Unsupervised Cross Spectral Thermal RGB Depth Estimation

A Project Report

submitted by

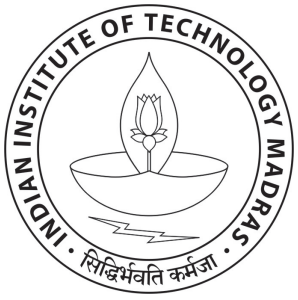
K VENKATA NARASIMHA KARTHIK

(EE15B092)

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY



DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

July 2020

THESIS CERTIFICATE

This is to certify that the thesis titled **Unsupervised Cross Spectral Thermal RGB Depth Estimation**, submitted by **K Venkata Narasimha Karthik**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Kaushik Mitra
Research Guide
Assistant Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date: 3 July 2020

ACKNOWLEDGEMENTS

I express my sincere gratitude to my guide Dr. Kaushik Mitra for giving me an opportunity to work under his supervision. His insights and guidance have helped me understand the subject in greater detail and kept me motivated at all times.

I would also like to thank Honey Gupta for helping me in my research. Her contribution is invaluable to the project. She has consistently helped me identify pertinent issues with my project.

Last but not the least, I thank my parents Mr. Umamaheswar and Mrs. Krishna Kumari and my sister Harika for their constant support.

ABSTRACT

KEYWORDS: Cross-modality, Stereo Matching, Deep Learning, Disparity Map Estimation,Matching Cost, Feature Map Comparison

Stereo Matching attempts to identify the depth of objects in a scene given two images of the said scene from different angles. Depth estimation becomes difficult if the modality of the two images are different, due to appearance differences. Currently, hand-crafted feature descriptors provide the best outputs for disparity map estimation.

The focus of this thesis is to provide a novel attempt to solve this problem by using a Siamese style CNN. The CNN takes in both the images, and generates a disparity map.The aim is to be able to estimate the disparity map for various scenes, both outdoors and indoors.

The different network architectures proposed for solving the problem are first discussed,along with the key idea behind each network.We then discuss the final proposed network that generates the best output.By comparing feature maps of the thermal and RGB images, we show that this is a plausible method for disparity map estimation.The method has been trained and tested on the CATS dataset.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

IITM	Indian Institute of Technology, Madras
FMN	Feature Map Network
DMN	Disparity Map Network
CNN	Convolutional Neural Network
conv	Convolutional layer block
Disp	Disparity

CHAPTER 1

INTRODUCTION

Stereo Matching is a fundamental problem of computer vision. Given two images, the aim is to estimate the depth of the scene. The central theme of this thesis is to compute the depth when one image is thermal infrared (LWIR) and the other is an RGB image.

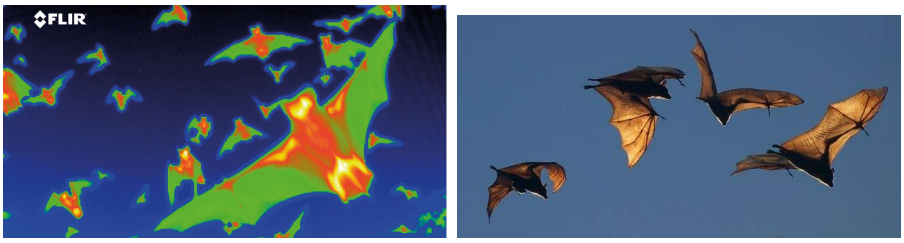


Figure 1.1: Different Modalities

The two images are from different modalities. As such, they contain different aspects of the same scene. The thermal image works well in case of dark surroundings, as objects can be easily identified on the basis of their temperatures.RGB images fail in such situations, since the brightness and contrast is low. However, the LWIR modality suffers when temperature differences are minimal. The RGB image, works well in such cases. The objective to leverage the strengths of the two , while ironing out the differences, and generate a disparity map to estimate the depth.

A disparity map is a 2D representation of the shift in pixels of an image of a scene to obtain the image taken by the other camera. Thus disparity map can also be viewed as input to a warping function W such that for every pixel I^l_{ij}in

the left image I^l, to obtain the right image I^r:

$$W(I^l) = I^l_{i+d_{ij}j} = I^r$$

Given the camera parameters, a depth map can be easily obtained from a 2D disparity map from the relation: $z = \frac{fb}{d}$, where f is the focal length, b the distance between cameras,z the depth and d the disparity value.

Cross-Modality immediately throws up several problems. Different imaging techniques capture different aspects of a scene. As a result, a naive comparison of images wouldn't be an appropriate method to generate an accurate disparity map. Regular networks used for RGB stereo matching would fail due to appearance differences in the two domains.

The problem is compounded by the fact that despite being in different domains, a common platform is required to bring the two modalities together to enable the comparison required to generate a disparity map.

In order to not constrain the network to only certain types of images, unsupervised training is adopted. Although harder to train, the resulting network is more general and can be applied to various scenes, which is the goal

Since the two images are of the same scene, the focus of this thesis is a novel approach by taking advantage of this fact and comparing feature maps of the thermal and colour images so as to generate the disparity map.

In short, the contribution is as follows:

- Train a reconstruction network from scratch such that aligned, rectified images have similar feature maps.
- Train a disparity map network by taking in the images and generating a

disparity map.Using the map, the images are warped and feature maps of the input are compared with those of the warped images to train the network.

Several architectures have been developed in order to solve this problem in an unsupervised manner. Each method has its merits and demerits. However, the final proposed network is the Correlation network. The following methods have been discussed:

- Naive encoder-decoder network
- VGG features based network
- Two Networks model
- Correlation Network

CHAPTER 2

RELATED WORK

2.1 Unsupervised Depth Estimation

Zbontar and LeCun was the first to use a CNN for stereo matching using imaged patches. Luo, Schwing, and Urtasun use a Siamese network and treat the problem as a multi-class classification, where the different classes are all possible disparity values.The data is modelled as a probability distribution. They also join the features of their siamese network with a inner product which produces very good results. One of the inspirations of the proposed network in this thesis is the architecture in this paper. Garg et al. was the first to use a warping-based unsupervised method to estimate the disparity map. The approach was to warp the right image to left using the disparity map and obtain the absolute difference between warped image and the left image. The error, also known as reconstruction error or photometric error, is minimized to learn the estimated disparity map.

Godard, Mac Aodha, and Brostow improved upon this by adding a left-right view consistency term. The model in the paper is closest to the model proposed by this thesis. However, the issue with these methods based on photometric loss is that, they fail when appearance differences are substantial, as in the case of cross-modality.

2.2 Cross-spectral stereo matching

Previous methods had mainly two approaches to this problem:

- Generate an invariant between the two modalities

- Generate a thermal image from the rgb image and vice-versa and compare.

2.2.1 Modality Invariance

Most of the methods adopted here are traditional. They are still the go-to methods for disparity estimation for RGB-LWIR. There are mainly three kinds of methods, as detailed in [2] and [9]:

- Similarity of pixels in two windows
- Squared Difference of pixels in two windows
- Comparison of binary vector representing windows

In the first catergory we have methods based on mutual information(MI) [14]. These compute the co-occurence of intensities of pixels in windows taken from each image.As such, these methods can generate a good result for the problem at hand.

