currently, LiDAR is being used only as on demand method for critical projects. Most often these captures can not be made public due to proprietary rights.

This problem can also be solved by transforming the available low-resolution DEMs (LRDEM) to super-resolved DEMs. Hence, we explore scalable solutions for generating high-resolution DEMs (HRDEMs) using their low-resolution counterparts (30m/pixel). These techniques are often discussed under the umbrella term Terrain Modelling and can be broadly classi ed as Terrain Ampli cation and Terrain Synthesis. Terrain ampli cation enhances the high-frequency 3D texture details of the scanned low-resolution terrain captured from the real world, thereby making it as close as possible to actual ground truth terrain. On the other hand, terrain synthesis deals with generation of terrain with speci c user controls giving a near-realistic appearance.

(a) Modelling of air eld for simulation (b) Modelling of real world terrain

Figure 1.2: Other Major Applications of Terrains: Figure (a) shows Bezmiechowa air eld 3D Digital Surface Model captured with Pteryx UAV. Image Source: Kbosak Wiki. Figure (b) shows 3D model of parks in Athence genereated via Google Earth. Source: Google Earth

1.2 Challenges

To summarise, we enlist some critical challenges in super-resolution of DEM as follows:

1. Diverse nature of the terrains makes the learning of these features a complex task. Figure 1.3 shows a few of such variations in terrains surfaces. Based on the geological phenomenon, terrains undergo various degradations. Some of which results in erosion, smoothing, cracking etc.

2. Talking in terms of feature scales, the features in DEM are often present at a smaller scale and vanish out easily even with simple degradation. While creating a dataset from existing high-resolution DEM by downsampling, features get vanished completely and recovering the high-resolution counterpart becomes more challenging. Figure 1.5 shows hillshade models for one such case where some of the smaller scale features are completely lost and making it a challenging case to recover the details.

3. Methods like [11] have tried to use elevation cues from high-resolution aerial images to aid the super-resolution process. However, as shown in Figure 1.6, sometimes aerial images can not provide true elevation cues, typically during the heavy snow-fall and areas with dense vegetation.

4. Terrains visualized from far distance exhibit the self-similarity pattern (Figure 1.4). Self-similarity patterns are patterns repeating themselves at different scales. Though these self-similarities seem easier to introduce in artistic terrains, recovering them from low-resolution terrain as close as to ground truth is dif cult. This is because the problem in this case becomes ill posed as many of the missing details have to be hallucinated unless any additional information cues from other modalities are given.

5. Terrain structures also exist in the form of arches and caves. DEMs captured from airborne devices or satellite are not capable of capturing such structures and thus super-resolution of these structures falls out of scope for this thesis.

(a) Highly diverse Terrain

(b) Planar terrain with slight variations

(c) Terrain with surface level variations

(d) Non-planar terrain with slight surface variations

Figure 1.3: Challenges: Various terrain structures.

(a) USGS[7]

(b) Moredo Mountain[2]

Figure 1.4: Challenges: Terrain showing self-similarity property

(a) Resolution 15m (b) Resolution 2m

Figure 1.5: Hillshades view de ning the features present in low and high-resolution DEMs. As it can be seen most of the linear structure are vanished out in 15m resolution and recovery of which becomes a challenging task.

1.3 Approach towards the solution

Some of the earliest methods for terrain ampli cation employed dictionary of exemplars to synthesize high-resolution terrains [26, 36] while some other efforts in literature used erosion simulations to mimic the terrain degradation effects [41, 58]. Owing to recent advancements in deep learning literature for super-resolution of real world RGB images like in [17, 33, 56, 32, 49, 39], some recent efforts have adopted these ideas for DEM super-resolution. One of the very few efforts on availing the power of deep-learning based solution to the super-resolution of DEM task is based on Fully Convolutional Networks (FCN) proposed in [11] referred as FCN hereafter. FCN [11] by its performance over generation of realistic terrain demonstrated the essence of deep learning solutions to DEM super-resolution. However, FCN is based on the use of additional modalities like aerial (RGB) images (Figure 1.7(c)) that are geo-registered with the underlying low-resolution DEMs. However, getting such pair aerial image tagged with DEM is an arduous task. Further, it can be shown that such methods can be misleading in cases of dense vegetation and heavy snowfall as the underlying variations of the terrain are hidden.

Inspired by these challenges, we attempt to solve the super-resolution problem of DEM with simplistic setup, i.e., without using any additional modalities. Our solution is based on the features found in the natural terrains. The features available in natural terrains are small scale structures mostly comprising of small edges and contour structures. Interestingly, the initial layers in Convolutional Neural Network (CNN) capture these smaller scale (high frequency) details, which need to be enhanced or strengthened for DEM super-resolution. This motivates us to focus on learning of features captured by the initial

(a) Snow covered terrain surface (b) Dense vegetation covering terrain surface

Figure 1.6: Challenges: While some techniques can use information from additional sources like aerial images, these methods can perform poor as aerial images of the terrains covered with snow or dense vegetation could not provide accurate elevation cues.

layers of the CNN. In CNN, however, the features are propagated in the forward direction and there is no explicit control over the features learnt by the initial layers. While the CNN itself inspired by the functioning of the human brain, CNN with its architectural limitations, lacks to avail additional features like feedback mechanism. Studies like [31, 23] from cognition theory, reveals the feedback connections in the visual cortical visual area, capable of transmitting the response signals from areas responsible for higher order recognition to areas requiring lower-order recognition for the complex recognition task.

