Image Deblurring By Using Variational MIMO-UNet

Under Guidance of Prof. Shanmuganathan Raman

Neel Patel - 18110114
Mechanical Engineering
IIT Gandhinagar
Gandhinagar, Gujarat, India.
neel.kirankumar@iitgn.ac.in

ABSTRACT

Coarse-to-fine strategies have been extensively used for the architecture design of single image deblurring networks. The conventional methods of image deblurring task typically uses the stack of subnetworks and it gradually increases the sharpness but the disadvantage is that it increases the computational cost as well. I propose a variational MIMO-UNet model that looks into the modifications of the existing MIMO-UNet model. The MIMO-UNet has three distinct features. First, the single encoder of the MIMO-UNet takes multi-scale input images to ease the difficulty of training. Second, the single decoder of the MIMO-UNet outputs multiple deblurred images with different scales to mimic multi-cascaded U-nets using a single U-shaped network. Last, asymmetric feature fusion is introduced to merge multi-scale features in an efficient manner. Comparing the 3I3O with 4I4O & 2I2O model shows that 4I4O gives a slightly higher psnr loss but with the worst time complexity. Fine tuning of syndata resulted in the increment of psnr loss on the pretrained model.

KEYWORDS

MIMO-UNet, Image Deblurring, Syndata, PSNR.

1. Introduction

The image deblurring task has been an evergreen problem in the field of computer vision because even the latest technology has not been able to capture the blurless image every time. Hence, the image deblurring process plays a crucial role in deblurring the blurred images. This process aims to restore the sharp latent image from the blurred image. The initial stage of the image deblurring task was mainly performed by the convolutional neural networks and its modifications. The results obtained were not satisfactory and a new model architecture with better performance was needed in the field of image processing. The below are some of the renowned network architectures of state-of-the-art models:

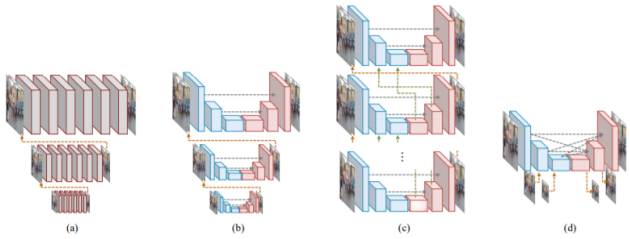


Figure 1. Comparison of different model architectures a) Deep Deblur b) PSS-NSC c) MT-RNN d) MIMO-UNet [1].

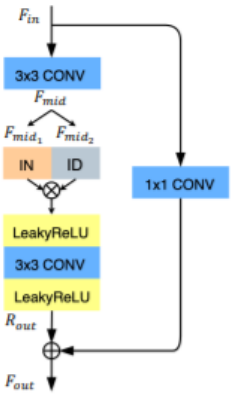
A light-weighted CNN model was also introduced to diminish the con effects of CNN like high computing requirement, time costly but it was not able to meet the expectations of good results [2]. The MIMO-UNet model has low computational complexity and uses a single encoder-decoder based U-Net model. The model has 3 important and distinct features:1. MISE(Multi Input Single Decoder). It takes multiple images and it encodes using a single encoder. 2. MOSD(Multi Output Single Decoder) It outputs multiple deblurred images using a single decoder. 3. AFF(Asymmetric Feature Fusion) It is used to merge multi scale features in an efficient way. All 3 features combinely give state-of-the-art results in image deblurring tasks on datasets like GOPRO and RealBlur [1].

2. Related Works

The past researchers have advanced the image deblurring task performance in both time and space complexity.

2.1 HINet

HINet model is the first positioned amongst all the participants of NTIRE CHALLENGE IMAGE DEBLURRING 2021 in track 2 [3]. This model uses a HIN Block carefully constructed considering the instance normalization and it adopts normalization directly with state-of-the-art performance. A network architecture made up of HIN blocks called as HINet model was made to deblur the blurry images. The HIN Block's is as follows:



where,
IN = Instance Normalization,
ID = Identity
+ = Add
x = Concat

Figure 2. HIN Block [3]