In the second method, we have methods relying on descriptors such as LSS[13],HOG[3],SIFT[10], etc. Heo, Lee, and Lee proposed Adaptive Normalized Cross-Correlation(ANCC)to tackle illumination changes and camera parameter differences. Pinggera12, Breckon, and Bischof showed that dense gradient features based on HOG achieved better performance than MI and LSS descriptors. Kim et al. proposedDense adaptive self-correlation descriptor (DASC) by im-proving LSS descriptor with random receptive field pooling.

In deep learning methods, Aguilera et al. learned a similarity measurement of cross-spectral image patches.

2.2.2 Image-to-Image translation

The approach taken Zhi et al. for estimating RGB-NIR disparity map is to train the network to convert RGB to NIR images and then perform the comparison.

The approach by Liang et al. is of a similar nature. A cycle GAN is used to train conversion of a RGB image to NIR and vice versa, and from this, a disparity map is learnt. However, these methods have been developed for RGB-NIR wherein the appearance differences are lesser as compared to Thermal-RGB.

2.2.3 Correlation Based

Recently, a third method has been proposed for Thermal Infrared(LWIR)-RGB Spectral Matching using a Siamese network modelled along the lines of [11]. Beaupre and Bilodeau attempt to compare image patches from the different modalities and model the disparity value as a probability distribution. However, the focus of the paper is disparity estimation for human silhouettes and may not generalize for other scenes. This thesis is an attempt to estimate the disparity in general, and not restricted to any particular object in question.

CHAPTER 3

NAIVE NETWORK

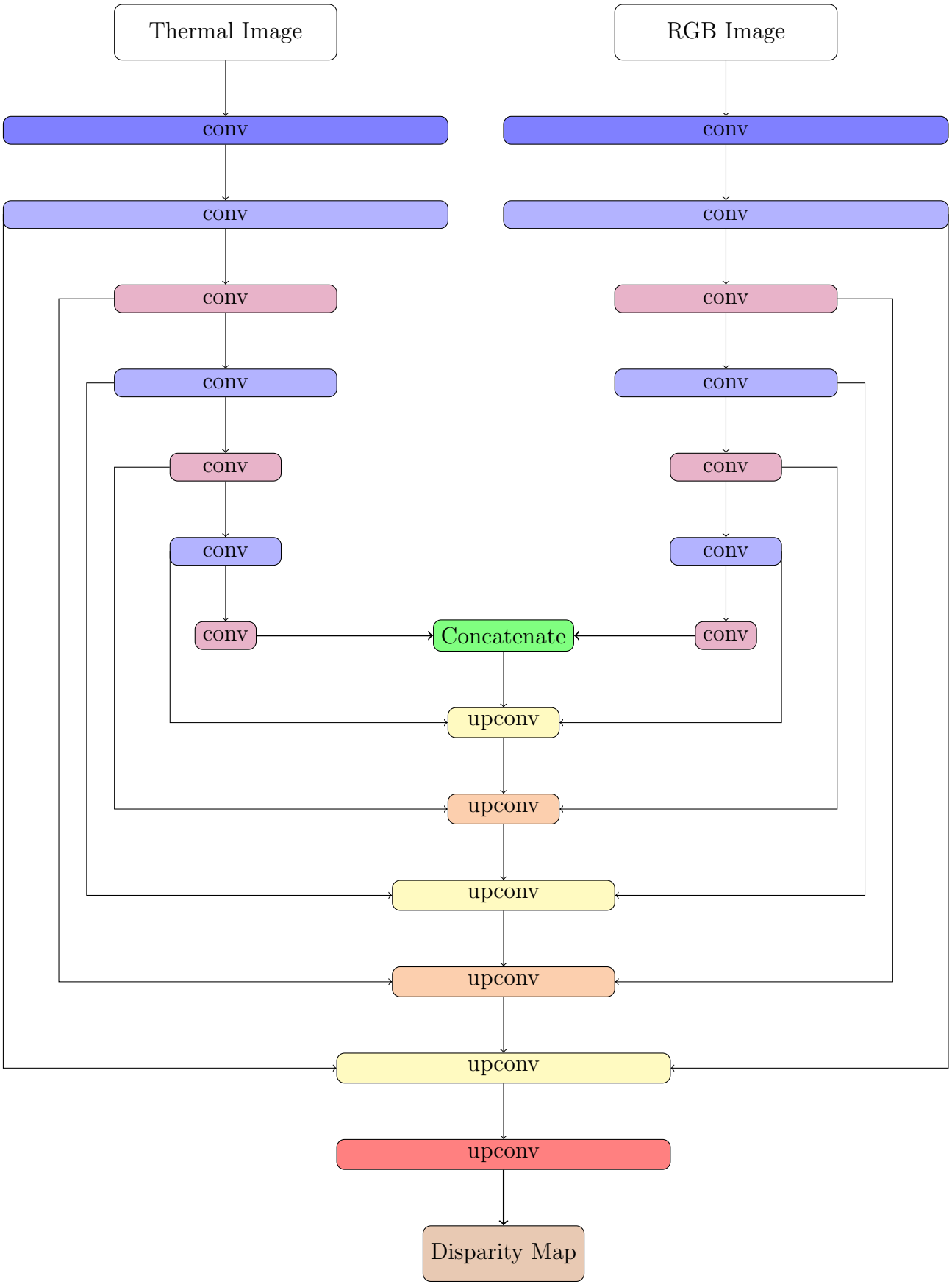
The Naive implementation of unsupervised disparity map estimation network consists of a simple Siamese style encoder-decoder network. The network takes in a thermal image and a RGB image. These are passed separately through several convolutional layers with shared weights and the results are concatenated . This is the encoder.

The output is then once again are passed through convolutional layers with upsampling and skip connections.This is the decoder. The output of the network are two disparity maps, one for each modality.

To enforce the similarity between the disparity map and the images, the loss function is a cosine similarity function between the output and the corresponding images.The training is unsupervised.

The aim is to make the network learn the relevant features with which both images can be compared and estimate the disparity map from those features.The disparity map is learned directly from the images, and thus contains the basic features present in both images. On the flip side, since both the images are radiometrically variant, the disparity map is not very accurate.

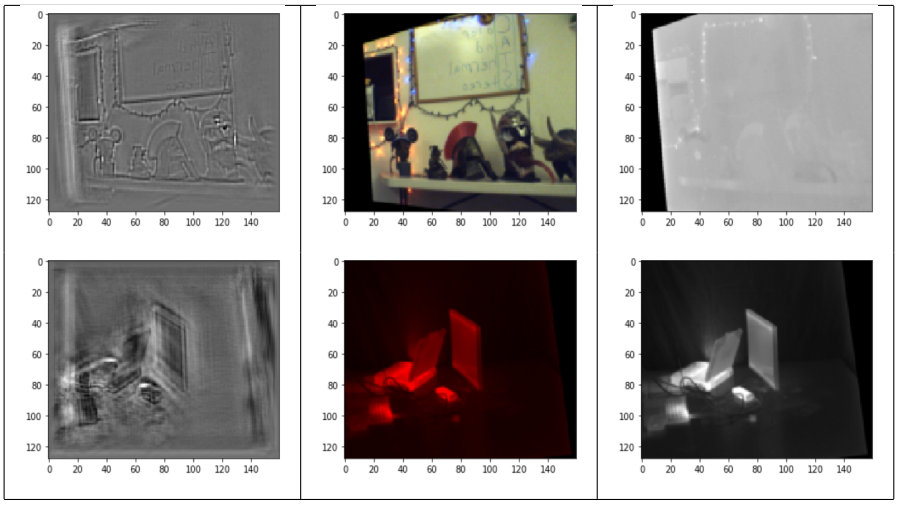
3.1 Network Architecture



Each 'conv' represents a block of two convolutional layers. The layers are designed in such a way that the image dimensions are halved after every alternate 'conv' block. The purple layers represent the encoder with shared weights.

