UNCONSTRAINED DYNAMIC SCENE DEBLURRING FOR DUAL-LENS CAMERAS

A Project Report

submitted by

NITHIN GK

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR AND MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JUNE 2020

THESIS CERTIFICATE

This is to certify that the thesis titled **Unconstrained Dynamic Scene Dual-Lens Deblurring**, submitted by **Nithin GK**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor and 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.

Prof. A.N. Rajagopalan
Research Guide
Professor
Dept. of Electrical Engineering
IIT Madras, 600 036

Place: Chennai

Date: May 2020

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep sense of gratitude to my guru and advisor **Dr. A.N. Rajagopalan** for his constant support, motivation and the wonderful ideas he has discussed that throughout the past year. I was able to learn a plethora of topics thanks to him. His ideas and guidance motivated me to explore new areas and understand the working of many real-life applications.

I would like to thank **Mahesh Mohan**, PhD student at IPCV lab, for his guidance throughout the thesis. This work wouldn't have been possible without Mahesh's deep commitment.

I would like to express my gratitude to **Kuldeep Purohit**, PhD student at IPCV lab, for his constant motivation and the intense brainstorming discussions on the working of certain state-of the art architectures.

I am very much thankful to **Kranthi Kumar Rachavarapu**, PhD student at IPCV lab, for his constant support from being my teaching assistant in ISP course to the time spent patiently explaining some difficult topics throughout the last year.

I would like to thank **Sharath Girish** for his help in providing and guiding through the code for the conventional methods.

I would like to thank **Gautam Sreekumar, Shyam Shankar, B Akhil, Kevin Cherian and Atul Antony** for their support and being there for me in the most difficult of times.

I would also like to thank **IIT Madras** for the wonderful five years of my life and providing me the opportunity to learn under admired professors and providing all the necessary resources

Finally I would like to thank my parents and the god all mighty who has been there throughout, guiding me helping me to stay motivated without which nothing would have been possible

ABSTRACT

KEYWORDS: Dual-Lens, De-blurring, Deep Learning

With the recent advances in augmented reality, autonomous driving and robotics, the need for good quality 3D images have increased. Despite improvements in 3D cameras, the image-pairs captured is still prone to dynamic scene blurs, combined with the difference in the resolution and exposure between the image pairs, the view-consistency of the image pairs are broken, and the image qualities vary. Thus, forbidding their usage in detail-oriented 3D reconstruction and scene understanding applications that require equal quality view-consistent image-pairs. We tackle this un-addressed problem of unconstrained dual-lens(DL) dynamic scene deblurring by an image adaptive multi-scale based coherent fusion approach. In this work, we take into account the facts that $1)$ The epi-polar error reduces when we down-sample the image pairs, $2)$ The image pairs could contain complementary features which when incorporated into each other for a win-win situation. In effect we address three important problems in the area of unconstrained DL deblurring. We also address the inherent problem in unconstrained DL deblurring that violates the epipolar constraint by introducing an adaptive scale space approach. Our signal processing formulation allows accommodation of different image-scales in the same network without increasing the number of parameters. We then address the root cause of view-inconsistency in the generic DL deblurring network using a coherent fusion module. Finally, we propose a filtering scheme to address the space variant and image-dependent nature of blur, with guaranteed stability. We also build an unconstrained-DL dataset with dynamic scene image pairs of different resolutions and exposures. Comprehensive experimental results on our dataset show that we achieve a new state-of-the-art performance for the unconstrained DL dynamic scene deblurring problem.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Network
DL	Dual Lens
RNN	Recurrent Neural Network
ASPP	Atrous spatial pyramid pooling
BF	Bootstrapped
SA	Scale adaptiveness
AF	Scale adaptive filter
MAE	Mean absolute error
CF	Coherent Fusion

CHAPTER 1

Introduction

Motion blur is a phenomenon that not only introduces artifacts to image but also renders it useless for many vision tasks. Motion blur can be caused due to camera motion, dynamic object motion or both. This makes methods that are exclusive only to deblurring due to camera motion ineffective in the presence of dynamic object motion and thus brings forward a need to tackle blur caused due to dynamic blurs

In this thesis, we propose a solution the problem of dynamic blur for dual lens images. As opposed to the single lens case, dual lens de-blurring requires additional attention to the stereo-cues. This renders methods for single lens deblurring to produce erroneous results when used for the dual lens case. A solution to tackle the dual lens blur must take into account the stereo-cues and information from both views before arriving at a result.

A yet un-approached problem is the case of unconstrained dual lens blur where the exposure times and resolutions of both views could be different. This results in the feature loss due to blur in both images to be different. Hence, unconstrained Dual lens-deblurring has to ensure consistency between the left right views, e.g., through fusing good complementary features

Another issue is the space variant and image dependent nature of blur, this aspect is rarely approached in deep learning networks. Ideally a network should be capable of having different receptive fields and adapt according to the blur.

For unconstrained dual lens deblurring there exists only one work in literature [Mohan *et al.* (2019)], but it works only for the case of camera induced blur and takes a lot of time(20 mins) for deblurring a single stereo pair. The advantage of using a deep network to solve the deblurring problem is that deep networks are extremely fast and results could be obtained in about 0.4 seconds

1.1 Contributions of this thesis

- For the first time in literature, we study the phenomenon of unconstrained blurring in dual lens configuration
- We bring out the issue of scene inconsistent depth and propose a multi-scale network in order to solve it
- We address the problem of view inconsistency and propose a coherent fusion module to solve it.
- We propose a filter module for addressing the space-variant and image dependent nature of dynamic scene blur.

CHAPTER 2

Theory Background

In this chapter we will discuss about some basic topics in the area of deep learning, CNNs and learning methods

2.1 Convolutional Layer

A convolutional layer in terms of deep learning is a layer that is used to extract features from an image. Mathematically, I is the input if W is a weight matrix, b as scalar bias value then the output of the convolutional layer will be

$$O = \sigma(W.I + b)$$

Where, σ is the nonlinearity used

2.2 Convolutional Neural Network

Deep Convolutional Neural Networks(CNNs) were first introduced in the year 2012 in the work of [Krizhevsky *et al.* (2012)], where it was able to obtain an improvement of 10.9% as compared to the second best entry. The network architecture was the first to use a deep network of convolutional layers and fully connected layer for the task of image classification,

After the success of Alexnet, CNNs have been used feature extraction in a wide variety of vision tasks like Image classification, Object Detection, Object Tracking, Image Segmentation, Super Resolution, De-blurring etc

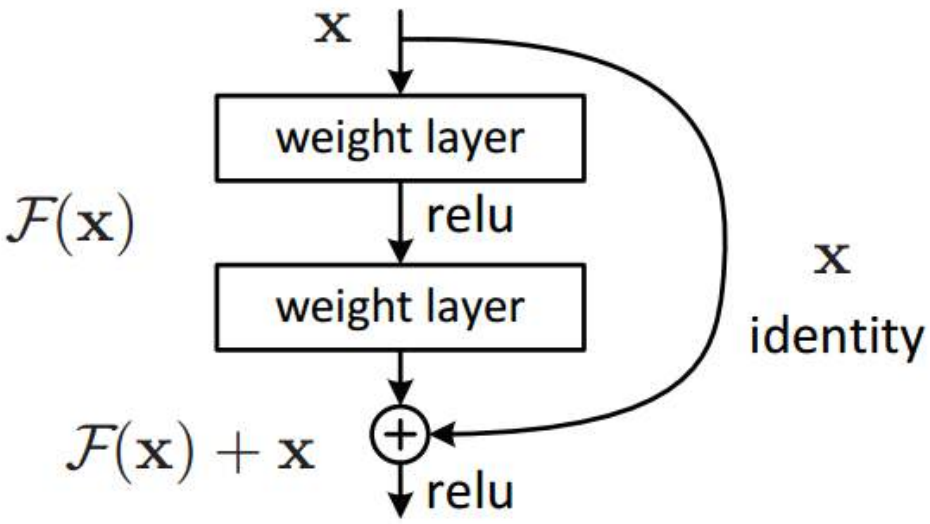

Figure 2.1: A Residual Block. Figure from:-He *et al.* (2016)

2.3 Residual Network

After the success of AlexNet, the state of the art results were obtained by increasing the depth of the network to obtain better accuracy, While AlexNet had 5 convolutional layers, VGG Network [Simonyan and Zisserman (2014)] and GoogleNet [Szegedy *et al.* (2015)] had 19 and 22 convolutional layers respectively.

But increasing the depth of the network further makes the network hard to train because of the vanishing gradient problem as the gradients propagated to further layers became very small because of multiple multiplications. The solution to this problem was introduced in the work of He *et al.* (2016) where identity shortcut connections were made for better gradient propagation and this can be seen in figure 2.1.

2.4 Dilated convolution

Dilated convolution for deep networks[Yu and Koltun (2015)] was an idea initially introduced for segmentation tasks , and later this was used in a wide variety of computer vision tasks. The main advantage of using dilated convolution is that it helps to give a larger receptive field without an increase in number of parameters. For example in the figure 2.2, a dilation rate of 2 and 3 in a 3×3 filter gives a receptive field if size 5×5

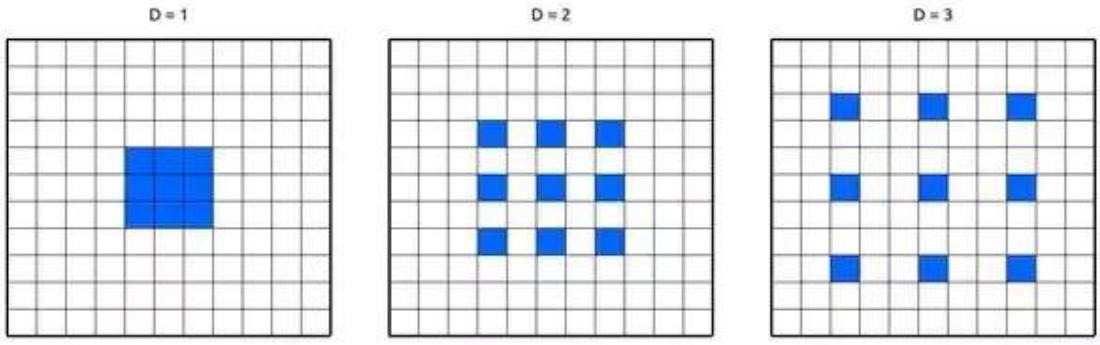

Figure 2.2: Dilated convolution

and 7×7 respectively.

