Deep Learning based Multi Frame Super Resolution

Advait Athreya, Akshay Mundra, Pankhuri Vanjani

Universität des Saarlandes, Saarbrücken 66123, Germany

Abstract. Recent trends in smartphones have shown remarkable improvements in their imaging quality. While some of it is due to improved optical capabilities, post-processing pipelines such as image super-resolution (SR) also play a critical part in getting high quality images. Various Single-Image Super-Resolution (SISR) approaches have been designed to address this task, but all of them are limited in their ability to reliably reconstruct the lost high-frequency details. A promising approach to handle this is Multi-Frame Super-Resolution (MFSR), which can leverage the natural tremors from hand-held photography to have a more reliable reconstruction. We take an MFSR model trained on satellite images, and extend it to this use case by performing transfer learning. Our results show significant improvements over SISR based approach such as SR-GAN, both quantitatively and qualitatively.

Keywords: Multi-frame Super resolution · Deep Learning · Residual Attention Networks · 3D Convolution

1 Introduction

Super-resolution (SR) algorithms reconstruct high-resolution (HR) images from either single or multiple low-resolution (LR) images. Since getting HR images from high-end sensors isn't always feasible or economical, the SR algorithms provides a viable opportunity to enhance and reconstruct HR images from LR images recorded by the sensors. Many approaches have been proposed in the literature to address the super resolution problem [14]. However, most of them focus on SISR, where the algorithm takes a single low-resolution input, and produces a corresponding high-resolution output. Since the finer details of the scene are already lost in the creation of the input, such approaches are limited to learning image priors in order to add high frequency details. In contrast, MFSR offers the possibility of reconstructing rich details by combining signal information from multiple images. This results in a more reliable super resolution.

In the following paragraph, we discuss the idea behind MFSR. The basic premise for increasing the spatial resolution in SR techniques is the availability of multiple LR images captured from the same scene. In MFSR, typically, the LR images represent different "looks" at the same scene. That is, LR images are sub sampled (aliased) as well as shifted with sub-pixel precision. If the LR images are shifted by integer units, then each image contains the same information, and thus there is no new information that can be used to reconstruct an HR image. If the LR images have different sub-pixel shifts from each other and if aliasing is present, however, then each image cannot be obtained from the others. In this case, the new information contained in each LR image can be exploited to obtain an HR image. To obtain different looks at the same scene, some relative scene motions must exist from frame to frame via multiple scenes or video sequences. Multiple scenes can be obtained from one camera with several captures or from multiple cameras located in different positions. These scene motions can occur due to the controlled motions in imaging systems, e.g., images acquired from orbiting satellites. The same is true of uncontrolled motions, e.g., movement of local objects or vibrating imaging systems. If these scene motions are known or can be estimated within sub-pixel accuracy and if we combine these LR images, SR image reconstruction is possible as illustrated in Fig. 1.

MFSR finds its application in problems like remote sensing and medical imaging where an accurate reconstruction of high resolution signal is critical, as well as in everyday use-cases like mobile photography. The 'Proba-V Super Resolution Challenge' [2] presents such a MFSR problem for the

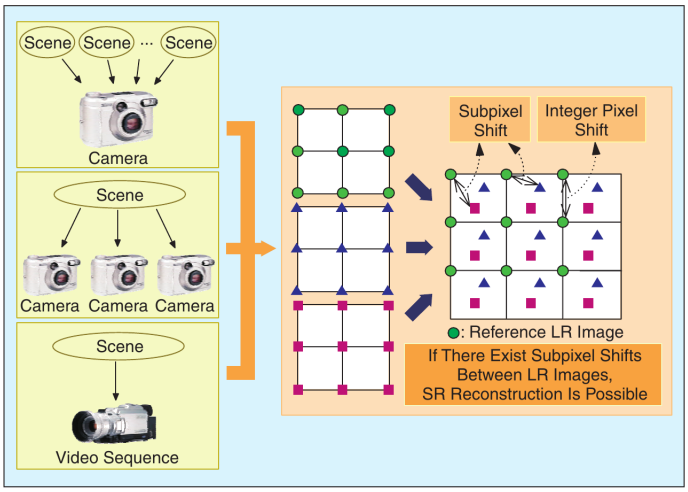


Fig. 1: Premises for super-resolution reconstruction. Author: S.C. Park et al. [14]

satellite images captured by the European Space Agency. Many deep learning based methods have been proposed for it which perform exceptionally well. However, very few work extends it to mobile photography. Thus, we intend to take a model trained on satellite images, and perform transfer learning such that it works for mobile phone photography. This is motivated by the fact that, compared to the DSLR cameras, smartphone cameras have smaller sensors,limiting their spatial resolution; smaller apertures, limiting their light gathering ability; and smaller pixels, reducing their signal-to-noise ratio. However, the natural hand tremor - typical in handheld photography - can be harnessed to acquire a burst of frames with small offsets. These frames can then be merged to form a single large resolution image.