The outputs of the encoder are concatenated and sent through the decoder. The decoder consists of 'upconv' layers designed in such a way that after every alternate block, the image dimensions are doubled. With such an arrangement in place, outputs from the layers in the encoder are concatenated using skip connections as shown above. This is to incorporate the higher-level encoded features while generating the disparity map.This architecture has been inspired from [6].

3.2 Network Results



As we can see, the network is able to capture the features of the given images. However,it struggles to learn the shift in the images.One of the reasons is that the network is forced to learn the disparity map as well as extract features from the input images.

A better approach is to separate feature extraction and disparity map estimation, which forms the basis of the successive networks.

CHAPTER 4

VGG NETWORK

The VGG network is an enhancement over the Naive implementation. Instead of forcing the network to transform both the images to a common feature space, the same network is trained to generate the disparity map.There are two ways to proceed.

First way is, the images are passed through the VGG network and features from the third layer of VGG are extracted.These extracted features are then warped using the disparity map. The warped feature maps are compared with the feature maps of the input images.

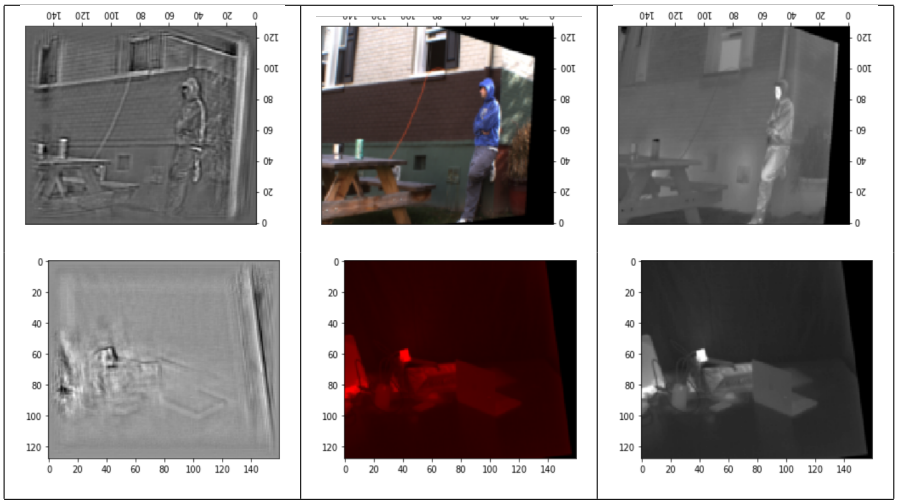
The significant advantage of this network is that the network need not learn the feature transformation and the disparity map at the same time. Given that a VGG network is used for the feature maps, the maps can said to have captured the salient features of the original images.This method is in line with the naive implementation.

The other way is that the images are passed through the network to generate a disparity map.The images are also passed through the VGG network, and the features from the thirs layer of VGG are extracted. The disparity map is then applied on those feature maps in order to warp and compare.

In this method, the network learning the disparity map is same as that of the Naive implementation as shown above.However the difference lies in the focus of the network as well as the approach to extract the features.

However, the VGG network has been trained for an image classification problem. Although the outputs contain salient features, they might not be the best suited for the problem at hand.

4.1 Network Results



The results are not much of an improvement over the Naive implementation. In order to address the issue of feature extraction, one of the ways is to train a network from scratch to extract features, and use it with the disparity map estimation network, which forms the crux of the following networks.

CHAPTER 5

SSIM NETWORK

5.1 Introduction

The SSIM loss network aims to address the issues of the VGG netowrk. The method consists of two networks viz. Feature Map Network(FMN) and Disparity Map Network(DMN).

The FMN is trained to reconstruct the images provided, and feature maps are extracted from the bottleneck layers. The aim of the FMN to generate similar feature maps when a thermal image and a RGB image is taken of the same scene, in the same perspective. This results in the thermal image and RGB image being projected onto a common feature space, where they can be compared more accurately.

The DMN is trained to generate the disparity map given a thermal image and a RGB image. The disparity map is used to warp the input to produce warped rgb and warped thermal images ,respectively. Feature maps are extracted from the input images and warped images using the FMN and are then compared. However, here the DMN is different in terms of the loss function used.

5.2 Feature Map Network

The FMN is an encoder-decoder style Siamese network, inspired by[6]. The input images are passed through shallow CNN layers. The output is then passed through two encoders with shared weights. They are then concatenated and passed to the

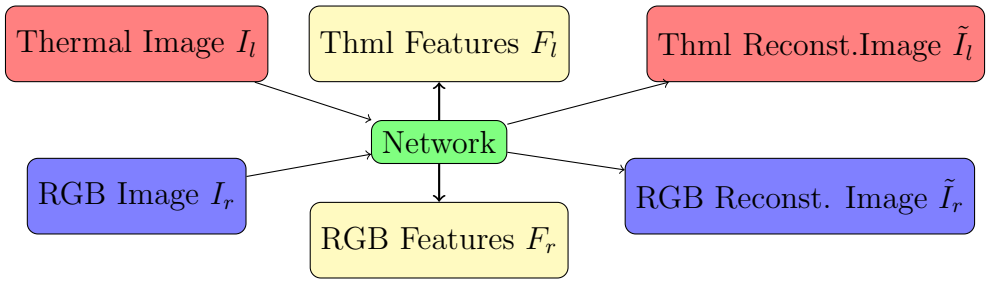
decoder, with skip connections. The decoder tries to reconstruct the original images. Feature maps are extracted from the bottleneck layer of the encoder. The training is unsupervised.

By enforcing reconstruction, we can be assured that the feature maps capture the requisite information from the images. From this information, the network tries to generate the original images.As such, this information can be seen as a representation of the underlying data. Therefore, since the thermal and rgb images are of the same scene from the same perspective, this information should also ideally be matched, which is ensured by the L2 loss between features

In short, the network takes in a thermal image I_l^{tml}, and a rgb image I_r^{rgb}, and generates output images $\tilde{I}_l^{tml}$ and $\tilde{I}_r^{rgb}$. Feature maps of the images extracted are F_l and F_r. The loss function is

$$\alpha_{rct} * (||I_l^{tml} - \tilde{I}_l^{tml}|| + ||I_r^{rgb} - \tilde{I}_r^{rgb}||) + \alpha_{ft} * ||F_l - F_r||$$

In order to emphasize the similarity of feature maps, $\alpha_r ct$ is taken to be 0.1 and $\alpha_f t$ as 1