2.5 Atrous spatial pyramid pooling

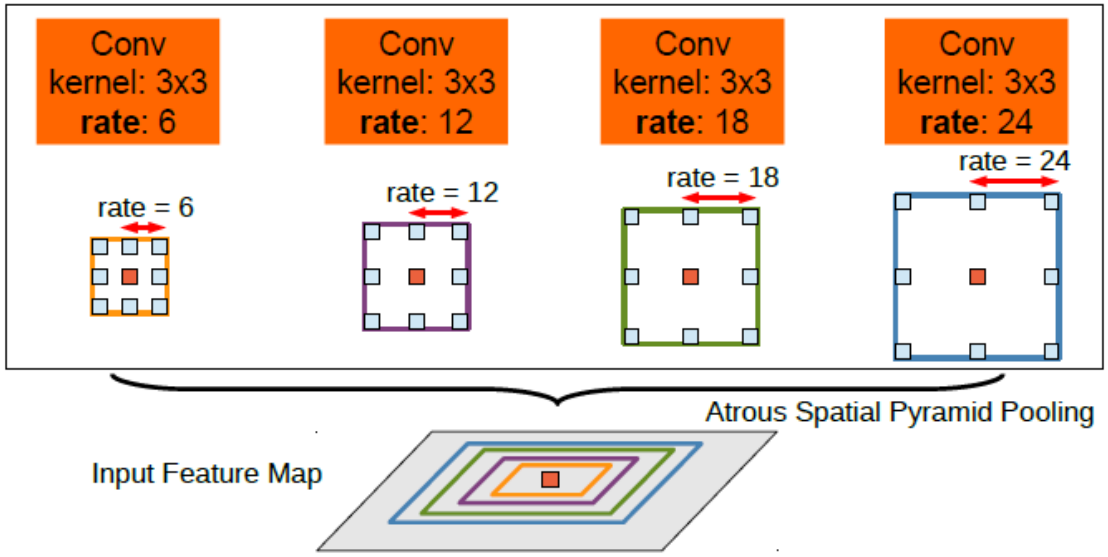

Figure 2.3: The ASPP module taken from Chen *et al.* (2017)

This idea was first introduced in the paper Chen *et al.* (2017) for the task of image segmentation. In tasks like segmentation, where the size of an object of interest may vary, using normal convolutional layers of a very small receptive size may fail to capture the whole image details. A solution for this would be to use deeper networks for the same task, but this causes vanishing gradient problem and difficulty in training. The proposed solution as shown in figure 2.3 uses concatenated features obtained after passing the input image through filters of different dilation rate, this not only reduces depth of network, but is also does not add much parameters to the network

2.6 Learning rate Scheduling

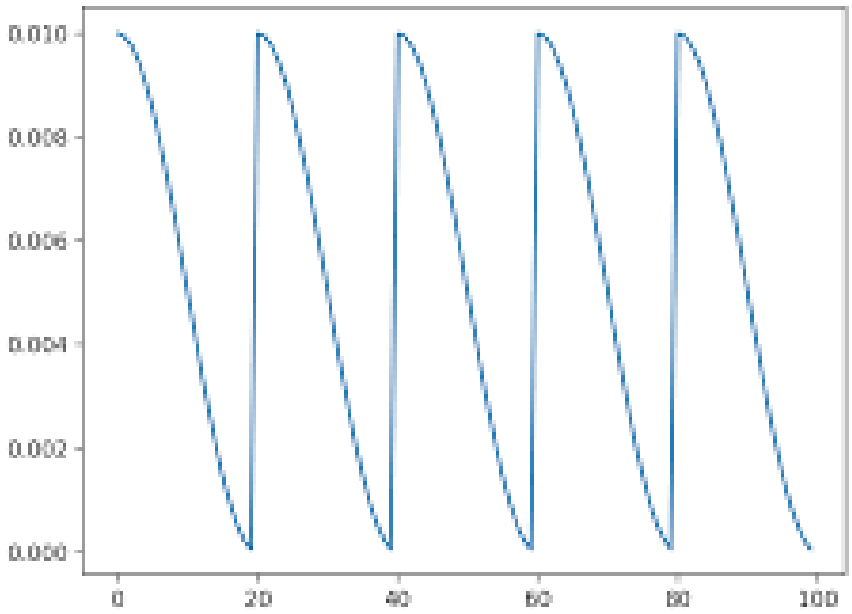

Figure 2.4: Cosine annealing

Learning rate scheduling is adjusting the learning rate for training the model in a predefined order. This technique has been used in networks in order to obtain lower mean square error in case of regression tasks and lower cross entropy error in case of classification tasks. In our work we use the technique of cosine annealing [Loshchilov and Hutter (2016)].

Unlike normal learning rate decay methods, In cosine annealing as shown in fig.2.4 we decay the learning rate from a max value to a min value in through a cosine curve over a fixed number of iterations and suddenly bring it back to its max value. This helps in avoiding the cases where the network is stuck in local minimas, and helps in in faster convergence

2.7 U-Net Architecture

The U-Net Architecture [Ronneberger *et al.* (2015)]initially devised for image segmentation tasks was further used for various computer vision tasks like deblurring, super

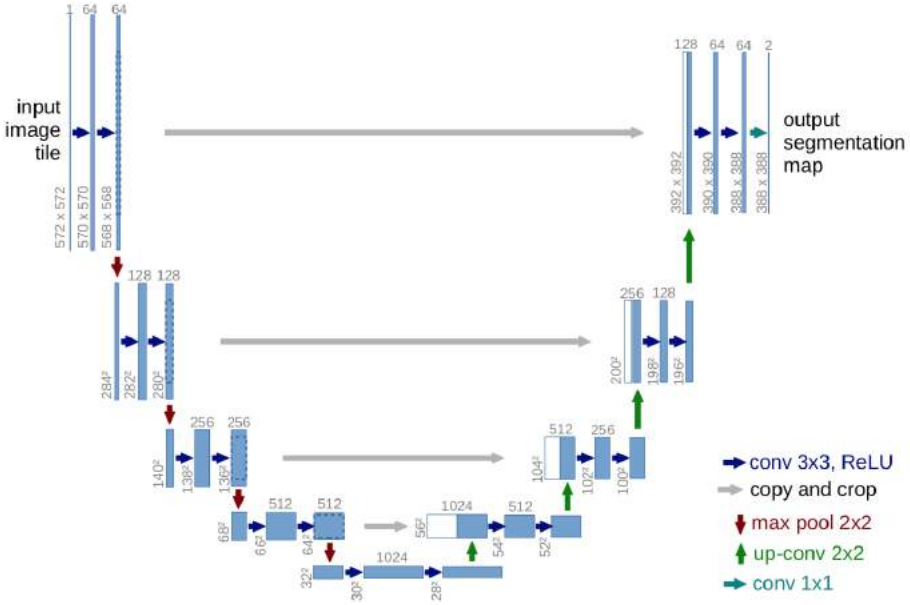

Figure 2.5: U-Net architecture taken from [Ronneberger *et al.* (2015)]

resolution etc. This network consists of a symmetric encoder and decoder. Also the input is appended to the subsequent part of the decoder symmetrically as shown in Figure 2.5

The loss is a computed by a pixel wise soft-max over the final feature map combined with the cross-entropy loss function. These skip connection enable better flow of gradients and makes training of the encoder easier

2.8 Video frame interpolation

In order to produce realistic dynamic scene blur, a common method used for the preparation of dataset is Video frame interpolation. This methods has been used for GoPro[Nah *et al.* (2017)] and Stereo-Blur Datasets[Zhou *et al.* (2019)]. In this technique the frame rate of a video is captured at captured at a low frame rate is increased by extrapolating using the optical flow information between subsequent input frames

2.9 FlowNet

FlowNet[Dosovitskiy *et al.* (2015)] was the first network to utilize CNNs to perform optical flow computation in a supervised manner. In FlowNet, multi channel features of two image frames are produced by passing through 3 independent sets of convolutional layers and a correlation layer further compare patches from each channel and concatenates them, which is further processed by a series of convolutional layers to find the optical flow. In our work, we use FlowNet 2.0[Ilg *et al.* (2017)] for all flow computations as it provides much superior results.

2.10 Stereo Style Transfer

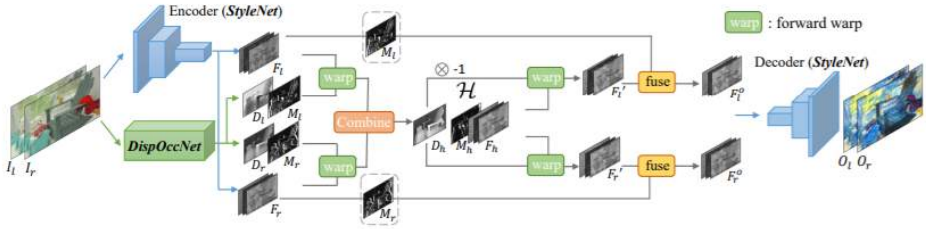

Figure 2.6: Network architechture taken from [Chen *et al.* (2018)]

In usual computer vision supervised tasks images or pair of images are concatenated and passed through a CNN and trained in a supervised fashion in order to obtain the results. The paper[Chen *et al.* (2018)] introduced a new method in which disparity between stereo pairs were estimated and the encoded features were warped and concatenated and decoded to obtain the subsequent left right image pair as shown in Figure. 2.6.

CHAPTER 3

Previous works

In this chapter we will discuss about some neural network architectures used for deblurring and also the previous works on dual-lens deblurring

3.1 Deblurring using Neural Networks

Deblurring is the task of removing artifacts from an images which may be caused due to camera motion or long exposures or some other phenomenon. The goal of deblurring is to recover the sharp image$B = K * S$ and where K is the blur kernel and B is the blurred image and the * operation is convolution.

Deblurring is basically of two types Blind-Deblurring and Non-Blind Blurring, Blind deblurring refers to the task of obtaining the clean image without the knowledge of the blur kernel or the point spread function,

Motion blur is phenomenon in computer vision that not only affects the aesthetics of the image but also affects many vision applications. Motion blur could be caused due to camera motion or dynamic objects or both. So the techniques used for the case of only camera motion doesn't work when there is dynamic scene blur. The task of obtaining the clean image in such cases is refereed to as dynamic scene deblurring.

The subsequent session will discuss in detail about some previous works in the area of dynamic scene deblurring,

3.2 Multi-scale deblurring

Paper summary: Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring

Nah *et al.* (2017) uses a multi-scale approach for the task of dynamic scene deblurring.

Along with a multi-scale approach to restore images in an end-to-end manner, the authors introduced a new large-scale dataset-GOPRO dataset containing realistic blurry images and their clean pair for the task of dynamic scene deblurring. In this technique, the authors have used a slightly modified version of the residual network architecture which enable them to use a deeper network compared to the normal deep CNN network.

The network consists of three stages and the input to these layers are the output of the coarsest stage is up-sampled and concatenated with the finer stage(stage-2) and the subsequent combination is passed through the network in order to deblur and the same is repeated for stage-2 and the finest scale.The mean-square error of all three stages are back-propagated together in order to obtain the clean images. The down-sampled images are Gaussian pyramid images. One important point to note is that all three scale reuses the same weight. The network also consist of a Discriminator which classifies whether the final image obtained is a blurred image or a deblurred one.