Our work follows the following flow:

1. Generation of a synthetic multi-frame dataset whose images resemble mobile photographs.
2. Training a deep learning architecture over the generated dataset.
3. Performance comparison of the trained network with the baselines.
4. Ablation study to understand the contribution of various modules of the network.

2 Related Work

In the recent works, research has moved from classical SISR to exploring deep learning based techniques for SISR [18]. This was demonstrated with SRCNN in [4] where a CNN model was trained end to end to map between low and high resolution images. Following this work, [5] came up with FSRCNN, which improves on the speed of the network, thus making it suitable for real time performance. In SRCNN [4] authors had also shown sparse-coding methods analogous to a CNN based architecture where SRCNN was outperforming those. But this inspired the researchers and in some of the works like [17] sparse coding was integrated in deep learning architecture to give an efficient as well as accurate results.

With the growing popularity of Generative Adverserial Networks(GANs) in Deep Learning for image restoration, researchers started exploring various GAN based models for super resolution. [11] introduced SR-GAN for this purpose which gave better results than other existing methods in SISR.

From SISR, the research has been extended to MFSR and even though few classical methods have been able to give good results, application of Deep Learning techniques in MFSR has leveraged the performance. Some of them have taken inspiration from SISR utilizing their algorithms and techniques.

[7] used multi frame CNN (MFCNN) architecture which was utilized from single frame CNN model and they were able to improve the results from SISR.

In MFSR, most of the research work has been for remote-sensing on Proba-V dataset. In last few years, some of the teams started experimenting with residual networks for this Proba-V Challenge and they were able to get stay on top of leaderboards. In [12] Enhanced Deep Residual Networks(EDSR) was introduced which emerged winner for NTIRE Super resolution challenge in 2017. Following this work [19] designed Wide Activation Super Resolution (WDSR) which took won the 3 tracks in super resolution in NTIRE 2018. These good results motivated the teams to experiment WDSR with multi frame super resolution too. Apart from residual networks, from different submissions it was also found out 3D convolution layers improved the results by taking took advantage of both small variations in frame positions as well as temporal dimension. RAMS model [15] which emerged as one the best performing model in this challenge was built by improving the work of [6] which used used 3D CNN residual architecture with WDSR blocks from [10].

However, this architecture was experimented with remote-sensing dataset. Researching through these works and the specific blocks in model architecture, they seem to have good potential in extending to other problems too. In our work we propose extending the RAMS model to a mobile photography based dataset 'Holopix' and experiments have been done to analyze the performance and scope of improvement.

3 Methodology

We propose the use of RAMS(Residual Feature Attention Deep Neural Networks)[15] model for super resolution of mobile photography images using multiple image frames. In the PROBA-V SR Challenge this has emerged as the best performing model as per the leader board statistics.

The problem can be formulated as:

$$I^{SR} = H_{RAMS} * (I^{LR}_{1,T}, \theta)$$

I^{SR} : super resolved image
$I^{LR}_{1,T}$: Set of T Low Resolution Images
θ : Model Parameters Learned

Initially, we have used the pre-trained model (with Proba-V dataset) for our dataset, and in later stages we have trained the model on our dataset to update the weights and analyze the performance. We have done an ablation study to understand the affect of each module in the final results. Finally, we have ćompared our results with the state of the art super resolution techniques working for single image frame super resolution.

3.1 Network Architecture

RAMS model uses nested residual blocks for extracting high-frequency components since, those are more important in super resolution task. While the attention mechanism allows keeping focus on important features which were extracted during fusion of multiple image frames.

The model consists of 3 main components:

1. **Global residual branch: Residual Temporal Attention Block**

This module provides a starting image output from a stack of multiple images and does the data fusion from them to give a super resolved image.

2. **Residual Feature Attention Block**

With 3D convolutions, the model takes advantage of both spatial and temporal correlations in the stack of multiple low resolution images. The feature attention mechanism feature in this block along with 3D convolutions allows focusing on high-frequency components and letting the low-frequency components flow through the model to the output.