2.2 DeepDeblur

DeepDeblur is another well-known image deblurring model which has maintained its dignity for years in the field of image deblurring. DeepDeblur uses the multiple stacks of sub networks as shown in Figure 1(a). Each subnetwork is made up of convolutional networks in order to store the latent input feature maps [4]. A deblurred image is constructed from the blurred image as follows:

$$\hat{S}_n = \mathcal{H}^{\mathrm{D}}_{\theta^{\mathrm{D}}_n}(B_n; (\hat{S}_{n+1})^{\uparrow}) + B_n,$$

where $\mathcal{H}^{\mathrm{D}}_{\theta^{\mathrm{D}}_n}$ is the n^{th} sub-network of DeepDeblur parameterized by θ^{D}_n. B_n and $\hat{S}_n$ are blurry and deblurred images at the n^{th} scale, respectively, and $\uparrow$ denotes the up-sampling operation.

3. Dataset

GOPRO and RealBlur are some of the renowned datasets in the field of image deblurring. But I used my own dataset which I named it syndata(synthetic dataset). I made a new dataset because I discovered that the top models like HINet and MIMO-UNet are not able to perform well on the plain/simple dataset. Hence, a need to fine tune on a simple dataset was there. I made a new dataset in which I used simple shapes like Circle, Square, Rectangle, Triangle etc. I changed the orientation of the shapes, their colors and their position. I altered the background color as well and hence made a new dataset. I used all the colors of the rainbow to make the dataset exhaustive. I also changed the size of these shapes/objects. Hence, a try of making a simple but comprehensive dataset was made. Training on these dataset can improve the model's performance on basic/simple images. The below images are part of the dataset which I have made:

Sharp Images:

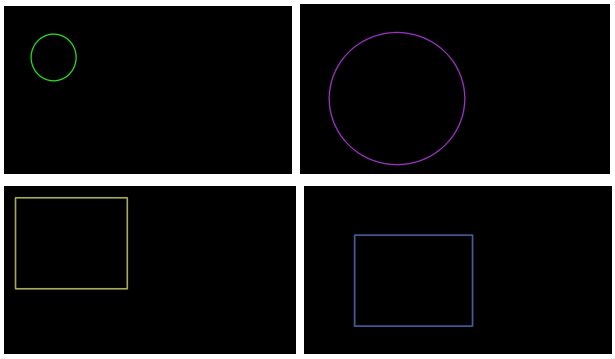


Figure 3: Sharp images(Target Images) of syndata

Blur Images:

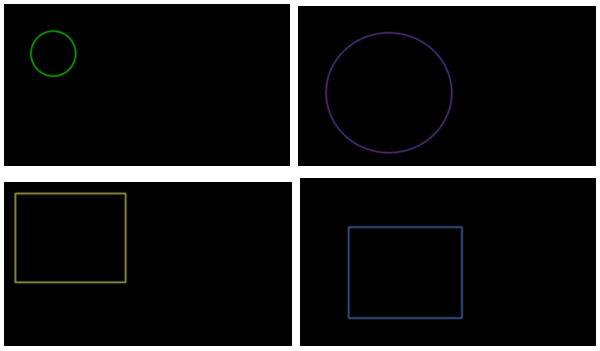


Figure 4: Blur images(Target Images) of syndata

Hence, a training was done with these dataset on MIMO-UNet model and its variations. The blur images were made by applying a gaussian filter of kernel size 5 and they were taken as input images. The sharp images were considered as the target images. A total of 960 images were made.

4. Method

The MIMO-UNet model as described above has 3 major parts MISE, MOSD and AFF. We will study in detail about the 3 major paradigms in order to understand how MIMO-UNet gives a state-of-the-art performance in image deblurring.

MISE: Multiple input single encoder is the part in which multiple images are encoded using a single encoder. The image 1 passes through the 3 Encoder blocks (to get EB_1) and it is then convoluted with the shallow convolutional module (SCM_1) output of image 2 to get the EB_2. Similarly the EB_3 is obtained by EB_2 and SCM_2. Generation of EBs from the input images is the task done by MISE.