Paper summary: Scale-recurrent Network for Deep Image Deblurring

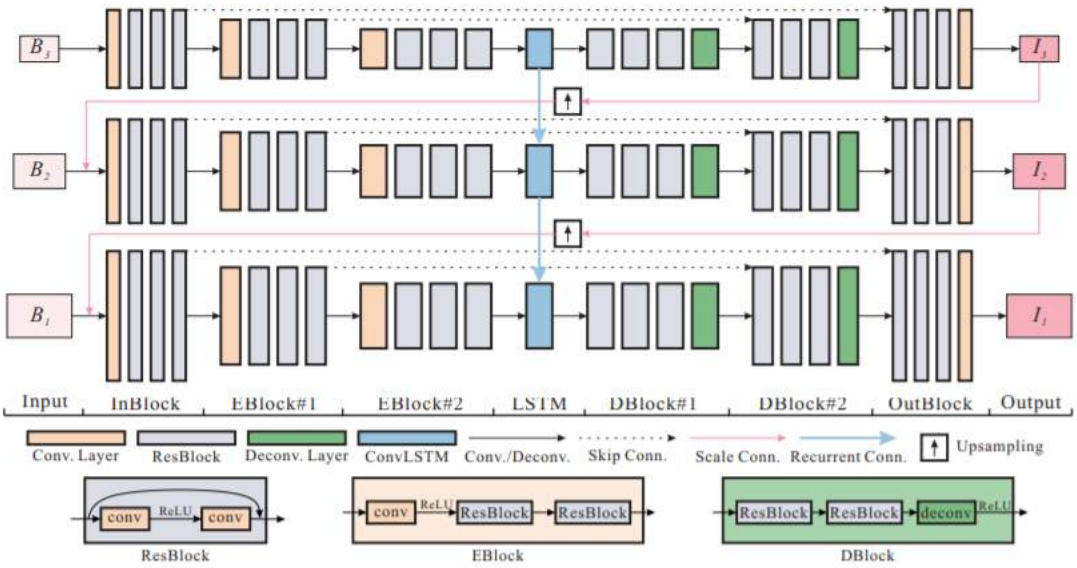

Figure 3.1: Network architecture taken from [Tao *et al.* (2018)]

After the success of the multi-scale approach for dynamic object deblurring. Tao *et al.* (2018) proposed a new network with much simpler architecture and small number of architecture in order to produce the state-of-the art result in dynamic scene deblurring. In this work the authors have used a network with recurrent structure that takes a

series of images down-sampled from the input blurred image and produces the corresponding sharp images.

The network as shown(Figure. 3.1 consists of three scales and along with concatenating the output or coarser scale to finer scale, The output of each scale is defined as

$$I^i, h^i = Net_{SR}(B^i, I^{i+1\uparrow}, h^{i+1\uparrow}; \theta_{SR})$$

Here B^i, I^i are the input, output of a particular scale, Net_{SR} is the proposed network , θ_{SR} the training parameters, and the hidden state h^i flows through layers. The authors have also added a convLSTM in between the encoder-decoder structure for flow of information between subsequent layers. The inputs of coarser scales are obtained by bi-linear down-sampling of the input blurred image

3.3 Other methods for monocular deblurring

Paper summary: Dynamic Scene Deblurring Using Spatially Variant Recurrent Neural Networks

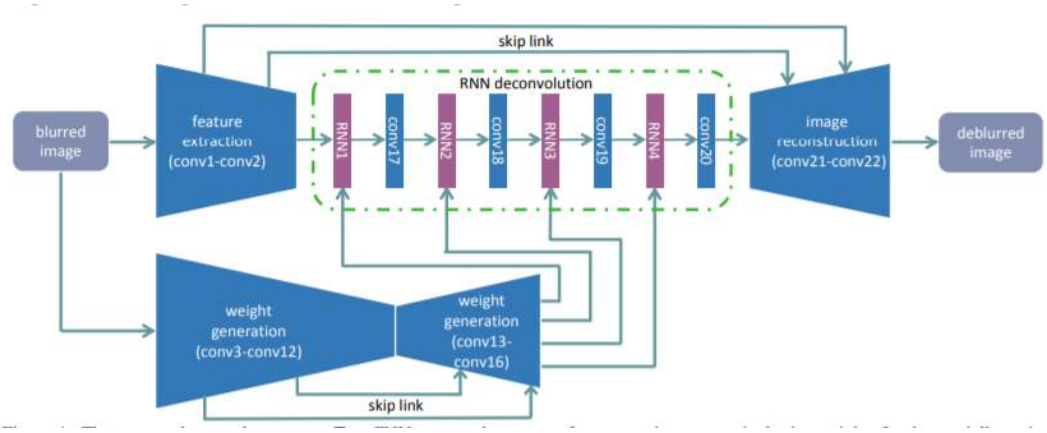

Figure 3.2: Network architecture taken from [Zhang *et al.* (2018)]

Zhang *et al.* (2018) proposed a network (as shown in Figure 3.2) consisting of three Convolutional Neural networks CNNs and on Recurrent neural network (RNN). The first CNN acts as an encoder. The RNN is used as a deconvolution operator on the features extracted by CNN. Another CNN learns the weights for the RNN at different

locations thus making it spatially variant and enables it to model the deblurring process with spatially variant kernel. The third CNN is used for image reconstruction and acts as a decoder that provides the output from features.

The whole network is end-to-end trainable and is able to produce large receptive fields with a small model size.

Paper summary: DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks

The first GAN based approach of dynamic scene deblurring was introduced by Kupyn *et al.* (2018) in their work DeblurGAN. Also the PSNR values does not reach state of the art for this work. The structural similarity and output visual appearance were both state of the art at the point of release of the paper. The main advantage of using a GAN based approach for deblurring is that it is extremely fast because of low number of parameters in the generator.

The DeblurGAN generator consists of two strided convolution blocks along with nine residual blocks each containing a convolutional layer, a normalization layer and ReLu Activation. The Wasserstein distance[Arjovsky *et al.* (2017)] with gradient penalty is used for training the discriminator

3.4 Dual-Lens deblurring

In this section we will discuss about the only existing architecture for dual lens deblurring and also the state of the art work for dual lens deblurring using conventional methods

Paper summary: DAVANet: Stereo Deblurring with View Aggregation

Zhou *et al.* (2019) proposed a network for the purpose of Stereo Image deblurring which utilizes the two-view nature of stereo images and incorporates the features into one another to deblur both images. The authors also introduced a large scale dataset which contains blurred stereo image pairs and their corresponding clean image pairs. The input

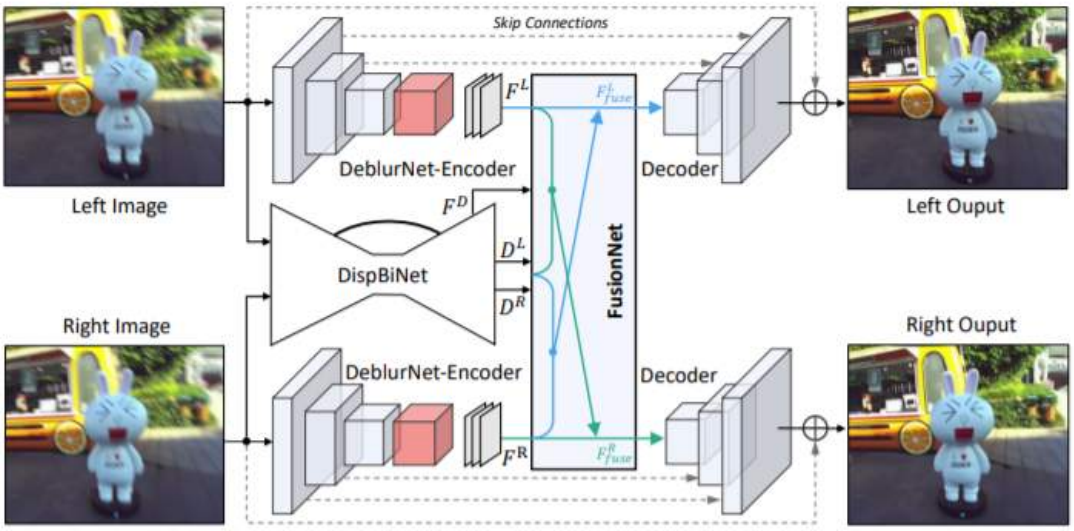

Figure 3.3: Network architecture taken from [Zhou *et al.* (2019)]

to this network are scenes captured using a stereo camera with same characteristics for each individual sub-camera.

The Networks as shown in figure 3.3 consists of three subnetworks, DeblurNet, DispBiNet and FusionNet.

DispBiNet:- This network is a variant of DispNet[Mayer *et al.* (2016)]. Different from DispNet , this network can find the bidirectional disparities in one forward pass. Three stages of down-sampling happens in the encoder part of the network and subsequent up-sampling in the decoder. The network also consists of a variant of ASPP module for enhanced feature extraction.

DeblurNet:- DeblurNet is comprised of a U-Net architecture, i.e., an encoder decoder with residual connections along with a context module. The encoder outputs features of 1/4th the scale of the input. In order to obtain a larger receptive field and extract richer features, the encoder output is passed through the context module

FusionNet:-FusionNet extracts the depth information and the information from the different views. The FusionNet consists of two stages, one for utilizing depth information and the other for extracting information from the other view,i.e The features from the right view is warped using a pooled version of disparity map to produce the left view and a weighed version of the resultant features is appended along with left view encoder features and this along with the depth feature obtained from the difference in views is

passed through the decoder to obtain the left clear image and the same is performed for the right view.

Paper summary: Unconstrained Motion Deblurring for Dual-lens Cameras

Mohan *et al.* (2019) proposed a novel method in order to obtain motion deblurred images with scene consistent disparities. The paper brings out an ill-posedness that is inherent to solving the equation to estimate the clean image pair.i.e.,

$$I_B^n = \Sigma_p w^n(p) P^n(R_p(X - I_c) + I_c + I_b) =$$

$$\Sigma_p w^n(p) P^n(R_p R_n^{-1}(R_n(X - I_c) + I_c - I_c) + I_c + I_b)$$

Here I_B^n is the blurred image I_c is COR, I_b is baseline, P^n is world space projection and $w^n(p)$ is proportion of time exposed to pose p. , as can be seen, the desired solution is $\{P^n(X + I_b), P^w(X)\}$, but $\{P^n(R_n(X - I_c) + I_c + I_b), P^w(X)\}$ is an apparent solution. In order to fix this a convex DL prior is introduced which curbs the relative shifts between MDFs

CHAPTER 4

Scene consistent depth

In the next two chapters we will discuss about some unaddressed problems in the area of unconstrained dual lens dynamic scene motion deblurring caused by existing methods and reason about a solution for the problem.