3. **Main branch: Long skip connections and Temporal Reduction Blocks**

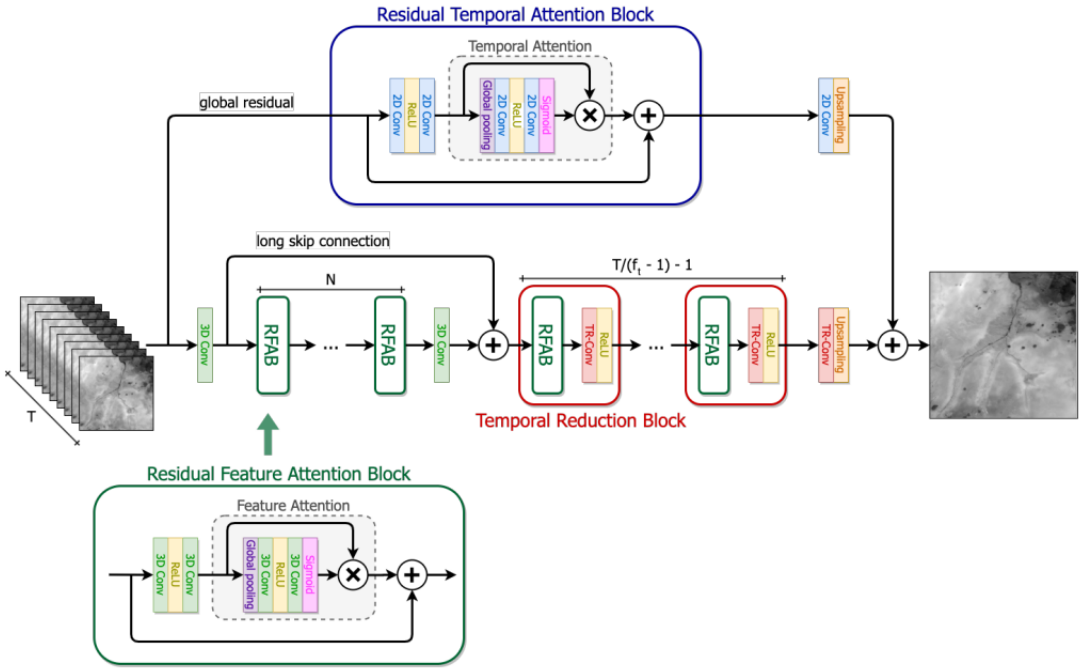


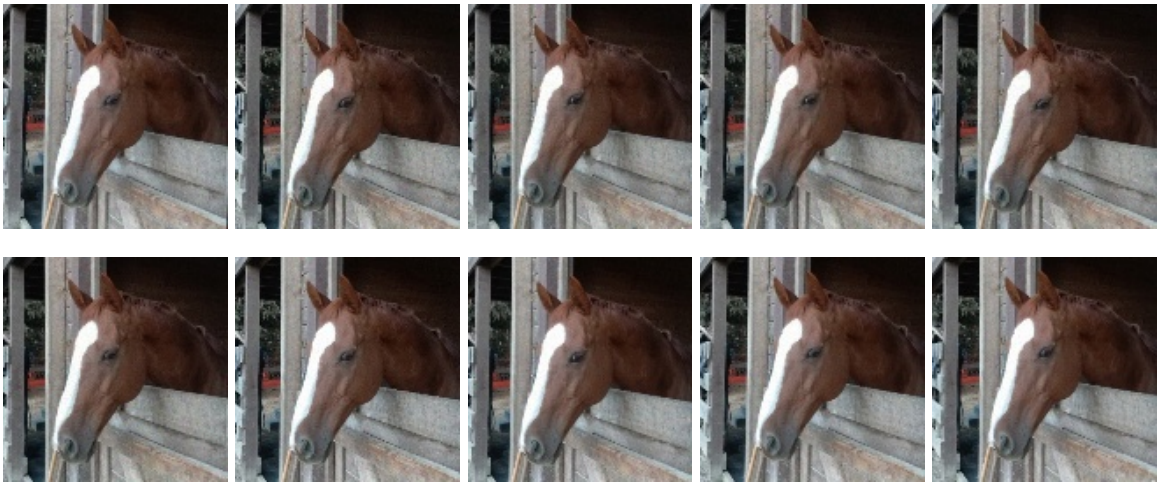
Fig. 2: RAMS model. Author: Salvetti et al. [15]

This branch has 3d convolution operations as well as temporal reduction blocks, they provide a refinement to the output obtained from the global residual branch. Since, there were T images initially in the input of the model, temporal reduction blocks are responsible for reducing the dimension so that we obtain a final single super resolved image.