Exploiting this idea of feedback, Zamir et al. proposed Feedback Networks [53] to learn complex tasks iteratively. In a feedback network typically implemented with a Recurrent Neural Network as the backbone, the output from an iteration is passed on to next one by using a set of hidden states. This setup comes up with parameter sharing and early prediction as additional bene cial properties. While the trend was for going deeper and using a complex structure with a large number of parameters, [46] introduce residual setup for super-resolution task for natural images. Further, to enable the re ning of features learnt by the initial layer, [39] unitized Feedback networks for super-resolution.

Inspired from the idea of feedback networks [53] and its success in super-resolution task [39], we design a super-resolution architecture for DEM, that iteratively re nes the features learned by initial layers. We call our architecture as DSRFB (DEM super-resolution using Feedback Networks). Our experiments show the effectiveness of DSRFB by its near to the state-of-the-art performance [11]. Further, we also attempt to mitigate the artefacts in the reconstructed Super-resolved DEM due to structure splits at patch boundaries.

The proposed network DSRFB performs comparable to the state-of-the-art and in some regions with dense vegetation or heavy snow, performs better than the state-of-the-art. Nevertheless, empirically we have found that DSRFB can perform poorly in areas where the information is completely lost in low-resolution DEM and use of some additional modality can be helpful here. This lets us study the limitations of FCN [11] (which avails complementary information from RGB images but can mislead

where terrains in the aerial images are covered with natural entities) and DSRFB which does not rely on RGB modality (but has limitations if the low-resolution input is severely degraded). We attempt to utilize these complementary modalities in a more ef ciently and effectively using the concept of selective fusion in feature space. Attention Networks, applied to applications like image captioning [51] allow such selective fusion in deep learning framework. Therefore, we aim to design an integrated attention module with our DSRFB, which we now call Attentional Feedback Module (AFM), that enables learning of selective information fusion from multiple modalities. In this setup, where we have two modalities viz Aerial image and DEM, we use attention mechanism to selectively pick high-frequency details from one modality and discard from the other. Our joint attentional module, i.e., AFM generates attention mask, which serves as a weight factor deciding the contribution of each modality.

We compare the performance of the proposed AFN architecture with state-of-the-art architectures for DEM super-resolution quantitatively and qualitatively. Experimental results show that AFN achieves better performance using a reduced number of parameters. More precisely, the proposed AFN shares the parameters across the feedback loop for incremental fusion in feature space and uses only learnable 7M parameters whereas other SOTA architectures like [11] use an order of 20M parameters.

1.4 Contributions

In this research, we present two methods for generating high-frequency terrains with more focus on terrain ampli cation of LRDEM (Figure 1.7 (a)) with the aim to obtain terrain models with high delity to the ground-truth (Figure 1.7 (b)) by optionally using any additional modality information like aerial images (Figure 1.7 (c)). To generate a high-resolution DEM from a low-resolution input, we provide DSRFB. To avail the information from other sources such as aerial images, we propose AFN. The overall idea of this thesis can be summarized by looking at Figure 1.8, where the modules shown in green de ne our key contributions.

We enlist our contributions from this research effort below:

- We propose feedback based super-resolution network architecture, i.e., DSRFB, for DEMs that performs near to the SOTA[11] even without using any additional modalities.

- We propose the use of prediction by incorporating multiple estimates for the pixels near the boundary areas while reconstructing high-resolution DEM.

- An Attentional feedback module that selectively extracts information from different modalities to generate high-resolution DEM.

(a) Input LRDEM

(b) Ground truth HRDEM

(c) Geo-registered aerial image

Figure 1.7: Low and high-resolution DEMs and corresponding aerial image

1.5 Thesis Delineation

The current section has covered the introduction to the research problem, the application areas and the key challenges in solving it. The remainder of this thesis is organized as follows:

Chapter 2 provides an understanding of relevant research that has been carried out in this domain. We focus on work that has been carried our for realistic terrain generation, terrain super-resolution with delity to ground truth and super-resolution methods for images in general. At the end of this section, we introduce the recently proposed feedback neural networks for super-resolution of natural images.

In chapter 3 we introduce our rst contribution, i.e., DSRFB, a feedback based neural network for super-resolution of DEM. We discuss the architectural details and training method. We compare the relative performance of DSRFB with FCN [11]. Finally, we discuss the limitations of FCN and DSRFB altogether and possible direction for overcoming those limitations.

We start chapter 4 by discussing the pros and cons of FCN and DSRFB, and introduce the need of selective feature extraction. We introduce our second contribution in the form of AFN and

Figure 1.8: An overview of the contributions of this thesis. The modules shown in the green boxes viz. DSRFB and AFN are the key contributions of this thesis.

discuss its performance over SOTA methods. Towards the end of this chapter, we highlight some limitations of AFN and conclude with possible directions of working around.

We conclude the research contribution in chapter 5. We also enlist a set of directions to explore with DSRFB and AFN.

Chapter 2

Background

2.1 Preliminaries

Before tracing back the research works related to DEM super-resolution, we brie y introduce key terms that will be used often down the line of this thesis, a few variants to elevation interpretation, currently available elevation datasets, etc.

1. A Geographic Information System (GIS)deals with methods that provide the ability to capture, store and analyze spatial and geographic data. GISs have enormous applications in elds related to mapping (cartography), analysis and prediction of landslide hazards, urban planning, agricultural applications, land use and land cover changes, disaster management and mitigation, ood management and the list goes on. Most of these applications can be built using a raster data structure, the Digital Elevation Model (DEM) .

2. A DEM is a raster, which stores the elevation values of the terrain across a spatial region (Figure 2.1 (b)). On an abstract level, a DEM can be considered as a depth image of a large terrain with depth captured from axis perpendicular to the terrain surface.