4.1 Problem Formulation

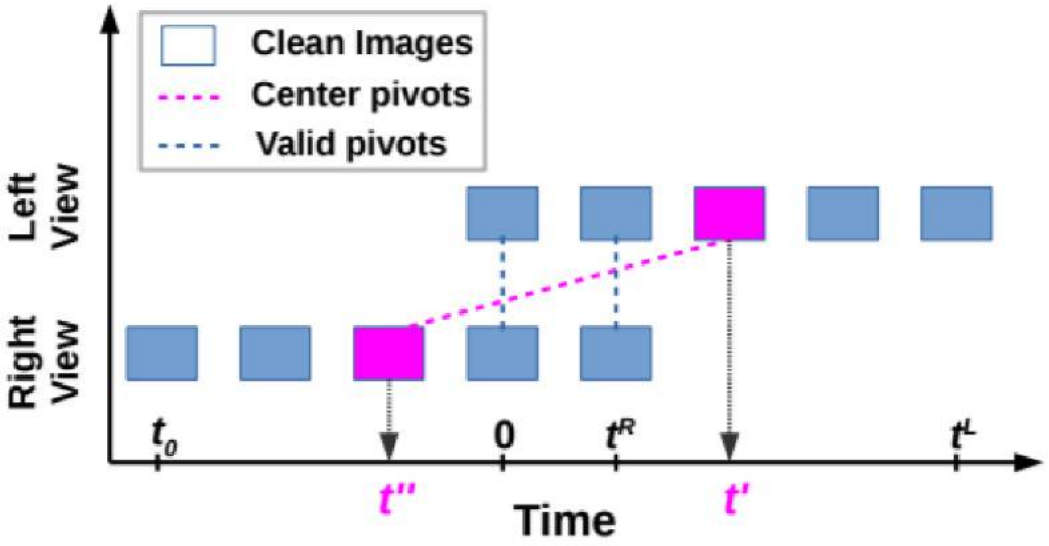

Figure 4.1: Scene Inconsistent Depth

For the cases in which there are dynamic objects, a clean dual lens pair with scene consistent depth is the one refers to pairs captured at the same time instant, in Figure 4.1 this is time 0 or t^R. If this is not the case, the assumption of an epi-polar constraint is violated and this causes many algorithms for 3D reconstruction and scene understanding and for applications such as augmented reality, robotics, and autonomous driving, to fail .

A motion blurred image can be modelled as a summation of clean images over a exposure time. For an unconstrained dual lens exposure setting, i.e., exposures may or may not be identical and fully-overlapping, blurred image-pair $\{B_L, B_R\}$ in the left-right views is given as

$$B^L = \frac{1}{t^L} \int_0^{t^L} F_t^L dt$$

$$B^R = \frac{1}{t^R - t^0} \int_{t^0}^{t^L} F_t^R dt$$

where $\{F_t^L, F_t^R\}$ is the clean dual lens image-pair at time-instant t, and $[0, t^L]$ and $[t^0, t^R]$ are exposure time in the left-right views. The constrained dual lens setting is a special case, where $t^0 = 0$ and $t^L = t^R$ and this means identical, fully-overlapping exposures.

The only existing Dual lens deblurring method using deep networks by utilizing supervised learning from blurred image pair to a clean image pair that is located at a particular time instant. This instant is typically selected at the middle of exposure time. Although this method is apt for the constrained setting, $t^0 = 0$ and $t^L = t^R$. and $\{F_{t'}^L, F_{t''}^R\}$, $t' = t''$being the clean image pair,in the cases of partially overlapping exposures, this causes serious binocular inconsistency as t' and t" are different.

Also, even if the pivots are chosen as the M^{th} and N^{th} fraction of exposure times, the deblurred image-pair can still exhibit binocular inconsistency, and the error increasing with the separation between the pivots $M.t^R$ and $N.(t^L - t^0)$. For an unconstrained exposure where timings$\{t^R, t^0, t^L\}$ can freely vary, there exists multiple choice of pivots which will produce scene-consistent depth.

A method to address the problem of scene-consistent depth has to adaptively select pivots that depends on the input blurred image-pair. A mutual agreement between left

and right-view signals to arrive at an intersecting pivot. Since using single-lens methods for deblurring works by reusing the same network for individual views, mutual information cannot be transferred from one another efficiently. Although the only existing network for dual lens deblurring network promotes a signal flow between views, by appending with a warped version of encoder output of one view to encoder-output of the other view for view-aggregation Zhou *et al.* (2019), the registration involved hinders the control on pivots, registration is necessary for coherently combining the encoder-outputs

4.2 Scale-adaptiveness for Scene-consistent Depth

Since there exists multiple choices of intersecting pivots in unconstrained dual lens setting, the left-right views must establish a mutual agreement to arrive at an intersecting pivot according to the input blurred image-pair. If we consider the standard choice of pivots, i.e., at the center of exposure time or the centroid of blurred images. As shown in Fig. 4.1, this choice results in deblurred left-right images at different time-instants$\{t', t''\}$. Hence, a scene-point with respect to one view can undergo different pose-changes in the other view due to object motion or camera motion or both. The image-coordinate discrepancy of a world-coordinate X in the right-view is given as [Mohan *et al.* (2019)]

$$\Delta x^R = K\left(\frac{X + I_b}{Z} - \frac{R_{\Delta t}X + T_{\Delta t} + I_b}{Z'}\right)$$

where K is the intrinsic camera matrix, lb is the stereo baseline, Z is the actual scene-depth, $\Delta t = t'' - t'$, and $R_{\Delta t}$ and $T_{\Delta t}$ model relative pose-change at t" due to rotation and translations (which include center-of-rotation and object motion) and Z' is the resultant scene-depth. Suppose that we down-sample the left-right blurred images by a factor of D(> 1), then the resultant image coordinate, this discrepancy becomes

$$\Delta x_R^D = D\Delta x_R$$

where$D = \{\frac{1}{D}, \frac{1}{D}, 1\}$ which implies that image-coordinate discrepancies get scaled down according decimation factors. This motivates our scale-space approach (as illus-

trated in Fig.4.2). We select a decimation factor that reduces the maximum discrepancy within a sensor-pitch (i.e., one pixel), so that the binocular consistency holds good in the discrete image-coordinate domain.

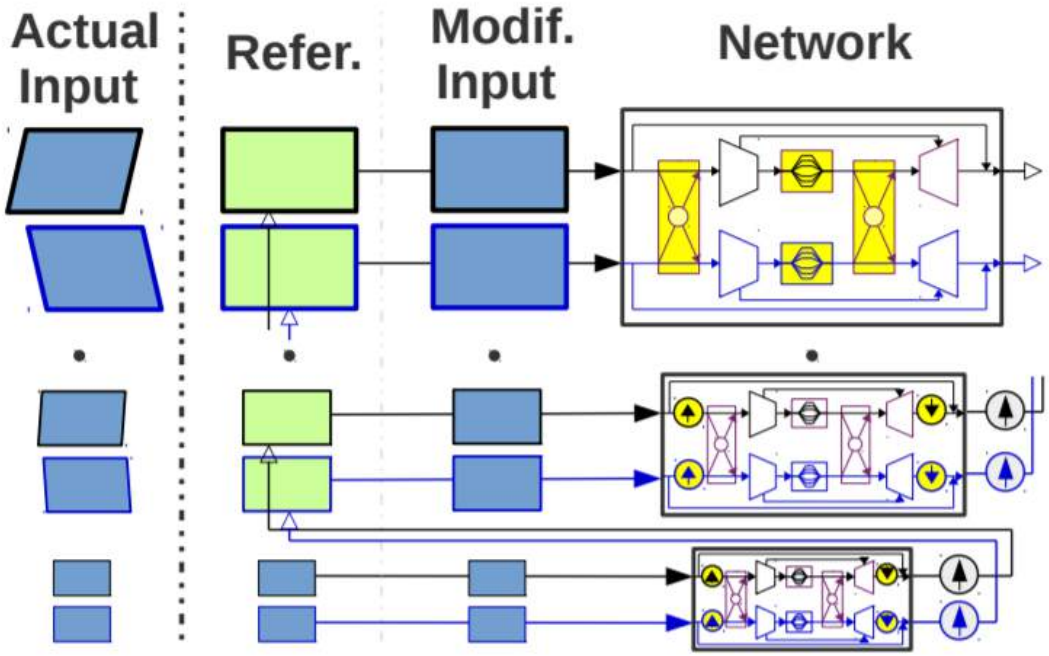

Figure 4.2: Adaptive Scale-space Approach

Next, we consider the coherent deblurred image-pair as the reference to centroid align the binocularly inconsistent blurred image-pair in the higher scale (via registration), which produces a coherent deblurred image-pair. This process is repeated till the fine-scale. Our registration approach is similar to the video deblurring method [Su *et al.* (2017)] where a blurred frame is used as the reference to centroid-align its neighbouring blurred frames, which together produce a coherent deblurred frame. Further, employing deblurred image from a coarse scale as the reference for higher scale is standard practice in conventional deblurring methods [Mohan *et al.* (2019), Whyte *et al.* (2012)]. More importantly, our scale-space approach has to be adaptive with input blurred images, e.g., an input with discrepancy of four pixels ideally requires a decimation factor at the coarser level to be five, whereas a constrained image-pair (i.e., no discrepancy) requires only the fine scale . An optimal method has to adaptively select the number of scales according to the input. Further, as the deblurred image from a lower scale is used as the reference for higher scale (for registration), the step-size between the fine and coarsest scale need to be small ($\frac{1}{\sqrt{2}}$ [Whyte *et al.* (2012)]).

CHAPTER 5

View Inconsistency

In this chapter we will discuss about some unaddressed problems in the area of view inconsistency and stability while using deep networks for dual lens deblurring.