5.2.1 Implementation Details

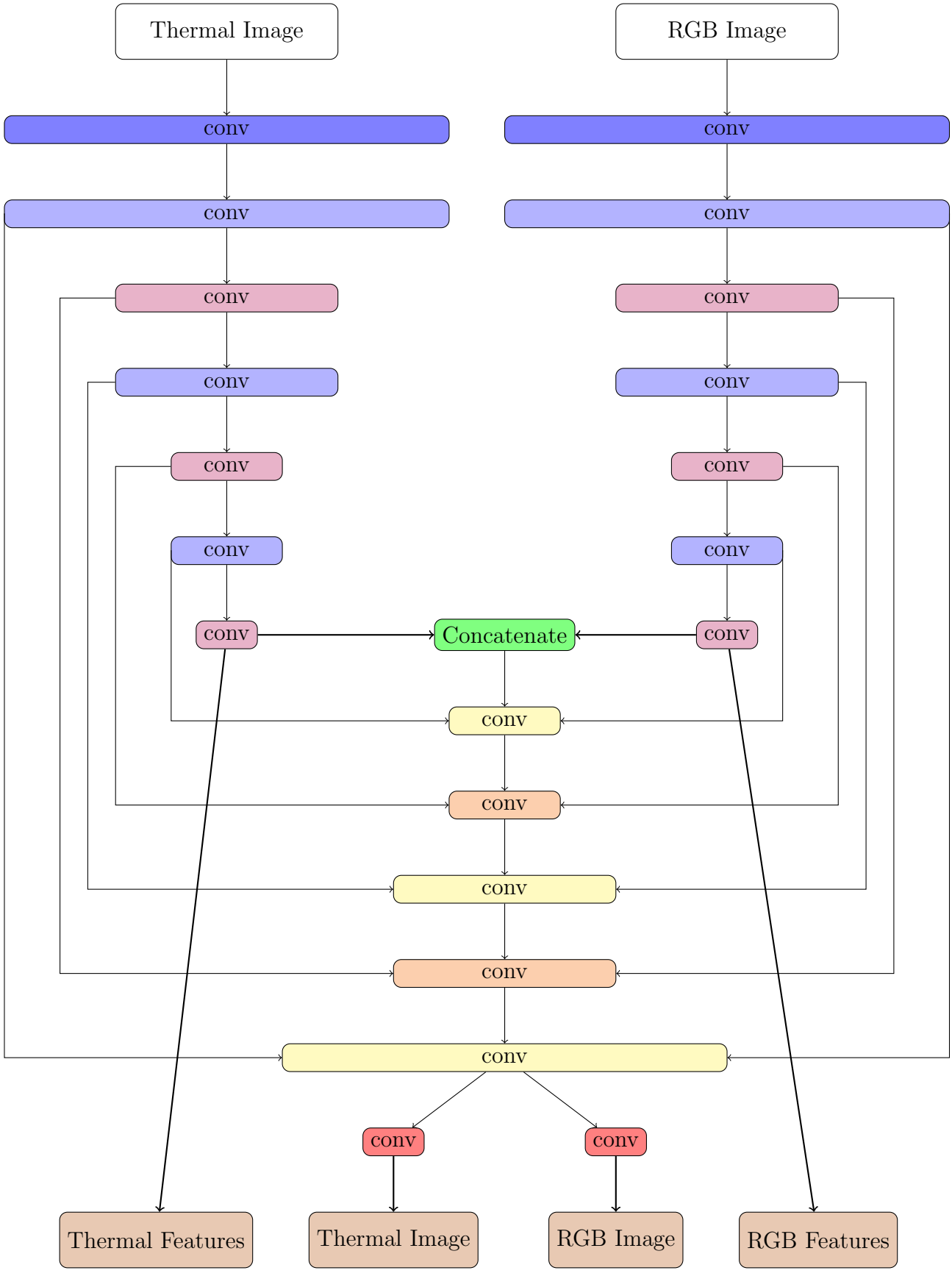
The decoder is a set of convolutional layers with Upsampling. Upconvolutional layers weren't used, unlike [6], in order to avoid checkerboard artifacts. ELU were used for activation, instead of ReLUsLearning rate is 1e-5 for the ADAM optimizer.

To ensure that thermal and rgb images undergo similar processing, the thermal image is concatenated with itself twice to obtain a 3 channel image, similar to the 3-channel rgb image.

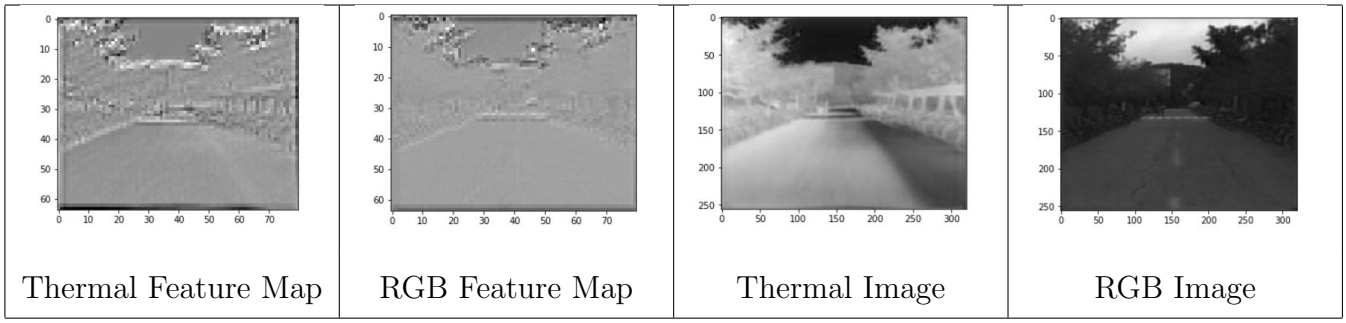
In order to force the encoder to generate similar feature maps for images of the same scene, the KAIST dataset is used. This dataset consists of aligned,rectified thermal-rgb image pairs. In the loss function, an MSE loss between feature maps extracted is used.

Data Augmentation is performed by random cropping of images, while ensuring that the centre of the image always remains in scene.This is to ensure that there will be a portion of the image that will be matched to a corresponding part in the other.

5.2.2 Network Architecture



5.2.3 Network Results

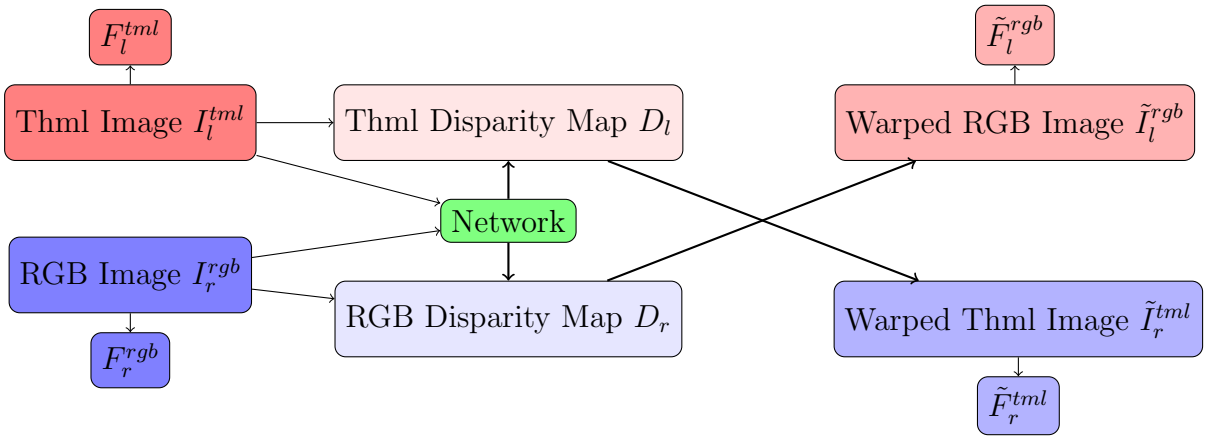


Thermal Feature Map | RGB Feature Map | Thermal Image | RGB Image

5.3 Disparity Map Network

The DMN is of a similar construct as the FMN. However, it differs in the initial and final output layers. The FMN takes in 3 channel inputs and generates 3 channel output so as to ensure similar processing. However, the constraint is relaxed for DMN. For the DMN, difference is processing is preferred in order to obtain slight differences in the two outputs generated.The dataset used is CATS, which has rectified but not aligned images, so as to produce a disparity map.The training is unsupervised.