3. Arc-meters and arc-secondsThese are the metrics to represent the latitude and longitude based on distance traversed on the earth's surface with respect to time. At sea level, one minute of an arc along the equator or a meridian (indeed, any great circle) equals exactly one geographical mile along the earth's equator which approximates to 1,852 metres. An arc second $\frac{1}{60^{th}}$ of this approximates to 30 meters [4].

4. Elevation: The elevation of any point in a DEM represents the height of that point with some reference elevation, most often, the sea level. The most common terms related to elevation are spatial resolution and vertical accuracy. A spatial resolution represents the surface area represented by the point. For instance, a spatial resolution of 30m/pixel (or 1 arc-second) denotes each pixel in the DEM represents an area of 30x30 square meters of the underlying surface. Vertical accuracy represents the average squared error in the elevation w.r.to to actual elevation value.

(a) Aerial view of the Terrain

(b) DEM

(c) Hillshade: View 1

(d) Hillshade: View 2

Figure 2.1: Various terrain structures. Figure (a) shows aerial view of the terrain. Figure (b) shows corresponding DEM. Figures (c) and (d) show the hillshade maps of the same terrain where the light direction has been chosen to be 45^0 and 315^0 respectively.

5. Ortho Images: Satellite imagery and aerial photographs, in the raw form, generally suffer from distortions due to natural topographical variations in the surface of the terrain, the angle between the terrain surface and the satellite sensor. The inherent distortion in the image has a direct impact over the diversity of the underlying terrain. Before these images can be used for mapping or analysis in a GIS system, these images must be pre-processed to remove the distortion. This removal of distortion aka `Orthorecti cation', recti es the images so that it can be used for various mapping and analysis in GIS by accurately by overlaying with layered data structures. Figure 2.1 (a) shows a recti ed ortho image.

6. Hillshade: A Hillshade is a gray-scale 3D representation of the surface where Sun's relative position taken into account for shading the image. Unlike DEM, the colour shades do not attribute to elevation; rather they represent the terrain variations visible from a virtual position of Sun. Figures 2.1 (c) and (d) show the hillshade maps for DEM shown in (b).

2.2 Tools and Libraries

The plot and diagrams in this work have been generated using MeshLab and QGIS. MeshLab is a standard 3D mesh visualization and processing software. QGIS is an open-source, cross-platform desktop application for visualization and processing of geospatial data in GIS. It supports multiple formats of raster images and is heavily used for georeferencing images. The hillshade plots in this thesis have been generated using QGIS.

2.3 Available DEM Datasets

Available DEM Datasets: Currently, we have publicly available DEM datasets, that have elevation models for the entire globe. We discuss a few major datasets here.

1. Space Shuttle Radar Topography Mission (SRTM): As part of the mission, the SRTM shuttle orbited earth 16 times in 11 days and covered over 80% of the earth's surface. It has used synthetic aperture radar and interferometry for capturing elevation. The dataset is available at two resolutions, viz 30m (1 arc-second) and 90m (3 arc-second). SRTM data can be accessed from the USGS (United States Geological Survey) Earth Explorer website[3]. SRTM has an average vertical accuracy of less than 16m.

2. ASTER Global Digital Elevation Model: ASTER, which stands for Advanced Spaceborne Thermal Emission and Re ection Radiometer was part of NASA and Japan's combined project. ASTER in its original version was available with a global spatial resolution of 90m and with a resolution of 30m for the United States. However, with artefacts found in cloudy areas, a more accurate version, i.e., ASTER GDEM version 2, was released in October 2011. ASTER dataset also is available on the USGS Earth Explorer[7].

3. Global ALOS 3D World by JAXA: ALOS 3D World[1] is also an 1 arc-second spatial resolution dataset provided by Japan Aerospace Exploration Agency (JAXA) with their Advanced Land Observing Satellite (ALOS) satellite.

4. OpenDEM[5]: Open Digital Elevation Model is the collection of one of the high-resolution elevation model datasets. It offers DEMs with as high resolution as 1m for some regions in Europe and various resolutions in between for rest parts of the world.

5. ICCG [2] and SBG[6]: These are some of the high-resolution datasets that cover mountain ranges from Pyrenees and Tyrol. These are particularly used in this thesis because of their high resolution. These datasets offer DEM with a resolution of 2m. By downsampling these high-resolution DEMs, a pair of low and high-resolution DEMs can be created for using in supervised machine learning tasks.

6. We also have would like to brie y mention about an Indian project dataset. This dataset comprises of patches of DEM from a small region in India. This is a collection of data from three dataset providers, namely SRTM (30m), CartoSat (10m) and a customized scan of 5m resolution. From Figure 2.2, we can see that the surface level features present in 5m resolution are barely visible at 10m resolution and are almost absent in 30m resolution. Thus this becomes a challenging case for super-resolution. Due to smaller size of this dataset, we do not train our network with this data but will use it for testing generalization of the network.

(a) Resolution: 5m (b) Resolution: 10m (c) Resolution: 30m

Figure 2.2: Indian Project Dataset: The surface level features present in image (a) are barely visible in image (b) and almost absent in image (c).

As can be seen from the above datasets, the highest resolution of DEM that is available for the entire world is 30m. However, the aerial or satellite imagery is available at a spatial resolution of as high as 50cm. For practical analysis of terrains and higher accuracies of the systems that use them, there is a considerable scope to use high-resolution DEM. Due to hardware limitations and the cost and time involved in capturing of such scans, getting a high-resolution DEM via improving the hardware based techniques becomes limited. We thus propose to apply super-resolution techniques from computer vision to generate the high-resolution DEMs from existing low-resolution DEMs.