5.1 Problem Formulation

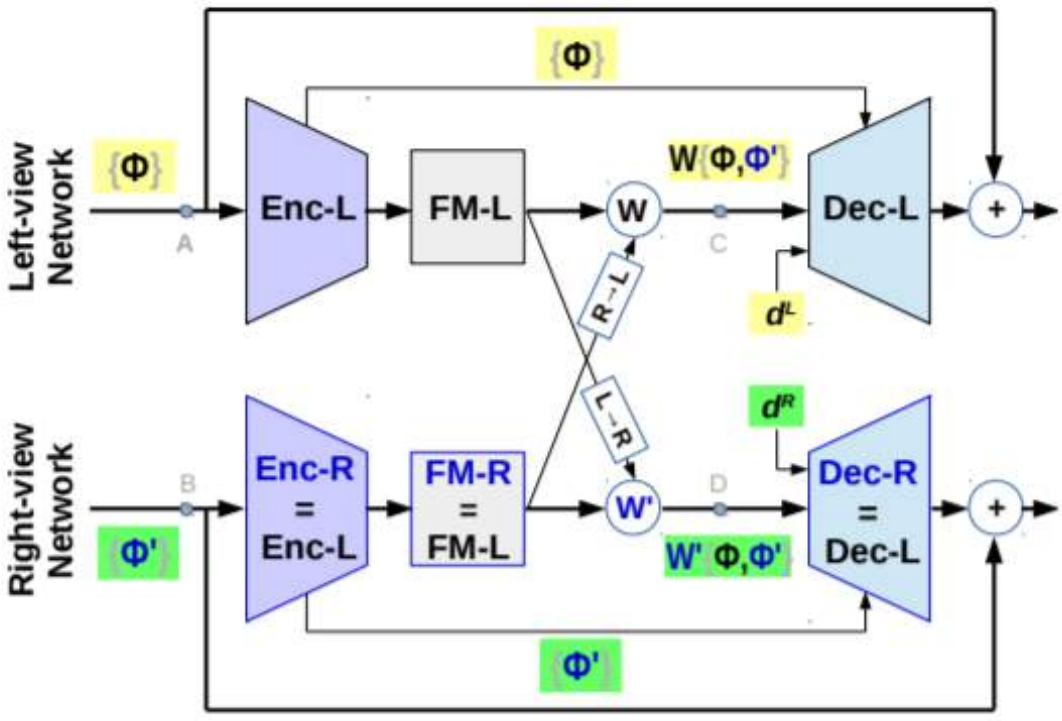

Figure 5.1: View Consistency

The issue of view inconsistency stems from unconstrained dual lens setting of different resolutions and exposure duration ([Mohan *et al.* (2019)]). Here, the feature-loss due to resolution and motion blur can be different in the input images in left-right views. This directly contradicts the assumption followed in the constrained dual lens deblurring methods [Zhou *et al.* (2019)] that the input images have identical resolutions and coherent blur (or identical exposures). Resultantly, these methods produce inconsistent deblurring performance in left-right views, i.e., disrupt view consistency.

We attempt to reason this inadequacy to arrive at a solution. To decouple this problem from the previous one, we assume that dual lens deblurring somehow produces intersecting pivots. As shown in Figure 5.1, a generic architecture for dual lens deblurring consists of symmetrical networks for left- and right-views, with both networks sharing identical weights (in order to not scale-up trainable parameters as compared to that of single-lens methods [Zhou *et al.* (2019)]). The mappings of the left-right networks to deblurred images $\{F^L_{t'}, F^R_{t''}\}$ can be respectively given as

$$T(B^L_{\phi}, f^L_{\phi,i}, W.f^L_{\phi,enc} + \bar{W}.f^{R\rightarrow L}_{\phi,enc}, d^L)$$

$$T(B^R_{\phi'}, f^R_{\phi',i}, W.f^R_{\phi',enc} + \bar{W}.f^{L\rightarrow R}_{\phi',enc}, d^R)$$

where the sets ϕ and ϕ' contains the resolutions and exposures of left-right views, respectively. Features f_i are the i^{th} intermediate-outputs of encoder which are fed-forward to decoder, and f_{enc} is the encoder-output. Bilinear masks W and $\bar{W} = 1 - W$ combine left and right-view encoder outputs after registration (denoted by $R \rightarrow L$) for view aggregation, and $\{d^L, d^R\}$ are depth-features for depth awareness.

The primary reason for the success of dual lens deep learning methods in producing view-consistent output [Zhou *et al.* (2019)], that is for the constrained dual lens set-up, is that the left- and right-view networks or mappings are identical, and more important, signal flowing in those networks are of identical nature (i.e., $\phi = \phi'$).

However, as shown in Figure.5.1 using yellow and green highlights, the same architecture leads to view inconsistency in unconstrained dual lens set-up because now, signal flowing in those identical networks are of different nature (i.e., $\phi \neq \phi'$). A method to address the problem of view-consistency has to ensure signal flowing in left- and right-view networks to be of identical nature irrespective of $\phi \neq \phi'$ or $\phi = \phi'$

5.2 Coherent Fusion for view consistency

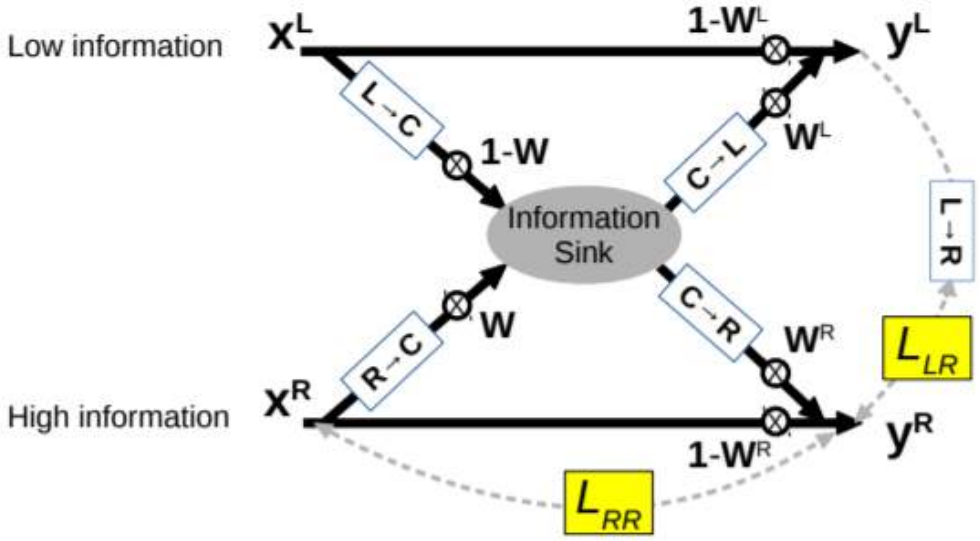

Figure 5.2: Coherent Fusion for view consistency

View inconsistency is caused in dual lens deblurring techniques because there does not exist a sub-part to enforce that the flow of signals in the left and right view networks are identical. As highlighted in Fig. 5.1, this inconsistency stems at nodes A, B and C, D, where the nodes A,B creates an imbalance in the encoder inputs and hence all feed-forward inputs to the decoder and network output, whereas the latter creates imbalance in the decoder inputs. In order to solve this, we introduce a coherent fusion module with two self-supervision costs(Fig. 5.2) . The module enforces the nature of signal in those two nodes to be identical, but equalizes with respect to the signal with more information. Mathematically, the fusion module maps the input left-right view signals $\{x_L, x_R\}$ to output $\{y_L, y_R\}$ and W, W_L, W_R are image-dependent bilinear masks [Zhou *et al.* (2019)]

$$y_s = W \odot x^{R \to C} + \overline{W} \odot x^{L \to C};$$

$$y^L = W^L \odot y_s^{C \to L} + \overline{W}^L \odot x^L;$$

$$y^R = W^R \odot y_s^{C \to R} + \overline{W}^R \odot x^R;$$

. W is a function of the error between $x^{L \to C}$ and $x^{R \to C}$, where $0 \leq W \leq 1, W + \overline{W} = 1$. In addition, the two self-supervision costs are $L_{LR} = ||y^{L \to R} - y^R||^2$and $L_{RR} = ||x^R - y^R||^2$

CHAPTER 6

Stability issues in deep networks

6.1 Problem Formulation

A dynamic-scene deblurring network requires spatially variant mapping (with varying receptive fields), and that has to adaptively vary with blurred images [Zhang *et al.* (2018)]. Intuitively, consider a scenario of static camera, and two dynamic objects at different depths, with the same velocity. Here, the static background exhibits no motion blur, whereas the nearer object exhibits more blur than the farther one (due to parallax [Zhou *et al.* (2019)]). Hence, an ideal deblurring network warrants an identity mapping for background and non-identity mapping for dynamic objects, with relatively larger receptive fields for the nearer one.

Also, those object-positions can freely vary in a fronto-parallel plane, and hence these mappings need to be image dependent. However, the only-existing dual lens deblurring network [Sim and Kim (2019)] do not account for this, whereas ([Zhang *et al.* (2018)]) restricts to fixed and very small receptive fields.

Filters employed in a deep learning network has to be stable, otherwise, finite energy signals like images or feature-maps can get mapped to unbounded or saturated signals, which erroneously steer the network. For the current problem, the existing solution is to perform an image-dependent causal IIR filtering [Zhang *et al.* (2018)]. But it is well known that recurrent filter can be easily unstable.

Specifically, a causal IIR filter is unstable if any of its poles lies outside the unit circle in pole-zero plot . Therefore, it is quite possible that some image-dependent IIR filters can be unstable. Next, we consider a best-case scenario that for all possible blurred images the generated poles lie inside the unit-circle. However, manipulation of network-weights, which is indispensable for model compression (e.g., quantization), can easily shift the poles leading to instability (see Fig. 6.1). Therefore, to guarantee stability while addressing the problem of space-variant and image dependent blur, a

method has to ensure that no poles lie outside the unit circle under any influence (e.g., input images and/or network-weight manipulations, etc

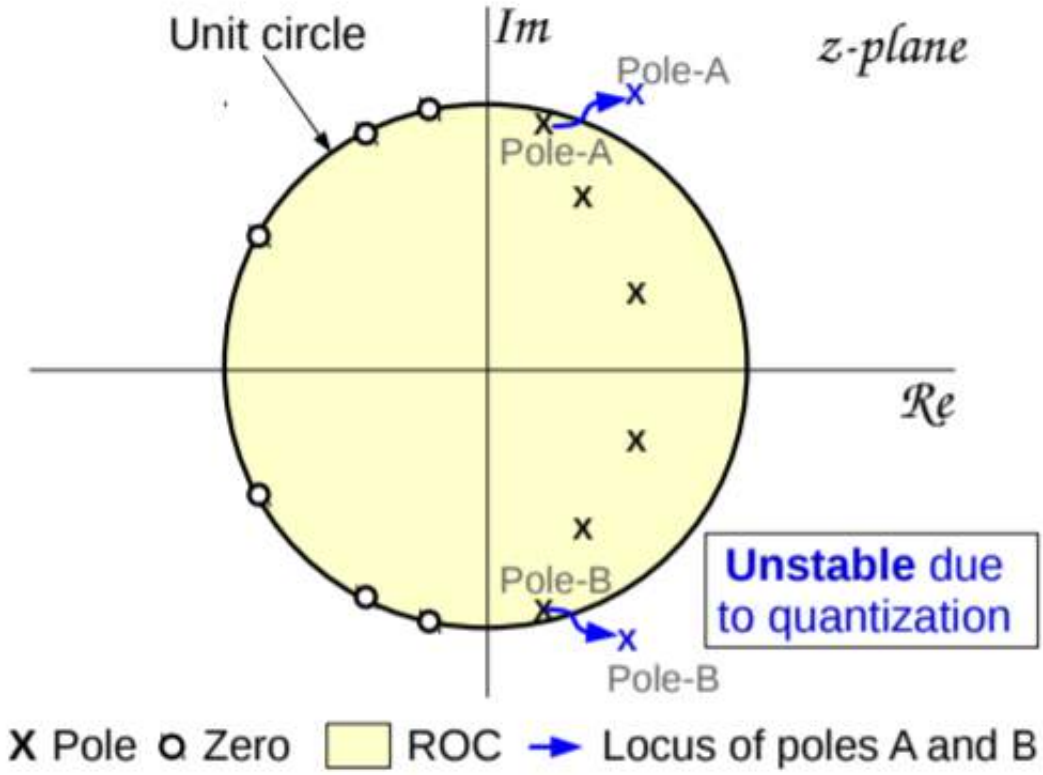

Figure 6.1: Instability due to quantization

6.2 Adaptive FIR Filter Module for Stability

The existing IIR approaches that address space-variant and image-dependent nature of blur exhibits stability issues due to its non-zero poles. Hence, we resort to finite impulse response (FIR) filters as they are inherently stable irrespective of any filter-weight manipulation. Since, FIR is the defacto filter in convolutional neural network (CNN). However, typical CNN filters have fixed receptive fields and filter-weights, and hence are not adequate for dynamic scene blur [Zhou *et al.* (2019)]. We introduce an adaptive filter module which produces spatially varying FIR filters with diverse receptive fields and weights in accordance with the input blurred image pair.