The DMN takes in the two images, and generates two disparity maps, D_l map for the thermal image I_l^{tml} and D_r map for the rgb image I_r^{rgb}. These maps are then applied to the images to generate warped images $\tilde{I}_r^{tml}$ and $\tilde{I}_l^{rgb}$.



Similar to [6], the loss for the network consists of three components

$$Loss = L_{fs} + L_{am} + L_{ds}$$

5.3.1 Feature Similarity Loss

By the construct of FMN, the outputs must have similar features to the input images, since ideally, the warped images are aligned to the input images. Thus, these 4 images are passed to the FMN to obtain the corresponding feature maps. The loss is:

$$Ł_{fs} = \alpha_{fs} * (||\tilde{F}_r^{tml} - F_r^{rgb}|| + ||\tilde{F}_l^{rgb} - F_l^{tml}||)$$

. This is in correspondence with the idea used for training the FMN.

5.3.2 Appearance Matching Loss

The loss used by [6] is used to ensure that the warped images generated are similar to the original images. Thus the loss is an SSIM loss between the pairs in consideration:

$$L_{am} = \alpha_{am} * [(1 - SSIM(F_l^{tml}, \tilde{F}_l^{rgb})) + (1 - SSIM(F_r^{rgb}, \tilde{F}_r^{tml}))]$$

5.3.3 Disparity Smoothness Loss

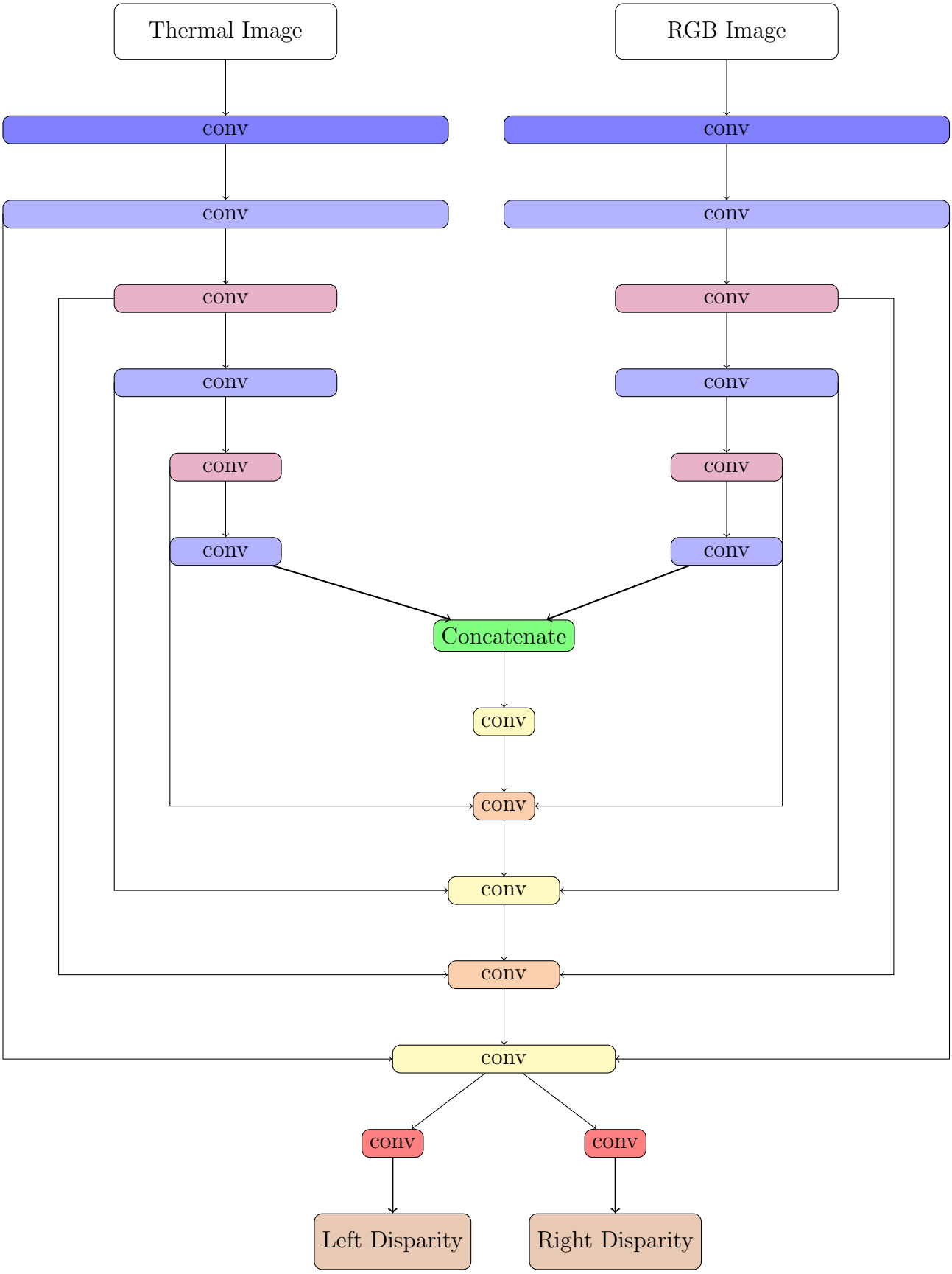
For this too, the loss used for smoothing the disparity map is taken from [6]. This is to ensure that the edges and corners in the disparity map are similar to those in the original images.

$$L_{ds} = |\partial_x(D_l)e^{-||\partial_x(I_l^{tml})||}| + |\partial_y(D_l)e^{-||\partial_y(I_l^{tml})||}| + |\partial_x(D_r)e^{-||\partial_x(I_r^{rgb})||}| + |\partial_y(D_r)e^{-||\partial_y(I_r^{rgb})||}|$$

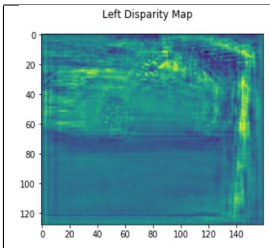
5.3.4 Implementation Details

Similar to FMN, CNN+Upsampling is used so as to avoid checkerboard artifacts,along with ELUs instead ReLUs. This is because the disparity map can have negative values. The ADAM optimizer is used with a learning rate of 1e-5, with Xavier initialization for the weights.