3.2 Dataset

We have generated synthetic datasets using the script from BURSTSR (Burst Image Super-Resolution Challenge, NTIRE, which is part of CVPR 2021 Workshop and challenge) [3]. This generates multiple images from Holopix Dataset [9] with different lighting and coloring conditions as well as frame of view. This multi-frame dataset has been be used for our experiments. Holopix Dataset has stereo image pairs captured with mobile phone. Different versions of our dataset have been created for different experiments to approach the problem systematically. Specifically, 2 sets of grayscale datasets 'Grayscale2' [Fig. 4] and 'Grayscale3' [Fig. 5] have been generated. It should be noted that the original paper operates on single channel images, thus starting our experiments with grayscale images keeps it tractable. To create both of them, we randomly crop a patch out of the original image, downsample by a factor of 3, and add noise. Grayscale2 also has additional geometric transformation applied to each of the LR images. Similarly, an RGB dataset [Fig. 3] is generated by applying the same downsampling, noise and geometric transformations. Each of the stereo pair images are used for generating 5 LR images (thus giving a total of 10 LR images per scene), and the original left-view image is considered ground truth. There are 4000 16-bit images in the generated dataset with Low resolution images of 128*128 dimension and High resolution images of 384*384.

(a) HR ground truth image



(b) LR burst images

Fig. 3: RGB synthetic burst data generated with BURSTSR (image visualisation not to scale).

3.3 Training Process

For the experiments done in this work, Nvidia Tesla P100 PCIE GPU with 16 GB GPU Memory has been used. For learning the H_{RAMS} function the network needs to estimate the θ model parameter appropriately, for this the loss function between the I^{SR}_{crop} (reconstructed super resolved image's cropped part with d pixels on sides of border) and I^{HR} (ground truth High resolution image's patch) has been minimized.

Loss function: The loss function used in this paper is L1 loss (Minimum Mean Absolute Error). After researching the original RAMS model's paper and relevant research work on loss functions for super resolution tasks it was concluded L1 loss function is suitable for working on experiments. This is due to the fact that L1 loss has been shown to give better results for image restoration problem experimentally. Unlike L2 loss it doesn't penalize the difference harshly(which has been shown to increase smoothing in the final image). The loss function is mathematically expressed as:

$$\mathcal{L} = \min_{u,v \in [0,2d]} \frac{\left\| I^{\mathrm{HR}}_{u,v} - \left(I^{\mathrm{SR}}_{u,v} + b_{u,v} \right) \right\|_1}{(sH - 2d)(sW - 2d)}$$

Here, (sH - 2d)×(sW - 2d) is the size patch is taken from ground truth where u,v $\in [0, 2d]$

(a) HR ground truth image

(b) LR burst images

Fig. 4: Garyscale2 synthetic burst data generated with BURSTSR (image visualisation not to scale).

Relevant parameters which were determined with experiments are summarized in Table 1

4 Experiments and Evaluation

In this section, we discuss the various training experiments, as well as an ablation study to understand the contribution of each network-path to the final result. We make use of the open source code[1] provided by [15].

Two quantitative metrics will be used for evaluating the performance: cPSNR and cSSIM. PSNR (Corrected peak signal-to-noise ratio) is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. SSIM (Corrected structural similarity index) measures the perceptual similarity between two images. A detailed comparison of the two can be found in [8]. cPSNR and cSSIM handle the possible shift in the pixels in low resolution and predicted super resolved image by taking shift correction, which is further explained in [13]. cPSNR

[1] https://github.com/EscVM/RAMS

(a) HR ground truth image

(b) LR burst images

Fig. 5: Grayscale3 synthetic burst data generated with BURSTSR (image visualisation not to scale).

and cSSIM are mathematically represented as,

$$\text{cPSNR} = 10 \log_{10} \frac{\left(2^{16}-1\right)^2}{\text{cMSE}}$$

$$\text{cSSIM} = \max_{u,v \in [0,6]} \text{SSIM}\left(I^{\text{HR}}_{u,v} \cdot M^{\text{HR}}_{u,v}, I^{\text{SR}}_{\text{crop}} \cdot M^{\text{HR}}_{u,v} + b_{u,v}\right)$$

where, cMSE (corrected Mean Square Error) is related with Mean square error (MSE) as:

$$\text{cMSE} = \min_{u,v \in [0,6]} \text{MSE}\left(I^{\text{HR}}_{u,v}, I^{\text{SR}}_{\text{crop}} + b_{u,v}\right)$$

$$\text{MSE} = \frac{\left\| I^{\text{HR}}_{u,v} \cdot M^{\text{HR}}_{u,v} - \left(I^{\text{SR}}_{u,v} \cdot M^{\text{HR}}_{u,v} + b_{u,v}\right) \right\|_2^2}{\left\| M^{\text{HR}}_{u,v} \right\|_1}$$