CHAPTER 7

Experiments

In this chapter we sill discuss about the various experiments and ablation studies that were done in evaluating our network. We divide this chapter into four parts,in the first part we will discuss about the data augmentations and training configurations used in the network, in the second section we will discuss about how our network is able to tackle view inconsistency and scene inconsistent disparity and finally we will compare some qualitative results of our network and other deblurring networks.

7.1 Dataset Preparation

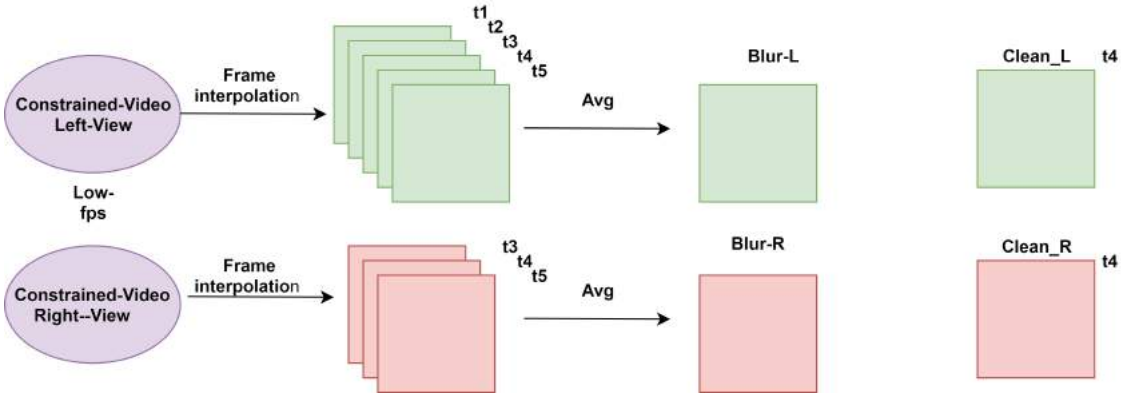

Figure 7.1: Dataset preparation

Till now the only public dataset for stereo deblurring is the Stereo Blur dataset. In this paper we introduce a new synthetic dataset which could be used for unconstrained stereo deblurring.

For this, as shown in Figure 7.1 we take 3 consecutive frames,from the clean images of Stereo Blur Dataset for both left and right image-pairs, the frames are interpolated using the video interpolation network [Niklaus *et al.* (2017)] to generate 3 interpolated frames between two clean frames.

In order to establish the unconstrained scenario, we consider three cases,i.e.(3,5), (1,3), (3,4) consecutive frame pairs which have at-least one overlapping frame for (1,3)

case and two over-lapping frames for(3,4)and(3,5)case between the 9 frames we have, here higher number of frames(more-blur)could be for left or right image, we average consecutive clean frames thus generated to create dynamic scene blur. The clean image pair location for both left and right image is aligned with the middle frame of the lower-blur case with vertical disparity close to zero.

By this method, we are able to generate blurred pairs which have a disparity along both x-y directions, and have different exposure times and the clean image-image pair retain the stereo constraint. Further the left image in each case down-sampled and up-sampled by a factor of 2 in order to generate different resolution inputs.The training dataset size is 43,642 pairs, where 17319 is from Stereo Blur Dataset, 9200 (3-5) case, 8806 (3-4) case,8317 (1-3) case and testing dataset size is 8221 pairs, where3318 is from Stereo Blur Dataset, 1614 (3-5) case, 1614(3-4) case, 1675 (1-3) case.

7.2 Training Configuration

Implementation details

Our network is implemented on pytorch 1.1.0 in a server with Intel Xeon CPU and an NVIDIA RTX 2080 TI GPU.We perform evaluations for our network with a modified version of Stereo Blur data set[Zhou *et al.* (2019)] containing 20475 stereo pairs with their ground truth(17158 for training and 3318for evaluation).

Data Augmentation

For increasing diversity in our model, similar to the works[Tao *et al.* (2018), Zhou *et al.* (2019)], we take 256×256 patches, create down-sampled image pairs and perform chromatic transformations(brightness, contrast and saturation sampled uniformly from[0.85,0.15]). To make our network robust to noise variations, we add random Gaussian noise with $\sigma = 0.01$ .Image pairs are also vertically flipped in random with a probability of 0.5 to generate new pairs.

Model Training

For training our model, we use Adam optimizer[Kingma and Ba (2014)] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.For both single image and dual lens deblurring, we set the batch size as 4 and 1 respectively. All the weights are initialized using Xavier initialization. For training we first train our network for single image setting,i.e. without our fusion blocks for 100,000 iterations and then add the dual lens setup. Convergence was observed for our dual lens setup in 400,000 iterations. The learning rate is decayed from 0.001 to 0 with power 0.3 for both single image and dual lens case

7.3 Performance of the coherent fusion module and scale-space approach

Scale adaptiveness

In order to test effectiveness of our scale adaptiveness technique, we perform a comprehensive study by comparing the PSNR values of left right views obtained at different scales for some standard networks

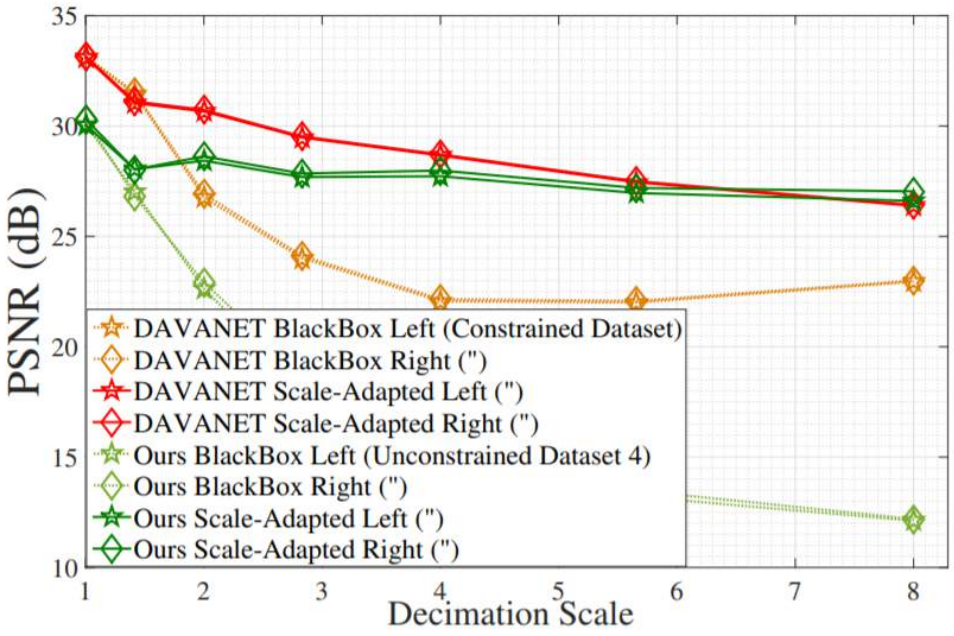

Figure 7.2: Scale-space approach results

In fig.7.2 we compare the effect on PSNR values for 7 different decimation scales for Zhou *et al.* (2019) and our work. As we can see, for both DAVANet and OUR work scale adaptiveness increases the average PSNR for more than 4dB.

Stereo quality

In order to find the subjective left-right consistency of the deblurred image pair we use the SAR evaluation metric from [Chen *et al.* (2013)] on the images obtained after deblurring from our network

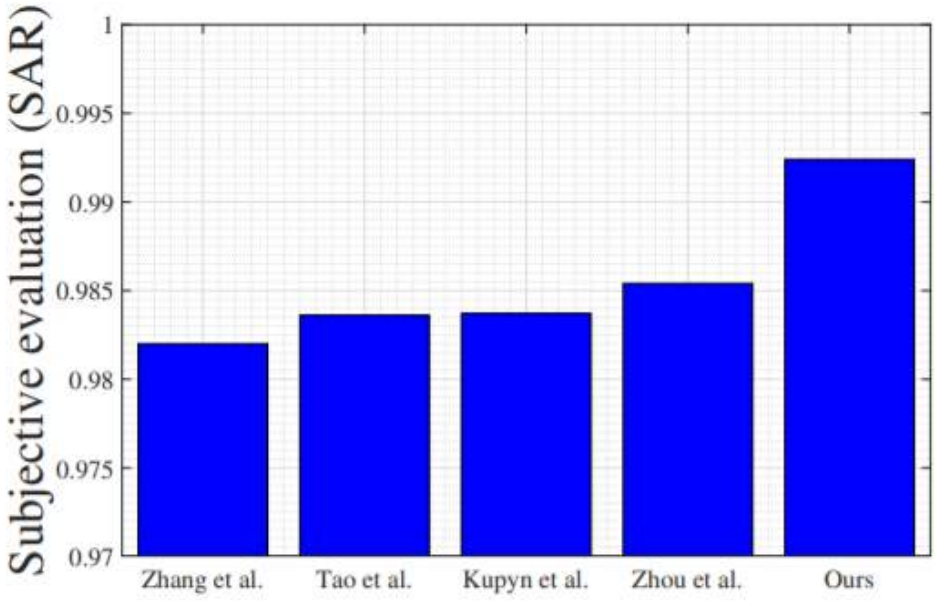

Figure 7.3: Subjective left-right consistency

As we can see in fig 7.3 the SAR values for Our network is much higher compared to other networks for deblurring.