2.4 Performance Metrics

Similar to natural images, the performance of super-resolution methods for DEM can be done with standard methods like Root Mean Squared Error (RMSE), PSNR (Peak Signal to Noise Ratio) metrics, etc. While RMSE helps to understand the cumulative squared error between the prediction and ground truth, PSNR helps to gain the measure of peak error. In a way, PSNR and RMSE are complementary measures to compare the performance of SR methods. The computation of RMSE for DEM is similar to the case of natural images, where we nd the squared root of the sum of squared errors.

$$RMSE = \sqrt{\frac{1}{S} \sum jGT \quad SRj^2} \tag{2.1}$$

where GT represents the ground truth elevation map, SR represents the super-resolved prediction from low-resolution input and S denotes the number of pixels in DEM.

However, the computation of PSNR becomes slightly trickier. In the case of images, we find the PSNR by,

$$PSNR = 20 \log_{10}(255.0/RMSE) \tag{2.2}$$

where the factor 255.0 indicates the maximum intensity value a pixel can take while the RMSE has been computed from Eq. 2.1. However, in the case of DEM, where we deal with elevation, this factor is variable. To set up the standard, authors have been using the maximum value present the ground truth (denoted as P) DEM while computing PSNR with various DEM degradation or reconstruction methods. Thus we use the following formula for computing PSNR for DEMs.

$$PSNR = 20 \log_{10}(P/RMSE) \tag{2.3}$$

Further, we also use metrics like SSIM and chamfer distance. SSIM is a perception-based model that conveys image degradation as perceived change in the structural contents. Treating DEMs as single-channel images, we can calculate SSIM scores for DEM with reference to a ground truth DEM. Chamfer distance has been used a comparison metric in ShapeNet's reconstruction challenge [3]. We have included chamfer distance in the evaluation metrics as DEM represents 3D terrain surface and can be easily converted to point clouds.

2.5 Related Work

Generating high-resolution DEM from a low-resolution terrain can be thought of as generating the high-frequency details like texture patterns, sharp edges, small curves which often are lost or absent in the low-resolution DEM. Image super-resolution is one of the core problems in computer vision. It is a complex task as it involves multiple plausible solutions. This is due to many high-resolution images, which after degradation by various methods, individually could result in the same low-resolution image. Considering super-resolution in this setup is highly dependent on our prior information about the degradation pattern.

Another way of seeing the super-resolution problem could be as an image translation task where source (input) image is a low-resolution image whereas the target (output) image is the high-resolution one.

With the recent success of deep learning based solution to super-resolution, super-resolution of natural RGB images have reached new heights. However, there are very few works [11, 13, 15, 40, 52] that have tried these techniques on DEMs. This could be attributed to the fact that the information present in the natural images are far different than what a DEM represents. The size of features, objects, the diversity of texture are profoundly different in these two modalities. Also, the datasets available for training such systems like DIV2K [8] have been standardized for super-resolution of natural images. However, such datasets are not available for DEMs. To understand how these challenges were tackled in this cross-domain task, we would like to highlight some of the signi cant works in respective domains.

This section presents a focused overview of terrain modelling, super-resolution methods for images in general and nally the feedback neural networks for image super-resolution.

2.5.1 Super-resolution of DEM aka Terrain Ampli cation

Increasing the details of terrain or Terrain Ampli cation has been an exciting direction of research. The community has followed two paradigms to enhance the terrain with high-frequency details. The rst direction deals with improving the texture of the terrain to give a realistic appearance to the terrain. A low-resolution terrain can be processed by playing with its texture, to introduce the erosion effects, to introduce biological effects such as vegetation. As you might have guessed, these work have applications in gaming industries, animation industries, AR-VR platforms, smart city planning etc. The second paradigm focuses on increasing the details of a real terrain by making it as close as possible to actual ground truth terrain. This is, in other words, super-resolution of terrain and is the main focus of this thesis.

Primarily, terrain ampli cation methods can be categorized into three 3 categories: Procedural Modelling, Use of Simulation, and Examples based learning.

One of the early works on terrain modelling by using Procedural modelling involves the use of fractal noises. Fractal patterns are self-repeating patterns. Natural terrains have this beautiful property of existing in self-repeating patterns at different scales. Perlin et al. [42] proposed the use of generating such fractal patterns and adding it to the low-resolution terrains to convert them into realistic terrains. By using combinations of octaves of such fractal noises and thereby creating various scales of noise and smoothness, [41] added new variations in the fractal dimensions.

Analogous to mountains, rivers can also be modelled with procedural modelling and incorporated into the landscape. Starting from designing a river network and using a classi cation of water bodies like springs, junctions and deltas etc., Genevaux et al. [21] used a modelling tree approach to produce broader terrains. Working directly with these methods will not let the user to interfere with the results, and hence results are often unrealistic sometimes. One of such works to enable the user to interact with and change the parameters was presented by [43]. They use painting and brushing on gray-scale images as the fractal's basis functions for editing. Availing the power rendering powers of GPUs, their editor gives real-time visual feedback, thereby enabling the user to enjoy the terrain generation interactively.

Gain et al. [20] used primitive features in the form of silhouette and shadows whereas Hnaidi et al. [29] used parameterized curves to de ne the land-form features such as ridgelines, riverbeds, cliffs etc. [22] presented an hierarchical combination of the primitives like riverbed, cliffs, hills etc., with fusing operations like blending by creating a tree of these object and operations. Terrains generated procedural techniques [20, 29, 22], still lack the effect of natural phenomenon like erosion in their appearances. Most often, a terrain generated by procedural methods are combined with simulation operations.