MISO: The MISO part's task is to generate the deblurred image from the AFF and Decoder Block(DB). The image reconstructed at each level is as follows:

$$\hat{S}_n = \begin{cases} o(\mathrm{DB}_n(\mathrm{AFF}_n^{\mathrm{out}}; \mathrm{DB}_{n+1}^{\mathrm{out}})) + B_n, & n = 1, 2, \\ o(\mathrm{DB}_n(\mathrm{EB}_n^{\mathrm{out}})) + B_n, & n = 3, \end{cases}$$

where $\mathrm{AFF}_n^{\mathrm{out}}$, $\mathrm{EB}_n^{\mathrm{out}}$, and $\mathrm{DB}_n^{\mathrm{out}}$ are the outputs of the n^{th} level asymmetric feature fusion (AFF) module, EB, and DB, respectively.

AFF: AFF is the model part which is used to merge multiple features at different stages: The formula used to obtain the AFFs are as follows:

$$\begin{aligned} \mathrm{AFF}_1^{\mathrm{out}} &= \mathrm{AFF}_1\left(\mathrm{EB}_1^{\mathrm{out}}, (\mathrm{EB}_2^{\mathrm{out}})^{\uparrow}, (\mathrm{EB}_3^{\mathrm{out}})^{\uparrow}\right) \\ \mathrm{AFF}_2^{\mathrm{out}} &= \mathrm{AFF}_2\left((\mathrm{EB}_1^{\mathrm{out}})^{\downarrow}, \mathrm{EB}_2^{\mathrm{out}}, (\mathrm{EB}_3^{\mathrm{out}})^{\uparrow}\right) \end{aligned},$$

where $\mathrm{AFF}_n^{\mathrm{out}}$ represents the outputs of the n^{th} AFF.

Loss Function:

Loss Function used during the training process is a combination of two loss functions: 1) Content Loss 2) Multi Scale Frequency Reconstruction (MSFR) Loss [1].

$$L_{cont} = \sum_{k=1}^{K} \frac{1}{t_k} \parallel \hat{S}_k - S_k \parallel_1,$$

where K is the number of levels. We divide the loss by the number of total elements t_k for normalization.

$$L_{MSFR} = \sum_{k=1}^{K} \frac{1}{t_k} \parallel \mathcal{F}(\hat{S}_k) - \mathcal{F}(S_k) \parallel_1,$$

where $\mathcal{F}$ denotes the fast Fourier transform (FFT) that transfers image signal to the frequency domain. The final loss function for training our network is determined as follows:

$$L_{total} = L_{cont} + \lambda L_{MSFR},$$

where λ =0.1 (obtained experimentally).

4.1 Experiments

On directly deblurring the syndata on MIMO-UNet model, it was not able to deblur correctly, hence a need of fine tuning the syndata on pretrained model was seen. When the syndata was fine tuned on a pre-trained MIMO-UNet model with train dataset size of 800 and test dataset size as 160, it was able to perform well. The time to train was very less as compared to the training on the GOPRO dataset because the syndata has a smaller number of instances than the GOPRO dataset.

The peak signal to noise ratio (psnr) loss obtained after fine tuning and before fine tuning on MIMO-UNet model can be obtained from the following table:

Model	PSNR Loss	
	Before fine tune	**After fine tune**
MIMO-UNet	31.56	46.59
MIMO-UNet plus	32.78	48.23

Table 1. Fine tuned models' psnr loss on syndata.

4.2 Variational MIMO-UNet

Modifying the MIMO-UNet and changing the network architecture may improve the results of deblurring the images. The one major parameter that we can change in order to modify the model architecture is varying the input images and output images. The MIMO-UNet proposed has used three images in the training process. Varying these numbers of images can change the architecture without the concept of the original model.

2 Input 2 Output UNet Model (2I2O Model):

Taking 2 images as input will bring the model architecture to a small scale but the concept of MIMO-UNet model will sustain and hence we can do the image deblurring task. It will follow the architecture of the original model but the only changes are: the input images and output images will be 2 and hence the network connections including the third image will be removed. The variational MIMO-UNet containing 2 images as input and output are as follows:

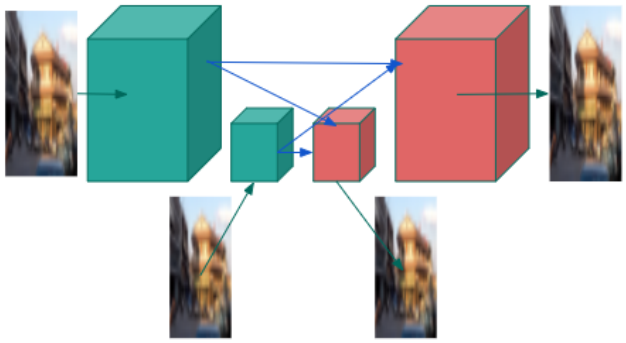


Figure 5. 2I2O UNet Model

4 Input 4 Output UNet Model (4I4O Model):

Another experiment which we can do is increasing the images while training in the network architecture. While doing training with 4 input images and 4 output images, may give us better results but the time complexity and space complexity will also increase. Hence, we have a trade-off between the complexity and results. If the results are significantly higher than the original model then we can

choose this model over the 3I3O-UNet model. The model architecture is of the 4I4O-UNet model is as follows:

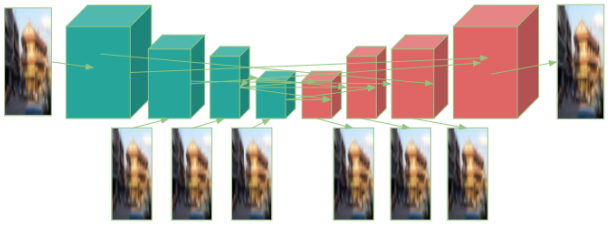


Figure 6. 4I4O UNet Model

5. Results

The training on variational MIMO-UNet model was done on 800 images of syndata and was tested on the remaining 160 images. After training with the syndata on variational MIMO-UNet model, the results of variational MIMO-UNet model obtained are as follows:

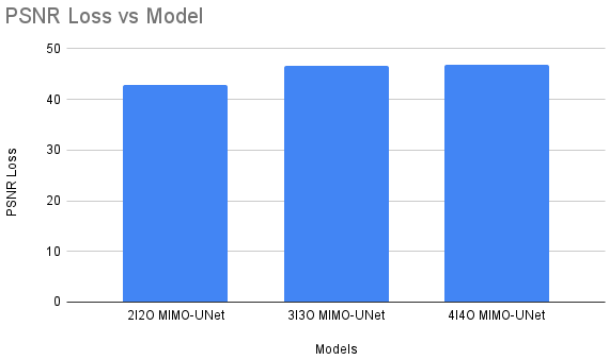


Figure 7. PSNR Loss vs Models

As you can see the 4I4O Model showed best results on syndata but the difference between the average psnr loss between 3I3O model and 4I4O model was not significant. The time taken by the 4I4O model in training was much higher and hence the best model considering the time & space complexity is the 3I3O model. If the psnr loss value is only the concern then we should go with the 4I4O model.

The reason behind such results can be the more detailed feature fusion in the 4I4O model as compared to other 2 models.

6. Conclusion

In this paper, I have experimented with the variations of the MIMO-UNet model. On complete analysis, I observed that the 4I4O is best in terms of results but the time & space complexity suggests that 3I3O is the best model. But a detailed analysis of the models on some famous datasets like GOPRO, RealBlur, etc. is needed.

ACKNOWLEDGMENTS

I am highly thankful to Prof. Shanmuganathan Raman for guiding me throughout the project course.

REFERENCES

[1] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, “Rethinking Coarse-to-Fine Approach in Single Image Deblurring,” *ArXiv210805054 Cs*, Sep. 2021, Accessed: Nov. 23, 2021. [Online]. Available: http://arxiv.org/abs/2108.05054

[2] Y. Yuan, W. Su, and D. Ma, “Efficient Dynamic Scene Deblurring Using Spatially Variant Deconvolution Network With Optical Flow Guided Training,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3552–3561. doi: 10.1109/CVPR42600.2020.00361.

[3] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, “HINet: Half Instance Normalization Network for Image Restoration,” *ArXiv210506086 Cs Eess*, May 2021, Accessed: Nov. 23, 2021. [Online]. Available: http://arxiv.org/abs/2105.06086

[4] S. Nah, T. H. Kim, and K. M. Lee, “Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring,” *ArXiv161202177 Cs*, May 2018, Accessed: Nov. 23, 2021. [Online]. Available: http://arxiv.org/abs/1612.02177