Here, I^{SR}_{crop} is reconstructed super resolved image, I^{HR} is ground truth High resolution image , M^{HR} is binary mask and their patches have been taken in formula equations. Bias in brightness is denoted with $b_{u,v}$.

Table 1: Training Parameter Values

Parameter	**Value**
Batch size	192
Epochs	70
Learning rate	1e-4
Filters (feature maps)	32
number of residual feature attention blocks	12
number of temporal steps	9
attention compression	8

The final results are compared against two baseline approaches: Bicubic interpolation [1] and SR-GAN based interpolation [11].

Grayscale2 data on Proba-V RAMS model: We evaluate Grayscale2 test images on the network trained on Prova-V dataset. Fig. 6 compares the output of the network with bicubic interpolation and ground-truth. We see that the generated image has handled the aliasing in the input well, but lacks sharpness. This is because the network assumes that the input images have been aligned in the pre-processing stage, which is not the case in this dataset. Thus, there is some merging of information by the network.

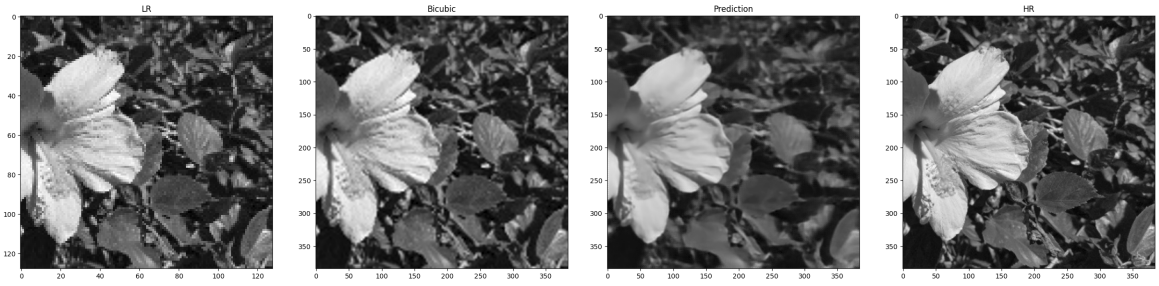


Fig. 6: Testing Proba-V RAMS model on a Grayscale2 image. Left to right: low resolution input (LR), bicubic interpolation of LR, prediction by the network, ground truth high resolution image.

Grayscale3 data on Proba-V RAMS model: To verify the performance of the network for well-aligned inputs, we run the above experiment on Grayscale3 test images. Fig. 7 compares the output. Though still not perfect, the network is able to remove the noise, while also keeping the output relatively sharper when the inputs are better aligned. This is promising enough to train a model on Grayscale3 dataset from scratch.

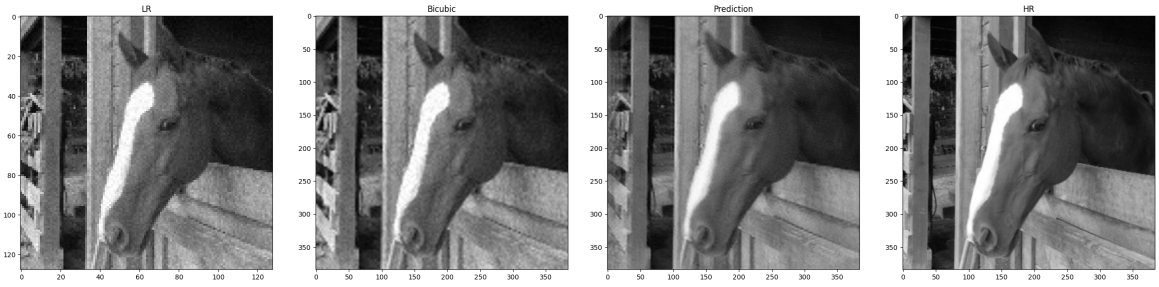


Fig. 7: Testing Proba-V RAMS model on a Grayscale3 image. Left to right: low resolution input (LR), bicubic interpolation of LR, prediction by the network, ground truth high resolution image.