Phenomenon like erosion of terrain are most often caused by physical factors like:

Temperature when terrains where frequent contractions and expansion of terrains results in breaking of the terrain

Hydrological factors like rivers passing through the materials resulting in smoothed valleys

Gradual abrasion as a result of overexposure to winds

Musgave et al. [41] presented hydraulic and thermal erosion in their work. In general, these simulations can be seen as moving the rocks and sediments from one location to another as an effect of one or more physical activities that dynamically results into a terrain. Such effects can be combined with ecosystems such as vegetation modelling. However, representation of arches and caves need more than mere height field because it can represent only topmost surface in a terrain. [12] introduced layered representation for such structures. These structural representations have also enabled stacking multiple layers for effects of various physical and biological phenomenon. One such integration has been represented by [14] where they fused the interaction between the growing vegetation and terrain erosion by representing them into different layers.

The third way to generate new terrains is to avail the terrain information from existing terrain datasets. Zhou et al. [58] combined an example terrain and a user sketch defined user desired terrain features to turn the sample terrain into terrain with sketch defined features. Work by [36] and [26] use patch-based terrain synthesis by using a dictionary of exemplars. Kwatra et al. [36] use graph cuts to match and fuse the potential patches. Authors in [26] make use of use of multi-resolution dictionaries to include various levels of details. With recently successful deep learning based Generative Adversarial Networks (GANs), [27] used conditional GANs to translate a sample terrain with interactive user sketch. Using fully convolutional networks as a backbone, Argudo et al. [1] proposed the use of elevation cues from high-resolution aerial images. Their architecture, referred as FCN for simplicity, is shown in Figure 2.3. Seeing terrains amplification as a style transfer in images, [57] designed the generative network to transfer textures from sample terrain to input low-resolution terrain.

2.5.2 Super-resolution of Images

Super-resolution of natural images has been a well-attempted problem. When it comes to finding the missing details, interpolation based on the neighbourhood information becomes the default choice. Methods such as linear, bilinear, bicubic or filtering methods like Lanczos, comprises of simple mathematical solutions and often reduce down the super-resolution problem to solving an equation. Also, since these methods try to average out the details based on neighbourhood, these methods end up smoothing out the surfaces. Such methods often fail on recovering the missing edges. A simple modification to this, as in Edge Directed interpolation [9] particularly focuses on edges and gives the images some sharp structures. Using a patch based learning of mapping between LR and HR patches, where a large scene or an image is generated by patch-wise replacement of corresponding high resolution patches from the learned mapping has been proposed by [19, 18]. By allowing small variations to the patches from the dictionaries, Huang

Figure 2.3: Architecture of FCN Terrains [11]

et al. [30] have included more flexible generation and even reduced the patch redundancies across various scales.

Observing that the overlapped patch-based reconstruction approaches ignore a critical aspect of patch consistency, [25] proposed to use a convolution based processing of image in order to tackle the patch inconsistency. Initially, an LR image is decomposed into a sparse feature map and an HR image into corresponding high-resolution feature maps. Followed by, a mapping is learned from sparse coded LR features to HR ones. Finally, a set of filters are learned that operate on LR image as convolutional operators to generate a new HR image. By working directly on the whole image, [26] removes the task of dividing the image into overlapped patches. This work also motivates to use convolution operations working on spatially varying data.

Another important direction in preserving edges while avoiding the edge artefacts was explored by [48]. They have extended the Directed Super-resolution by combining it with Gradient Profile Prior [45]. Just a variant of learning the mapping from LR images to HR images by using LR-HR pairs, Kernel Ridge Regressions (KRR) was used by [34]. However, on observing blurring and ringing artifacts (which are often found due to insufficient sampling rates for major edges as sharp changes) while using simple KRR, they include a prior model of a generic image class that considers the discontinuity property of images. These works that we have seen till now are based on using handcrafted features or algorithms based on semantic rules. However, with the boost in utilising deep learning techniques in computer vision, super-resolution using deep learning have raised the benchmark considerably higher. One of the very first work in this approach was by Dong et al. [16]. They applied a bicubic interpolation on LR images to make it equal in dimension to HR and applied a three layered fully convolutional neural network to push the image the interpolated image towards the ground truth higher-resolution image.

Using learned iterative shrinkage and thresholding algorithm (LISTA), [74], Wang et al. [50] designed a neural network such that each layer strictly corresponds to a step in sparse coding based super-resolution. In the alternate direction where the network itself was allowed to learn the upsampling filters in order to increase the accuracy and times was experimented by [17].

After the introduction of ResNet [23] where the use of a residual block to overcome the long-range pixel dependencies inside very deep neural networks, Kim et al. proposed DRCN [38], SRResNet [37] uses 16 residual blocks in its architecture. With their residual blocks similar to the ResNet [23] architecture except the activation function used as PReLU instead of ReLU. However, the major proposal from authors of [37] was to use SRResNet in a generative adversarial set up (termed as SRGAN) where, it can generate photo-realistic images. In SRGAN, authors have used perceptual losses proposed by [32]. They have also proposed the use of patch discriminator. Following this generative setup, ESRGAN [49] have shown improved performance by using Residual of Residual (RoR) [55] networks in their generator architecture. This Residual of Residual networks is a neural network architecture with multiple levels of skip connections. They modify the basic ResNet [26] architecture by introducing various skip connections across the residual blocks to form `block of residual blocks' and this hierarchy can be increased to any level.