Effect on Super Resolution

As we have mentioned before, scene inconsistent disparities could adversely affect the usage of the deblurred image in other vision tasks. In order to throw more light into this, we downscale the deblurred images obtained from the standard networks and super resolve it using Wang *et al.* (2019)

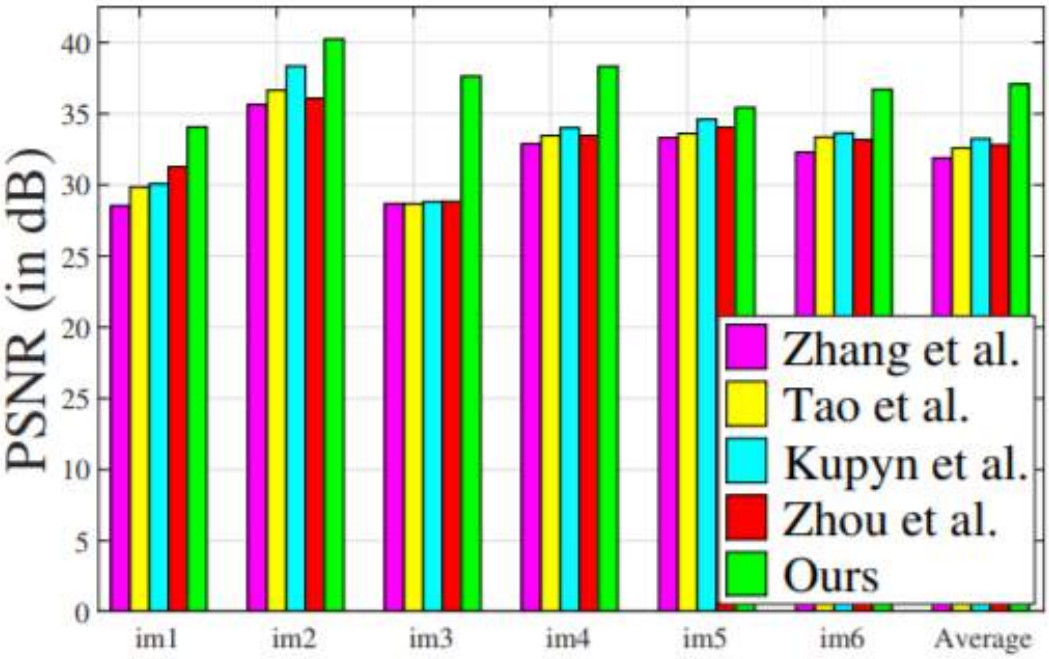

Figure 7.4: Dual Lens super resolution

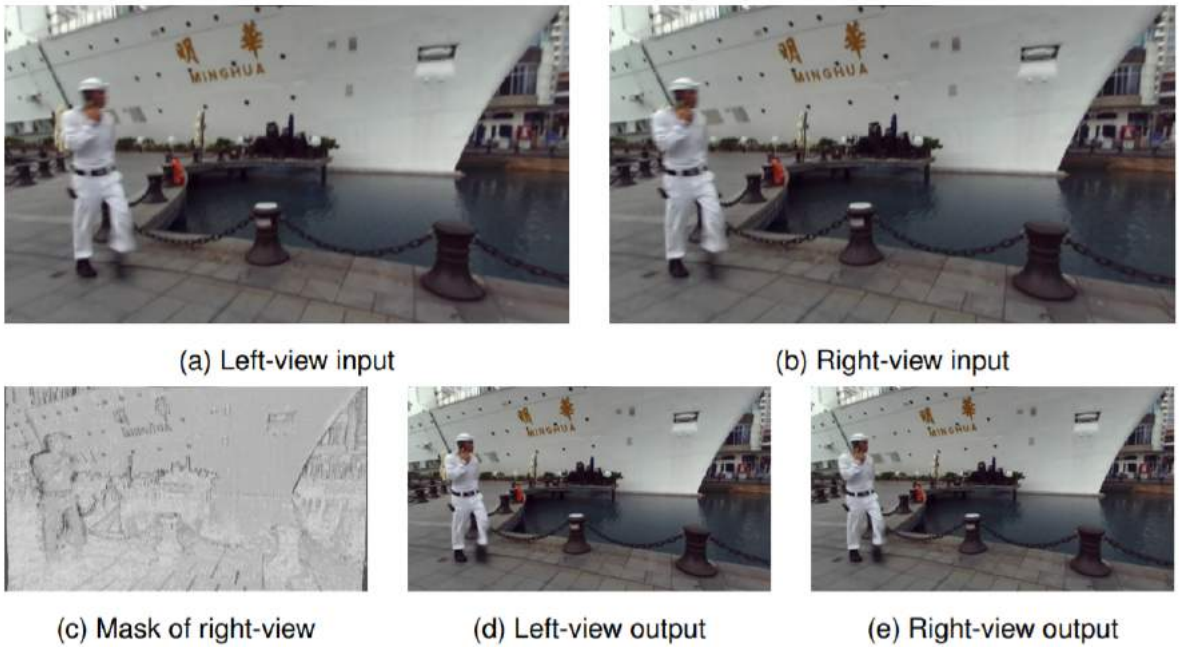

Figure 7.5: Coherent fusion module Visualization

As we can see the PSNR values of the SR images obtained from deblurred images of our network is higher. These is because of the view consistency of the output deblurred image, thus preserving stereoscopic property.

Visualization of Coherent fusion module

Fig. 7.5 shows a visualization of the coherent fusion module used in our network. The mask refers to the occlusion mask and black marks of the mask gives us areas where right view is present by the subsequent left view is not there.

7.4 Quantitative Results

In this section we present some quantitative results and ablation studies, For the below tables, scene-consistent disparities in unconstrained dual lens deblurring can be judged by MAE (lower values are better).

Table 7.1: Table containing different metric value on removing certain stages of our network for Exposure 1:3

Method	Ours	Ours(BS)	Ours(No SA)	Ours(No CF)	Ours (No AF)
MAE	**0.7718**	0.7838	1.7782	0.7846	0.7952
PSNR	**30.132**	30.127	29.852	28.456	29.177
PS:OFFSET	**0.1580**	6.3260	0.2531	6.1211	0.2541
SSIM	**0.915**	0.895	0.908	0.900	0.899
SS:OFFSET	**0.0070**	0.0350	0.0080	0.0710	0.0076

Table 7.2: Table containing different metric value on removing certain stages of our network for Exposure 4:3

Method	Ours	Ours(BS)	Ours(No SA)	Ours(No CF)	Ours (No AF)
MAE	**0.8465**	0.8533	1.97118	0.8572	0.8318
PSNR	**30.581**	30.560	30.118	27.11	28.32
PS:OFFSET	**0.8450**	5.1810	0.8971	5.2181	0.8677
SSIM	**0.917**	0.913	0.915	0.894	0.899
SS:OFFSET	**0.0030**	0.0290	0.0081	0.0581	0.0083

Table 7.3: Table containing different metric value on removing certain stages of our network for Exposure 3:5

Method	Ours	Ours(BS)	Ours(No SA)	Ours(No CF)	Ours (No AF)
MAE	**1.0043**	1.0066	2.2731	1.0076	1.0068
PSNR	**28.801**	28.724	28.402	26.181	27.65
PS:OFFSET	**1.0050**	4.1380	1.1139	3.254	1.0178
SSIM	**0.904**	0.901	0.898	0.885	0.891
SS:OFFSET	**0.0090**	0.0310	0.0131	0.0413	0.0454

Table 7.4: Table containing comparison of different metric value for different standard networks for Exposure 1:3

Method	Mohan	Tao	Zhang	Zhou	Ours
MAE	1.3504	1.8256	1.9312	1.7658	**0.7718**
PSNR	27.724	27.844	27.665	28.102	**30.132**
PS:OFFSET	2.6440	6.3310	5.4890	6.0460	**0.1580**
SSIM	0.888	0.874	0.871	0.890	**0.915**
SS:OFFSET	0.0070	0.0820	0.0770	0.0680	**0.0030**

Table 7.5: Table containing comparison of different metric value for different standard networks for Exposure 4:3

Method	Mohan	Tao	Zhang	Zhou	Ours
MAE	2.273	1.9704	12.0488	1.9328	**0.8465**
PSNR	25.169	26.536	26.406	26.437	**30.581**
PS:OFFSET	1.0600	6.4970	5.6060	5.6360	**0.8450**
SSIM	0.816	0.860	0.858	0.863	**0.917**
SS:OFFSET	0.0130	0.0860	0.0730	0.0740	**0.0070**

Table 7.6: Table containing comparison of different metric value for different standard networks for Exposure 3:5

Method	Mohan	Tao	Zhang	Zhou	Ours
MAE	3.021	2.2524	2.3518	2.2444	**1.0043**
PSNR	26.348	26.593	26.431	26.040	**28.801**
PS:OFFSET	1.9870	4.1550	3.2460	3.3640	**1.0050**
SSIM	0.876	0.862	0.858	0.868	**0.904**
SS:OFFSET	0.0140	0.0570	0.0470	0.0440	**0.0090**

Table 7.7: Table containing comparison of different metric value for different standard networks for Exposure 1:1

Method	Mohan	Tao	Zhang	Zhou	Ours
MAE	1.215	0.8869	0.9921	0.8672	**0.7380**
PSNR	26.815	26.984	25.854	29.198	**32.052**
PS:OFFSET	1.8090	1.2520	1.2960	3.9320	**0.2580**
SSIM	0.854	0.861	0.828	0.892	**0.905**
SS:OFFSET	0.0300	0.0330	0.0330	0.0480	**0.0070**