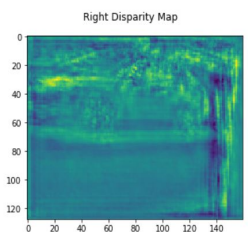
5.3.5 Network Architecture



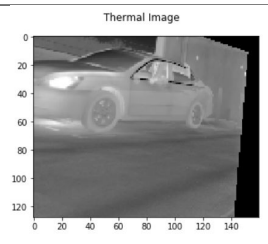
5.3.6 Network Results



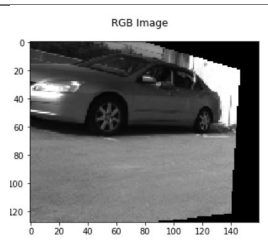
Left Disparity



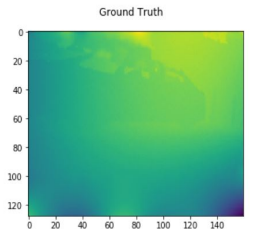
Right Disparity



Thermal Img



RGB Img



Ground Truth

As can be observed, the disparity map is a significant improvement over the previous estimated maps. The network is successful in learning the features as well as trying to estimate the shift of objects. However, it is still far from the actual ground truth.

Given that the output is a reasonable improvement over the previous approaches, the focus is now on ensuring the network is able to predict the shift in the pixels. One of the ways is to shift the feature maps and perform correlation. When a particular object is shifted by the right amount, the correlation between its thermal feature map and RGB feature map would be high.

Since the warping function uses the disparity map to shift the pixels and that different map shifts would result in highest correlation for different patches, by using shift and correlate, an estimation of the disparity map can be made. An enhancement can be achieved by providing the network with this additional information. The proposed network builds on this idea to generate the disparity map.

CHAPTER 6

CORRELATION NETWORK

The correlation network is an improved modification to the SSIM model. This network consists of a Feature Map Network and a Disparity Map Network.

The Feature Map Network(FMN) is essentially the same as that of the SSIM network. The modification is in the Disparity Map Network(DMN) by adding correlation and a method for coarse-to-fine resolution.The training is unsupervised for both the networks.

6.1 DMN

The DMN takes in the thermal and RGB images, and passes both through encoders with shared weights. The output maps are then concatenated. The output maps are also shifted multiple times and each time, they are multiplied with each other. The resultant maps are all concatenated with the output maps. This follows the correlation scheme in [4]. This is then passed to the decoder.

The decoder consists of several convolutional blocks. The first block applies convolutional layers on the maps obtained from the operation described above. The rest of the blocks apply convolutional layers on the input to generate the output required for the final disparity map. At the same time, they also generate another output that is passed through a convolutional layer to obtain an intermediate disparity map.

The intermediate disparity maps provide a useful way to improve training by providing the loss gradient directly to upper layers of the network, since the

loss function is also applied on these maps. The scale of these maps are half or one-fourth the scale of the final disparity map. These maps are also upsampled, and are given as input to the next convolutional block of the decoder, along with the output from the previous blocks. This provides refines the coarser, lower scale feature maps to provide a high resolution output. This approach has been taken from [4] and [6]. In total, three sets of intermediate disparity maps are generated

Following [6], Skip connections from previous layers are also provided to the decoder blocks. The multiplication process is not fool-proof. The skip connections and the concatenations of feature maps helps mitigates issues.

The loss function from the previous network is applied to each scale of disparity map. Each disparity map is upsampled to match the dimensions of the feature maps of the input images.

$$Loss^i = L^i_{fs} + L^i_{am} + L^i_{ds} \quad i\epsilon\{1,2,3,4\}$$

$$Ł^i_{fs} = \alpha_{fs} * (||\tilde{F}^{tml}_r - F^{rgb}_r|| + ||\tilde{F}^{rgb}_l - F^{tml}_l||)$$

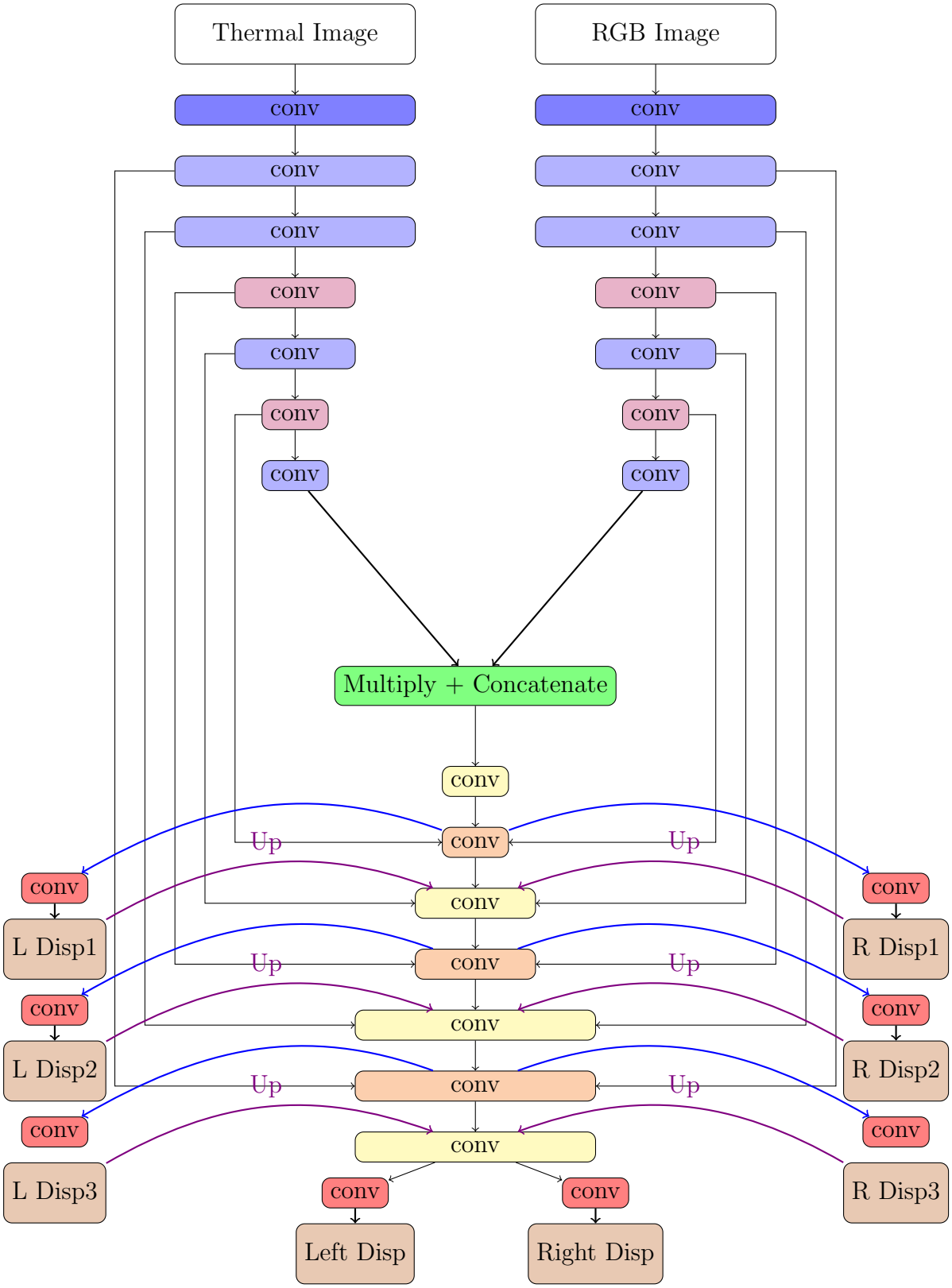
$$L^i_{am} = \alpha_{am} * [(1 - SSIM(F^{tml}_l, \tilde{F}^{rgb}_l)) + (1 - SSIM(F^{rgb}_r, \tilde{F}^{tml}_r))]$$

$$L^i_{ds} = |\partial_x(D_l)e^{-||\partial_x(I^{tml}_l)||}|+|\partial_y(D_l)e^{-||\partial_y(I^{tml}_l)||}|+|\partial_x(D_r)e^{-||\partial_x(I^{rgb}_r)||}|+|\partial_y(D_r)e^{-||\partial_y(I^{rgb}_r)||}|$$