Figure 2.4: SRFBN's recursive setup. Image source: [39]

More speci c to the super-resolution task, Zhang et al. proposed the use of global and local skip connections in their architecture named Residual Dense Network (RDN) [56]. They have proposed Residual Dense Block (RDB) and use dense skip connections to extract abundant local features that are captured by earlier layers. RDB allows direct skip connections from all the preceding RDB to all the layers of the current RDB, forming a contiguous memory mechanism. This setup makes low-level features available towards the deeper layers as well (and not limited only for immediately following layer, as it is in case of simple ResNet). At the end of obtaining the dense local features from the last RDB, features bypassing all the RDBs via a global skip connection, are fused with the output features of last RDB to learn the global hierarchical features as well.

Though RDN was able to utilize the low level features effectively, the ow of information was only from initial layers to deeper layers. These low-level features are reused again and again limiting the reconstruction ability of lower features in the super-resolution task of the network.

SRFBN (Super-Resolution Feedback Network) [39] was proposed to tackle this problem. SRFBN used a feedback mechanism adapting from Feedback Networks [53] in their architecture. SRFBN has also been implemented in a recursive setup, where the output features from the current step are stored and used for the successive steps. As shown in Figure 2.4, the output feature maps from deeper layers are again introduced to the initial layers and thus enabling the relearning of the initial layers with assistance from the output of previous layers. Authors for SRFBN [39] have also shown a distinction between the features learnt from feedforward and feedback layers (refer Figure 2.5)

Using a recurrent structure and thereby reusing the parameters has been one of the major techniques in while working with sequential data in deep learning. Recurrent structures also help to design a feedback

Figure 2.5: Features learnt with feedforward and feedback neural networks. Image source [39]. In the feedforward setup, the network forms a hierarchy of information through the layers where the earlier layers extracting edges and deeper layers suppressing the smoothed structures. On the other hand, with more negative values at the early iteration itself, feedback network tries to suppress the smoothed features at earlier iterations itself and later allows to generate more accurate features.

mechanism easily as recurrent structure can save states of a layer which helps in implementing the feedback component. FCN [11] extracts complementary information from aerial images. However, in their feed-forward setup, there is no control over features learned by initial layers of the network. We thus propose a feedback based super-resolution architecture named DSRFB for generating high-frequency features in DEM. Further, nding the limitations of DSRFB in cases of utterly degraded terrains, where the features are entirely absent low-resolution DEM, we seek to utilize information cues from other modalities. Though FCN [11] makes use of such additional cues from aerial images, there is little control on the extraction of features from aerial images. This motivates us to explore solutions that enable selective extraction of features from aerial images while focusing more on learning of initial layers of the network. For image captioning task, [51] successfully demonstrated the application of attention from one modality to others. Generalizing such attention mechanism, we propose Attentional Feedback Network for adaptive utilization of features selected respectively from aerial images and DEM. Integrating the attention mechanism with feedback network enables the proposed network to learn more re ned lower level features.

Chapter 3

Feedback Neural Network based Super-resolution of DEM

3.1 Introduction

Existing methods for generating HRDEMs, also known as terrain ampli cation, have primarily followed two paradigms: one aimed at reconstructing DEMs with higher delity to the actual terrain (e.g., dictionary-based reconstruction [26]), and others focused on adding plausible details and enhancing visual appearance, apathetic to deviation from real terrains, e.g., [27, 57].

While generating the terrain with delity to ground truth while trying to generate realistic terrain, [11] proposed an architecture based on fully convolutional networks (FCN) architecture that uses LRDEM along with registered ortho-photo (RGB) to generate an HRDEM. However, generating registered ortho-photo and LRDEM pairs is an arduous task. Additionally, aerial image modality could be misleading due to dense vegetation, heavy snow and timely changes in the landscape appearance. Similarly, by using timely changes in the terrain, authors of deepSUM [40] proposed a multi-input to single-output (MISO) system to generate high-resolution terrains. In the relevant eld of natural image super-resolution, deep learning solutions have shown exceptional improvement in performance.

We leverage the power of deep neural networks and propose a novel supervised approach to super-resolve the DEMs from LRDEMs. In DEMs, majority of the higher frequency details are available as small discontinuous edges (fractal structures with no global structures), unlike natural images which often contain larger structures with some global context. This motivates us to focus on learning features at smaller scales. Interestingly, the initial layers in Convolutional Neural Networks (CNNs) capture these smaller scale (high frequency) features which need to be enhanced or strengthened for DEM super-resolution. Recently, [39] proposed a feedback module in their super-resolution architecture (called SRFBN) for RGB images to re ne the features learned by initial layers.

3.2 Contributions

Our main objective is to enhance the lower level DEM features (features captured by initial layers in a typical CNN) DEM features, we propose to use a feedback module, similar to [39], in our architecture. As

Figure 3.1: Outline of the proposed DSRFB network that transform input low resolution DEM to high resolution DEM. The hashed red path denotes the ow of feedback information. Blue circles depict the recovered high delity features.

shown in Figure 3.1, the features extracted from LRDEM, F_{in}, are processed along with successive outputs of Feedback block, i.e. F_{out}^{t-1} to generate ner features as F_{out}^{t}. Hereinafter, we call the proposed network as DSRFB (DEM Super-Resolution using Feedback Network). The results from DSRFB using only LRDEM as input shows improved results in comparison to the state-of-the-art approach [11] using LRDEMs combined with aerial imagery.