RAMS model trained from scratch on Grayscale3 data We train the RAMS model on Grayscale3 training data for 70 epochs, whose output can be seen in Fig. 8. As evident, the prediction outperforms bicubic interpolation, and is perceptually very close to the ground-truth. Later, we also compare the results quantitatively.

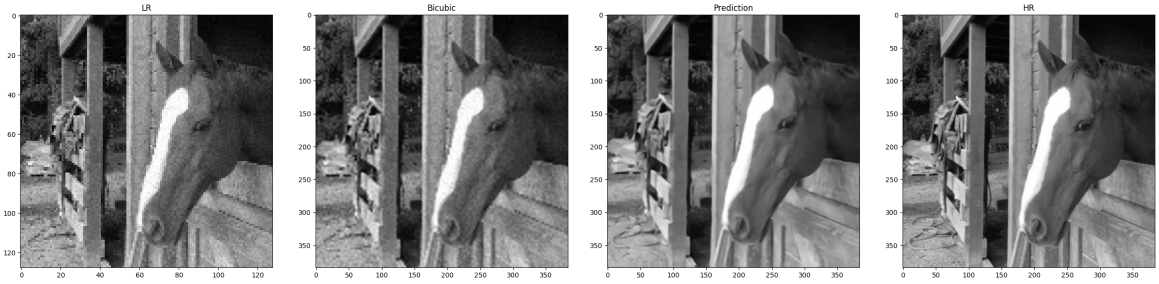
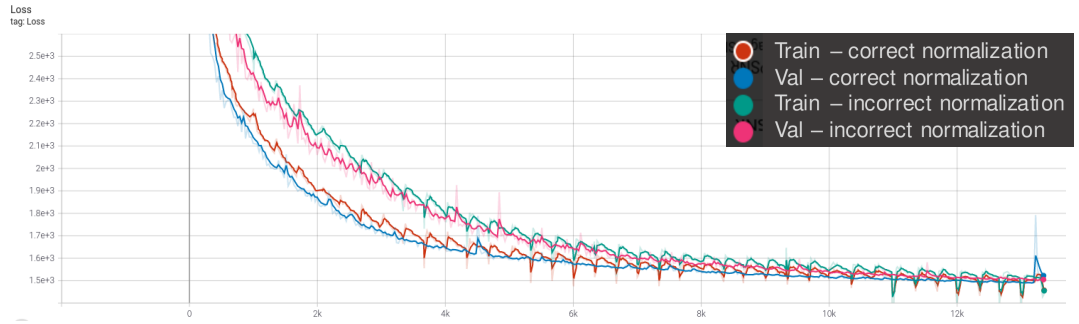
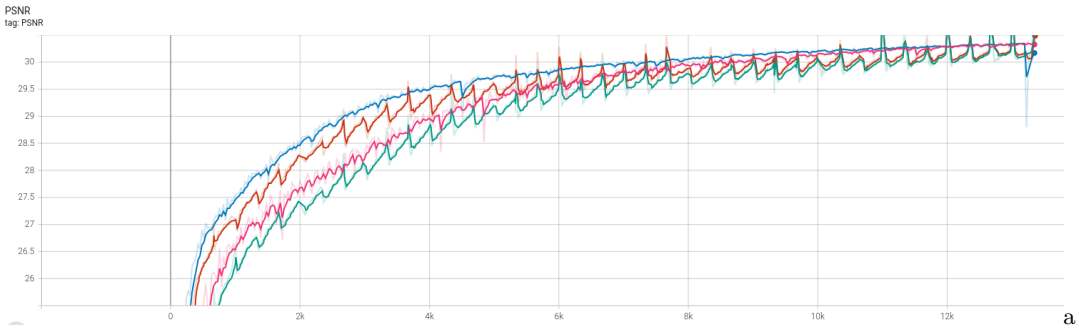


Fig. 8: Testing our RAMS model on a Grayscale3 image. Left to right: low resolution input (LR), bicubic interpolation of LR, prediction by the network, ground truth high resolution image.

Training with the correct normalisation parameters While running the above experiment, it was realised that the authors have hard-coded the normalisation parameters corresponding to the Proba-V images. Since this could lead to sub-par training, the parameters were corrected for our dataset and the experiment were re-ran. Fig. 9 compare the training plots. It clearly demonstrates that the correct data normalisation leads to a faster convergence of the training, which also corroborates the theory behind data normalisation.