7.5 Qualitative Results

Some Qualitative results and comparison are shown below:-

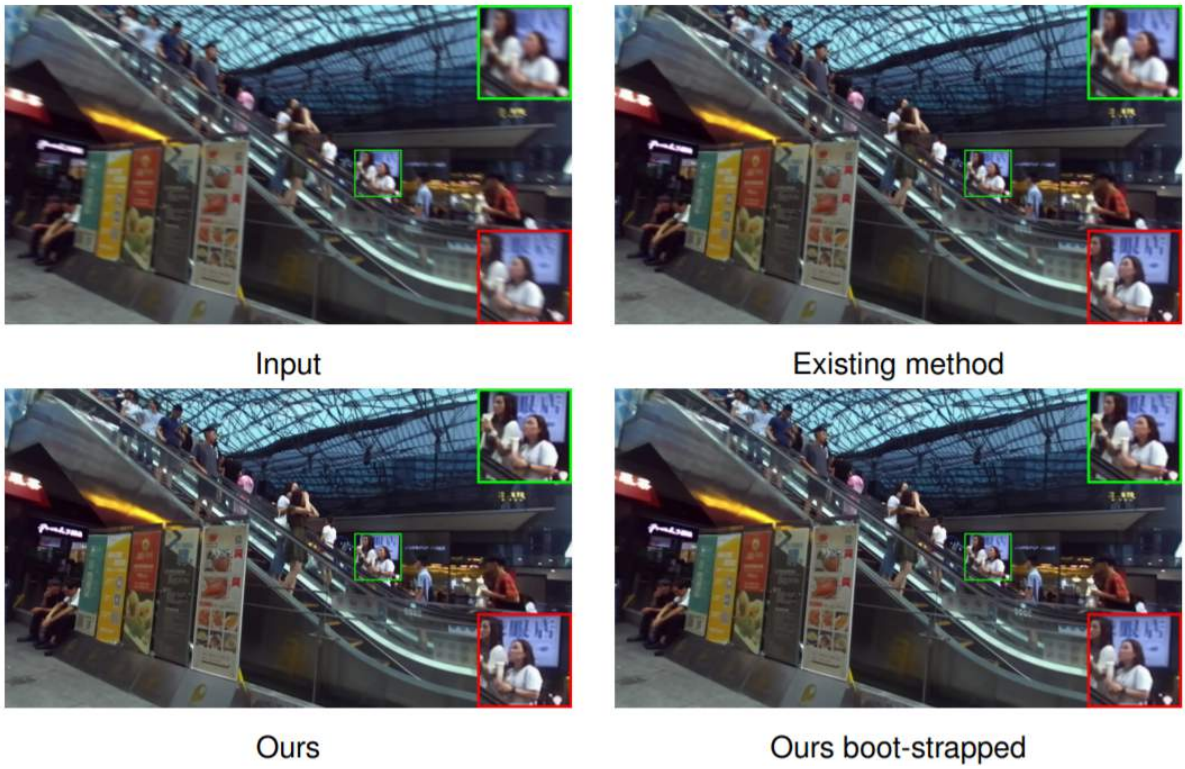

Figure 7.6: Synthetic image: Qualitative:-1

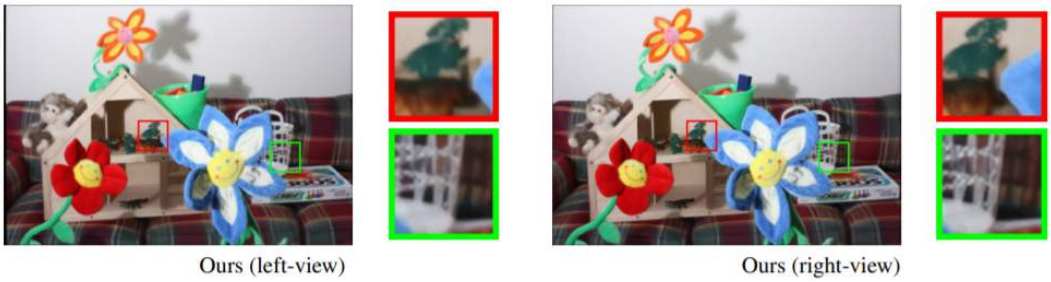

Figure 7.7: Real image: Qualitative:-2

Figure 7.8: Real image: Qualitative:-3

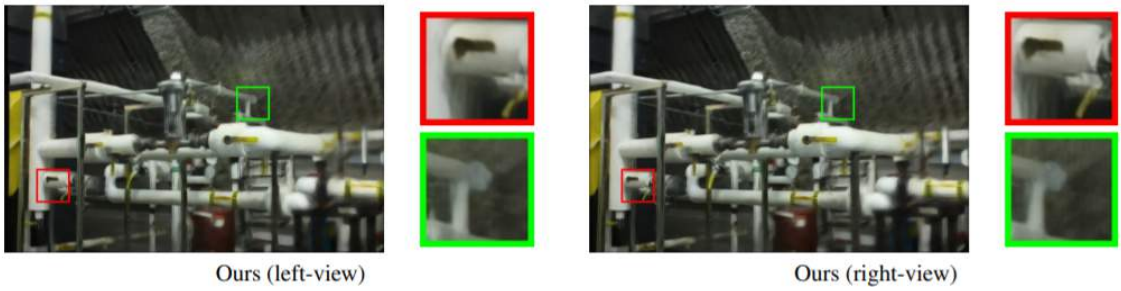

Figure 7.9: Real image: Qualitative:-4

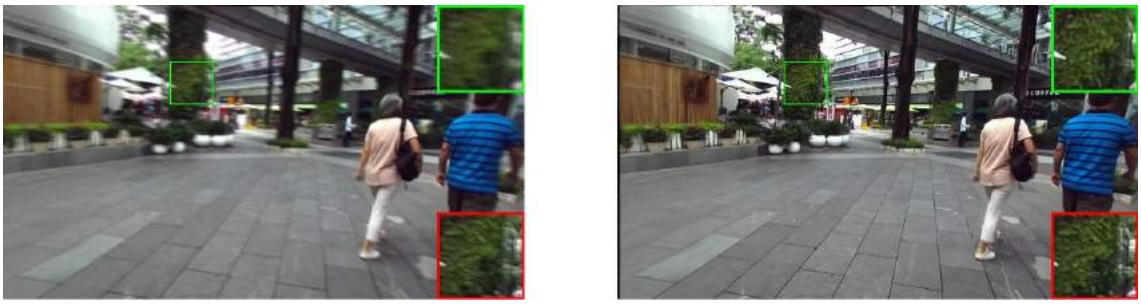

Figure 7.10: Synthetic image: Qualitative:-5

Figure 7.11: Synthetic image: Qualitative:-6

Figure 7.12: Synthetic image: Qualitative:-7

CHAPTER 8

Conclusion

In this work, we introduce a novel network for tackling unconstrained dual lens dynamic scene blurring. The proposed network incorporates an adaptive multi-scale approach to obtain scene-consistent depth in the image pairs. A new image adaptive feature extraction block using dilated convolutions is introduced which has the capability to use receptive fields of different sizes on different images. We also propose a coherent fusion block to to address the problem of view inconsistency.

Using the proposed method, image-pairs having different resolutions and different exposures making them blurred image pairs having significantly different blurs, could be deblurred to obtain equal quality pairs as compared to other networks where the low blur image gives better results.

We also built a new large dataset for unconstrained dual lens deblurring with dynamic scenes using frame interpolation and averaging. The dataset contains left-right views with different resolutions, and three different exposures, as well as the unconstrained scenario.

Comprehensive evaluations with the existing state-of-the-art monocular and dual lens techniques shows the superiority of our network for solving the unconstrained dual lens dynamic scene deblurring problem.

Possible future work to improve the disparity and quality of registration on improving the quality of optical flow values and coming up with more accurate flow estimation networks. Our proposed modules can be easily adapted to future deep learning methods that handle unconstrained dual lens cameras.

We conclude our work by again thanking everyone who have directly or indirecly aided in the completion of this project.

CHAPTER 9

List of Papers to be Submitted based on this thesis

Scale-adaptive Coherent-fusion for Dual-lens Dynamic Scene Deblurring.

Mahesh Mohan M. R., Nithin G. K., and Rajagopalan A. N.

(under preparation for IEEE Journal of Selected Topics in Signal Processing (IJSTSP))

REFERENCES

1. **Arjovsky, M.**, **S. Chintala**, and **L. Bottou** (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

2. **Chen, D.**, **L. Yuan**, **J. Liao**, **N. Yu**, and **G. Hua**, Stereoscopic neural style transfer. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

3. **Chen, L.-C.**, **G. Papandreou**, **I. Kokkinos**, **K. Murphy**, and **A. L. Yuille** (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, **40**(4), 834–848.

4. **Chen, M.-J.**, **C.-C. Su**, **D.-K. Kwon**, **L. K. Cormack**, and **A. C. Bovik** (2013). Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication*, **28**(9), 1143–1155.

5. **Dosovitskiy, A.**, **P. Fischer**, **E. Ilg**, **P. Hausser**, **C. Hazirbas**, **V. Golkov**, **P. Van Der Smagt**, **D. Cremers**, and **T. Brox**, Flownet: Learning optical flow with convolutional networks. *In Proceedings of the IEEE international conference on computer vision*. 2015.

6. **He, K.**, **X. Zhang**, **S. Ren**, and **J. Sun**, Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

7. **Ilg, E.**, **N. Mayer**, **T. Saikia**, **M. Keuper**, **A. Dosovitskiy**, and **T. Brox**, Flownet 2.0: Evolution of optical flow estimation with deep networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

8. **Kingma, D. P.** and **J. Ba** (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

9. **Krizhevsky, A.**, **I. Sutskever**, and **G. E. Hinton**, Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*. 2012.

10. **Kupyn, O.**, **V. Budzan**, **M. Mykhailych**, **D. Mishkin**, and **J. Matas**, Deblurgan: Blind motion deblurring using conditional adversarial networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

11. **Loshchilov, I.** and **F. Hutter** (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

12. **Mayer, N.**, **E. Ilg**, **P. Hausser**, **P. Fischer**, **D. Cremers**, **A. Dosovitskiy**, and **T. Brox**, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

13. **Mohan, M.**, **S. Girish**, and **A. Rajagopalan**, Unconstrained motion deblurring for dual-lens cameras. *In Proceedings of the IEEE International Conference on Computer Vision*. 2019.

14. **Nah, S.**, **T. Hyun Kim**, and **K. Mu Lee**, Deep multi-scale convolutional neural network for dynamic scene deblurring. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

15. **Niklaus, S.**, **L. Mai**, and **F. Liu**, Video frame interpolation via adaptive separable convolution. *In Proceedings of the IEEE International Conference on Computer Vision*. 2017.

16. **Ronneberger, O.**, **P. Fischer**, and **T. Brox**, U-net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

17. **Sim, H.** and **M. Kim**, A deep motion deblurring network based on per-pixel adaptive kernels with residual down-up and up-down modules. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

18. **Simonyan, K.** and **A. Zisserman** (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

19. **Su, S.**, **M. Delbracio**, **J. Wang**, **G. Sapiro**, **W. Heidrich**, and **O. Wang**, Deep video deblurring for hand-held cameras. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

20. **Szegedy, C.**, **W. Liu**, **Y. Jia**, **P. Sermanet**, **S. Reed**, **D. Anguelov**, **D. Erhan**, **V. Vanhoucke**, and **A. Rabinovich**, Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

21. **Tao, X.**, **H. Gao**, **X. Shen**, **J. Wang**, and **J. Jia**, Scale-recurrent network for deep image deblurring. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

22. **Wang, L.**, **Y. Wang**, **Z. Liang**, **Z. Lin**, **J. Yang**, **W. An**, and **Y. Guo**, Learning parallax attention for stereo image super-resolution. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

23. **Whyte, O.**, **J. Sivic**, **A. Zisserman**, and **J. Ponce** (2012). Non-uniform deblurring for shaken images. *International journal of computer vision*, **98**(2), 168–186.

24. **Yu, F.** and **V. Koltun** (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

25. **Zhang, J.**, **J. Pan**, **J. Ren**, **Y. Song**, **L. Bao**, **R. W. Lau**, and **M.-H. Yang**, Dynamic scene deblurring using spatially variant recurrent neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

26. **Zhou, S.**, **J. Zhang**, **W. Zuo**, **H. Xie**, **J. Pan**, and **J. S. Ren**, Davanet: Stereo deblurring with view aggregation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.