We have also observed that while reconstructing the high-resolution DEM by splitting large tile in small patches to get the prediction, the system introduces structural discontinuities at the patch boundaries. We propose to use a patch-based overlapping while getting inference from the network and then using an aggregated response for the overlapped patch areas. We refer DSRFB with overlapped inference as DSRFO. Overall, our contributions from this work can be summarized as follows:

- A new Super-resolution network for DEMs that iteratively re nes lower level features in DEM for better delity. Our experiments demonstrate the use of the backward ow of higher-level features helps to re ne the lower-level features in the DEM.

- The proposed method overcomes the feature split across tile boundaries by incorporating multiple estimates in HRDEM reconstruction. This technique can be coupled with any patch-based inference method.

While the two branch architecture by FCN [11] gives state-of-the-art performance, it is limited by the availability of pair of georegistered DEM and aerial image. Similar is the case with DeepSUM [4], where capturing timely varying terrain maps are required. DSRFB, however, gives the state-of-the-art performance by consuming only LRDEM.

3.3 Method

We treat super-resolution as an Image Translation problem. As in [16], we upsample LRDEM with bicubic interpolation as a preprocessing step and denote it as ILR and then add higher frequency details

Figure 3.2: Network unrolled in time steps. Each data sample goes through T iterations. The output of feedback module at current iteration F^t_{out} will be passed as feedback for the next iteration. At the end of T iterations, system generates T super-resolved versions of input DEM

by using the proposed network. Our network was inspired by SRFBN [29] where we wish to learn more accurate low-level features by receiving feedback from higher-level features learnt by deeper layers.

Our main task is to propagate higher level feature information to lower levels. Our proposed network achieves this using a recurrent neural network (RNN). An RNN is a neural network with cyclic connections that make them capable of handling sequential data. An RNN with T hidden states can store network activation up to T states. We model our feedback network such that with each time step, the unrolled network reconstructs a new Super-resolved (SR) image and also store the layer activations in the hidden state which are used for the next time step. Figure 3.2 shows the network unfolded in time steps. Each sub-network consists of three parts: Feature Extraction(FE) block to capture low-level details from ILR image, followed by a Feedback module(FB) where we re ne the low-level features using high-level features and lastly a Reconstruction block(RB) to fuse the high-frequency details with low-resolution DEM features to generate Super-resolved DEM (SRDEM).

The Feature Extraction(FE) module consists of two convolution layers $Conv(m; 3)$ and $Conv(4 * m; 1)$, where m is the base number of lters. DEMs up-sampled with bicubic interpolation(ILR) are used as input for feature extraction. FE module encodes ILR, denoted as F^t_{in} which is then forwarded to Feedback(FB) module once for each time step $t \in \{1; T\}$.

Our Feedback(FB) module consists of a stack of residual units. Each residual unit consists of a $Conv(m; 1)$ followed by a $Conv(m; 3)$ layer. The purpose of $Conv(m; 1)$ layers is to adaptively fuse the number of input channels to base number of lters. We denote the residual units as B_i, where $i \in \{1; G\}$. We use $G = 16$ in our case. At a particular time step t, the output of residual unit B_i is denoted as L^t_i. At time step t, the state of our feedback module is shown in Figure 3.3. Inspired from DARN [47], we use multi-path skip connections. However, instead of using the residual connections from a current residual unit to all its subsequent units, we use them in subsequent alternate units. Thus the information L^t_1 would be passed to $\{B_2; B_4; ...; B_{16}\}$, L^t_2 to $\{B_3; B_5; ...; B_{15}\}$, L^t_3 to $\{B_4; B_6; ...; B_{16}\}$ and so

Figure 3.3: Layers and skip connections in Feedback Module.

on. Eventually, we get two groups of residual paths, as shown by red and green arrows in Figure 3.3. The feedback module receives a concatenation of F^{t-1}_{out} and F^{t}_{in} as input, denoted as $[F^{t-1}_{out}; F^{t}_{in}]$, where F^{t-1}_{out} is the hidden state output features from previous time step ($t-1$) and F^{t}_{in} is the input feature vector at time step t. We use $Conv(m; 1)$ to compress $[F^{t-1}_{out}; F^{t}_{in}]$ along the channels and denote the compression output as L^{t}_{0}. We also use residual connections for L^{t}_{0} to all the alternate layers except the last layer. Lastly, the outputs of $\{B_2; B_4; ...; B_{16}\}$ are compressed by a $Conv(m; 1)$ into F^{t}_{out}, which gets forwarded to reconstruction block for SRDEM generation and also used as feedback information for the next iteration which we concatenate with F^{t+1}_{in}. As there will be no feedback available at iteration $t = 1$, we use F^{1}_{in} as F^{0}_{out}.

So, at the start of each iteration, generalized input will be $[F^{t}_{in}; [F^{t-1}_{out}]$ which will be compressed by the $Conv(m; 1)$.

Inside Reconstruction block(RB) the output of feedback module F^{t}_{out} at iteration t is operated with $Conv(1; 3)$ layer to generate a residual DEM image I^{t}_{res}. We pass the interpolated LRDEM (ILR) to reconstruction block via a global skip connection. The SRDEM at iteration t is then reconstructed as,

$$I^{t}_{SR} = I^{t}_{res} + ILR \tag{3.1}$$

Finally, for T number of hidden states, the model will generate a collection of T SR images $(I^{1}_{SR}; I^{2}_{SR}; ...; I^{T}_{SR})$. For each of the time step, we evaluate $L1$ loss (mean absolute error) between generated SRDEM and ground truth HRDEM.

$$loss = \sum_{t=1}^{T} |I^{t}_{SR} - HRDEM| \tag{3.2}$$

As shown in Eq.(3.2), we accumulate losses over all T steps and at the end of T steps, we back-propagate the loss to optimize the parameters of the network.