(a) Loss curves.



(b) PSNR curves.

Fig. 9: Training and validation curves with different normalization parameters.

Comparison with baselines: Table 2 shows the comparison of PSNR and SSIM metrics with the baselines. Note that both Bicubic interpolation and SR-GAN based interpolation perform single frame super-resolution. While the former uses a simple mathematical model to do so, the latter is a deep learning based approach. We also compare our results with the original RAMS model trained on Proba-V dataset, to highlight the quantative improvement achieved with transfer learning. We show that our model outperforms the baselines by a significant margin on Grayscale3 dataset.

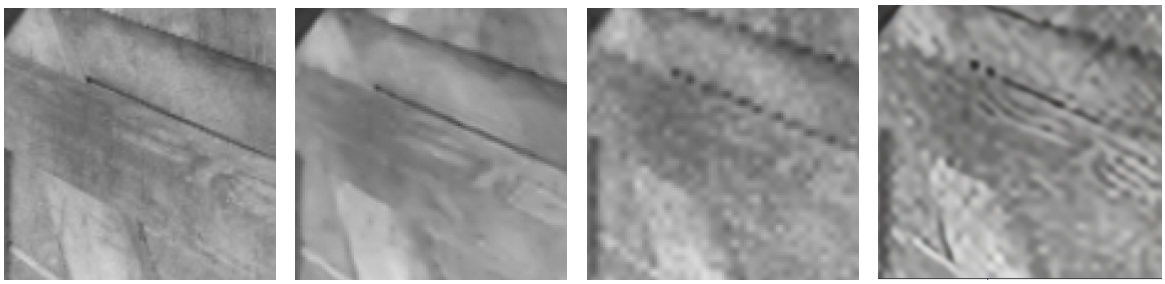
Table 2: Table captions should be placed above the tables.

Scope	Approach	PSNR	SSIM
Single frame	Bicubic Interpolation	19.76	0.43
Single frame	SR-GAN	28.91	0.79
Multi frame	RAMS + Pretrained weights	23.02	0.64
Multi frame	RAMS retrained	**29.20**	**0.85**

A qualitative comparison between the baselines is shown in Fig. 10 and 11. The results clearly show that our method gives cleaner and sharper results as compared to the baselines, and is very close to the ground truth. It is also worth mentioning that SR-GAN does include more high-frequency details as compared to Bicubic interpolation, but most of the details don't resemble the ground truth, but is instead the network trivially injecting high-frequencies in the result. Compared to this, we show that out network can learn to combine images to reconstruct reliable high frequencies.

(a) Ground truth (b) Ours (c) Bicubic (d) SR-GAN

Fig. 10: Qualitative comparison with the baselines.



(a) Ground truth (b) Ours (c) Bicubic (d) SR-GAN

Fig. 11: Qualitative comparison with the baselines on a zoomed-in patch.

Ablation Study In this section, we evaluate the contribution of each of the two network modules - global residual path and central path - to the final output. To that end, we pass a test Grayscale2 and Grayscale3 image separately through each module, and then through the entire network. The results are visualised in Fig. 12. We see that the global residual path produces grid-like artefacts in both cases, whereas, the central path produces over-smoothed results. When we combines the two, we get improvements in both these aspects.

5 Conclusion and Future Work

This work has explored the deep learning techniques for multi frame super resolution which leverages the additional information from different frames of an image scene. Currently, most of the applications in the multi frame resolution is towards remote-sensing based satellite images, but with the introduction of NTIRE challenges, researchers have found significant use cases with the mobile photography too. In this work, RAMS model which comprises of residual networks and 3D convolutions has been used for multi frame burst images generated from mobile photography based Holopix dataset. This network architecture has given promising results both qualitatively and quantitatively while comparing with baselines for the grayscale3 dataset.

For other datasets with pixel misalignments, the generated results lack sharpness. After researching through the problem, we realised the the network is not equipped to handle the inherent misalignment in the image frames. To deal with this, we can either use dense optical flow to warp the input images during pre-processing stage, or have dedicated stereo misalignment handling blocks in the architecture. For example, for the NTIRE MFSR challenge, teams have used PCD (Pyramid, Cascading and Deformable module) [16], Feature Enhanced PCD module[modified PCD], deformable convolutions or