6.1.1 Implementation Details

Similar to FMN, CNN+Upsampling is used,along with ELUs instead ReLUs. This is becuase the disparity map can have negative values. The ADAM optimizer is used with a learning rate of 1e-5, with Xavier initialization for the weights.

6.1.2 Network Architecture



CHAPTER 7

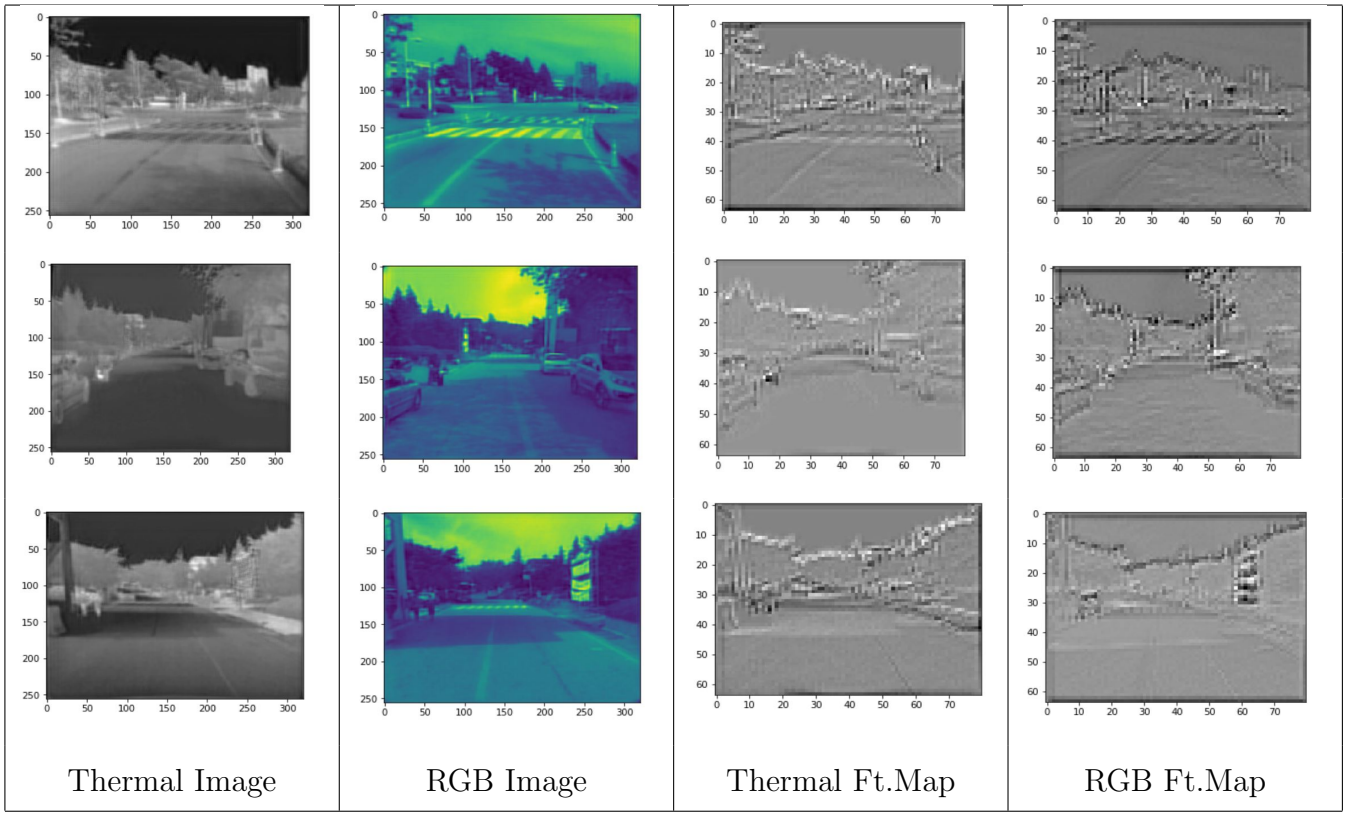
RESULTS

The final proposed network is the correlation network. It combines all the ideas presented so far. The central idea is still the same: images of the same scene should have similar features