3.4 Experiments and Results

3.4.1 Datasets

We use the same dataset used in [11] which is part of publicly available high-resolution DEMs of mountain ranges named Pyrenees [2] and Tyrol [6], respectively. The dataset comprises of DEM of large tiles of mountain ranges. It has the highest resolution of 2m. The dataset has been downsampled to 15m to create the low-resolution counterpart. We explicitly mention that, with a LR-HR pair of resolutions 15m and 2m, the effective scale of super-resolution in our experiments will be 7.5x rather than 8x. This dataset also has satellite images (alternately referred by aerial images within the scope of this thesis) with a spatial resolution of 1m which are geo-registered with the DEM. However, in our setup we do not need these aerial images. For a fair comparison, we use the same distribution as used by [11] by combining tiles from both Pyrenees and Tyrol. These tiles have been sliced into patches of 200x200. Further, 22000 patches have been selected for training and 11000 patches for validation. Four regions named Bassiero, Forcanada, Durrenstein and Monte Magro with their unique nature have been set aside for testing.

3.4.2 Implementation Details

In our experiments, we set the kernel size for all convolution layers is 3x3 except for the compression layers where we use 1x1 kernels. We use PReLU as activation layer after each of the convolutional layer, except in the reconstruction block. We choose SR factor as 8X, the base number of lters, is $n = 64$. The number of steps in feedback module, is $T = 4$ and the depth of the feedback module is $G = 16$. We use the SR image at $t = 4$ as the nal SR image. We have also experimented with an ensemble of SR images for all iterations from $t = 1$ to $t = 4$. However, in case of DEM, it has marginal improvement over the reconstruction. We explore two ways to reconstruct the HRDEM patches from the network. With the simplest setting in DSRFB, we reconstruct each patch independently and place them in the larger tile. However, to effectively recover the features split across the boundary regions, we propose DSRFO to process patches with an overlap. We use an overlap of 25% on each side of the patch and use the aggregated response for the pixels in the overlapped region.

In our proposed network, we have used the entire patch of 200×200 for training. We experimented with smaller patch sizes as well and found that because structural continuities, the larger the patch size the better is the prediction. We use a batch size of 4, an optimizer with a learning rate of 0.0001, and weights initialized with kaiming initialization. We use multi-step learning degradation with gamma set to 0.5. We have implemented our algorithm using PyTorch framework. The network was trained for 100 epochs on NVIDIA 1080Ti GPUs.

3.4.3 Comparison with state-of-the-art method

We use standard root mean squared error(RMSE) and peak signal-to-noise ratio (PSNR) measures to compare the performance of our proposed method with existing SOTA methods, i.e. TFCN [11]. We also

compare the performance with a variant of FCN, i.e., FCND which uses only LRDEM for reconstructing HRDEM and hence a fair comparison can be made in between DSRFB and FCND. Quantitative results are shown in Table 3.1 and 3.2. Even without using RGB channel, DSRFB is able to generate high-resolution DEM with similar accuracy. Our method performs better in regions of Dürnstein and Monte Magro, where the aerial images respectively show them as snow-covered and with dense vegetation, indicating that FCN method may mislead in varying terrain landscapes. However, DSRFB network performs consistently and is more stable across all regions. Further, our variant DSRFO with multiple estimations on tile boundaries performs even better and has the best performance for Dürnstein and Monte Magro regions.

For qualitative performance check, we pick a large patch from Bassiero region. As shown in Figure 3.4, we can see that DSFB is able to recover the missing details as close as to FCN. From the areas highlighted by vertical ellipse, we can see the the artefact introduced by simple patch based reconstruction methods (e.g., FCN, DSRFB) can be reduced by using DSRFO. Figures 3.5 and 3.6 show additional results on patches from Dürnstein and Bassiero regions. Figure 3.7 shows prediction errors for FCN and DSRFO on a patch extracted from Dürnstein region. We can see that DSRFB has lesser prediction error.

Region	PSNR (in dB (↑))				RMSE (in meters (↓))			
	Bicubic	FCND	Ours		Bicubic	FCND	Ours	
			DSRFB	DSRFO			DSRFB	DSRFO
Bassiero	60.5	62.261	62.687	62.752	1.406	1.146	1.091	1.083
Forcanada	58.6	60.383	60.761	60.837	1.632	1.326	1.270	1.259
Dürrenstein	59.5	63.076	63.766	63.924	1.445	0.957	0.884	0.868
Monte Magro	67.2	70.461	71.081	71.196	0.917	0.632	0.589	0.581

Table 3.1: PSNR and RMSE for the test regions. While comparing with the methods that use only LRDEM, our networks (DSRFB and DSRFO) outperforms FCND (the variant of FCN with only LRDEM as input)

Region	PSNR (in dB (↑))			RMSE (in meters (↓))		
	FCN (with RGB)	Ours (without RGB)		FCN (with RGB)	Ours (without RGB)	
		DSRFB	DSRFO		DSRFB	DSRFO
Bassiero	63.4	62.687	62.752	1.005	1.091	1.083
Forcanada	62.0	60.761	60.837	1.097	1.2702	1.259
Dürrenstein	63.6	63.766	63.924	0.901	0.884	0.868
Monte Magro	71.1	71.081	71.196	0.587	0.589	0.581

Table 3.2: PSNR and RMSE comparisons between FCN (which uses aerial images) and our proposed DSRFB and DSRFO (which use only LRDEM)