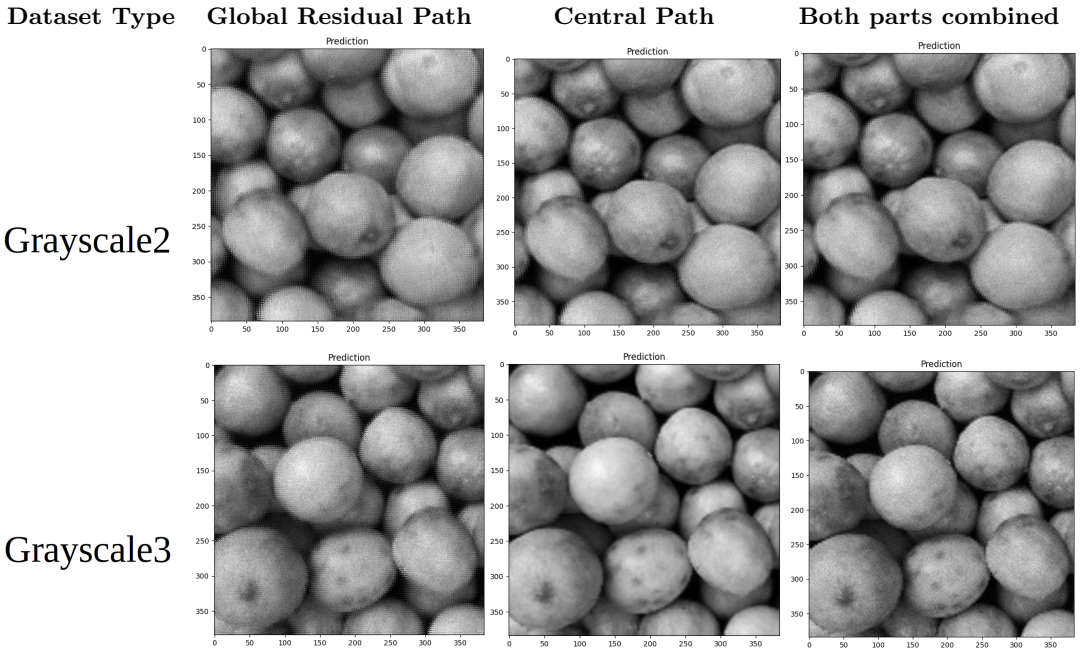


Fig. 12: Qualitative results from ablation study of the model architecture

Kernel Prediction Networks. For the future work, we propose incorporation of stereo feature alignment module like PCD in the RAMS architecture before the data fusing Residual Temporal Attention Block in the global branch. Currently the work has been done with grayscale images, we plan to extend the work to RGB images with further experiments.

References

1. Bicubic interpolation — Wikipedia, the free encyclopedia, https://en.wikipedia.org/wiki/Bicubic_interpolation, [Online; accessed 31-July-2021]
2. Proba-v super resolution challenge, https://kelvins.esa.int/proba-v-super-resolution/, [Online; accessed 4-Aug-2021]
3. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. arXiv preprint arXiv:2101.10997 (2021)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
5. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. pp. 391–407. Springer (2016)
6. Dorr, F.: Satellite image multi-frame super resolution using 3d wide-activation neural networks. Remote Sensing **12**(22), 3812 (2020)
7. Greaves, A., Winter, H.: Multi-frame video super-resolution using convolutional neural networks (2016)
8. Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition. pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579
9. Hua, Y., Kohli, P., Uplavikar, P., Ravi, A., Gunaseelan, S., Orozco, J., Li, E.: Holopix50k: A large-scale in-the-wild stereo image dataset. In: CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, 2020. (June 2020)
10. Kim, S.Y., Lim, J., Na, T., Kim, M.: 3dsrnet: Video super-resolution using 3d convolutional neural networks. arXiv preprint arXiv:1812.09079 (2018)
11. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)

12. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
13. Märtens, M., Izzo, D., Krzic, A., Cox, D.: Super-resolution of proba-v images using convolutional neural networks. Astrodynamics **3**(4), 387–402 (2019)
14. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine **20**(3), 21–36 (2003). https://doi.org/10.1109/MSP.2003.1203207
15. Salvetti, F., Mazzia, V., Khaliq, A., Chiaberge, M.: Multi-image super resolution of remotely sensed images using residual attention deep neural networks. Remote Sensing **12**(14), 2207 (2020)
16. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
17. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE international conference on computer vision. pp. 370–378 (2015)
18. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. IEEE Transactions on Multimedia **21**(12), 3106–3121 (2019)
19. Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., Huang, T.: Wide activation for efficient and accurate image super-resolution. arXiv preprint arXiv:1808.08718 (2018)