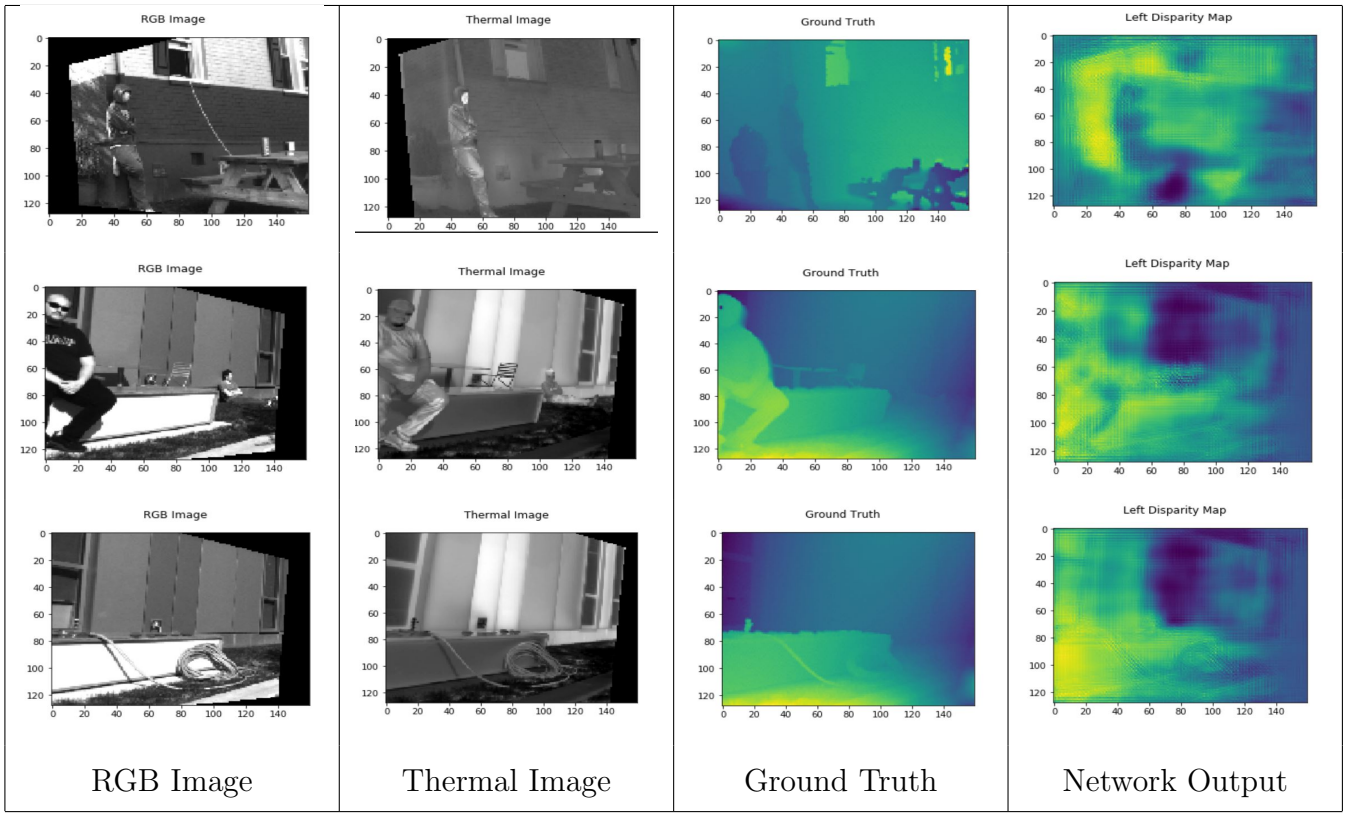
7.1 Results

The results obtained are as follows:

7.1.1 Feature Map Network



7.1.2 Disparity Map Network

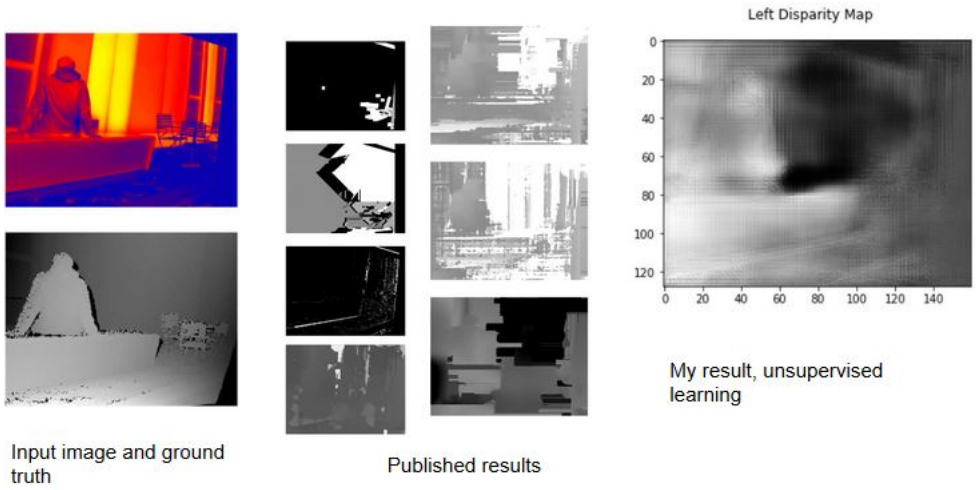


The FMN takes in images from the KAIST dataset, wherein the thermal and RGB images are same. This is successfully reflected in the results. The feature maps look almost similar. The slight differences are due to the fact that the network has been trained for image reconstruction. As a result, the radiometric variations generate slight dissimilarities in the feature maps.

The DMN takes in images from the CATS dataset. The outputs resemble the ground truth. This shows that the central idea is indeed a viable method for cross-modality disparity estimation. However, the output map is still not perfect. There is room for a lot of improvement. The network struggles when either or both the images are almost completely dark due to either lighting variation, or temperature uniformity.

7.2 Comparison

The RGB-Thermal depth estimation field is a relatively new one. As such, given the nature of the problem, not much progress has been made yet. A comparison of this network output with the results of other networks, as published in the CATS paper, is given below.[**CATS**]



Compared to the outputs of other networks, the output of the proposed network is significantly better. The network is able to estimate the disparity of both the human as well as the slab below. Most outputs struggle to identify both successfully.

CHAPTER 8

CONCLUSION

The results demonstrate that cross modal disparity estimation using similarity of feature maps generates a reasonably better disparity map as compared to the previous works. Since the network has been trained in an unsupervised manner on images containing objects in different surroundings, the resulting neural network is more general and can be applied to different scenes.

The drawbacks are still significant.The output can be improved vastly. The network struggles when appearance differences are large. Thus there is a lot of room for improvement. However, it is hoped that this work spurs more research in this field.

This work can be extended to multi-modal imaging.The key idea remains the same: images of the same scene in different modalities will contain similar features. An advantage of multi-modal imaging as comapred to thermal-rgb stereo matching is that more information is available at the network's disposal. The network can leverage this extra information to produce better disparity maps.

Bibliography

[1] Cristhian A Aguilera et al. "Learning cross-spectral similarity measures with deep convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 1–9.

[2] David-Alexandre Beaupre and Guillaume-Alexandre Bilodeau. "Siamese CNNs for RGB-LWIR Disparity Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2019.

[3] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1.

[4] Philipp Fischer et al. "FlowNet: Learning Optical Flow with Convolutional Networks". In: *CoRR* abs/1504.06852 (2015). arXiv: 1504.06852. URL: http://arxiv.org/abs/1504.06852.

[5] Ravi Garg et al. *Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue*. 2016. arXiv: 1603.04992 [cs.CV].

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised Monocular Depth Estimation with Left-Right Consistency". In: *CoRR* abs/1609.03677 (2016). arXiv: 1609.03677. URL: http://arxiv.org/abs/1609.03677.

[7] Y. S. Heo, K. M. Lee, and S. U. Lee. "Robust Stereo Matching Using Adaptive Normalized Cross-Correlation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.4 (2011), pp. 807–822.

[8] Seungryong Kim et al. "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2103–2112.

[9] Mingyang Liang et al. “Unsupervised Cross-Spectral Stereo Matching by Learning to Synthesize”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 8706–8713. DOI: 10.1609/aaai.v33i01.33018706.

[10] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110. ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000029664.99615.94. URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[11] W. Luo, A. G. Schwing, and R. Urtasun. “Efficient Deep Learning for Stereo Matching”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 5695–5703.

[12] Peter Pinggera12, Toby Breckon, and Horst Bischof. “On cross-spectral stereo matching using dense gradient features”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. Vol. 2. 2012.

[13] E. Shechtman and M. Irani. “Matching Local Self-Similarities across Images and Videos”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8.

[14] P. Viola and W. M. Wells. “Alignment by maximization of mutual information”. In: *Proceedings of IEEE International Conference on Computer Vision*. 1995, pp. 16–23.

[15] Jure Zbontar and Yann LeCun. “Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches”. In: *CoRR* abs/1510.05970 (2015). arXiv: 1510.05970. URL: http://arxiv.org/abs/1510.05970.

[16] Tiancheng Zhi et al. “Deep Material-Aware Cross-Spectral Stereo Matching”. In: June 2018, pp. 1916–1925. DOI: 10.1109/CVPR.2018.00205.