IFT 6145 - Project

Multiple-Image Super-Resolution

Bowen Peng

22/12/2020

0 Abstract

Image super-resolution aims to generate a high resolution image from a single low resolution image (in the case of single image super-resolution, or SISR) or from a sequence of low resolution images (in the case of multiple image super resolution, or MISR). Having multiple images for super-resolution can be beneficial as more detail can be captured, but it is often hard to align the multiple images when there are moving objects. In this project we will use optical flow to correct moving objects for MISR and compare its effectiveness to SISR.

1 Introduction

It is known that information is lost forever when a discrete image is downsized and it would be theoretically impossible to recover the original high resolution from a downsized image as most downscaling operations are non-injective functions (its exact inverse does not exist). However, by exploiting the fact that nearby pixels in an image are usually correlated to each other, knowing that most images contain salient features (and not just random noise) and by tricking the human visual system, we can design super-resolution algorithms that try to find the most accurate or most visually pleasing approximate inverse to the downscaling function.

Furthermore, if we have access to more information, such as multiple low resolution images taken at slightly different positions, we can infer much more information about the high resolution image. However, if the multiple images are taken at different times (which is usually the case if we only have a single camera), objects in movement will appear in different locations, and confuse the SR algorithm.

2 Proposed Method

To solve this problem, we use optical flow to compensate the movement of objects from each frame to the next. This allows us to take a sequence of images from a video, and correct the movement using the motion vectors. The method used in this project is similar to the one used in [2], but we use existing motion vectors to pre-process the images in advance instead of estimating them within our SR model. We then use a convolutional neural network to predict the high resolution image using eight aligned low resolution images.

2.1 Dataset Generation

To generate our custom dataset, we used the SINTEL optical flow dataset. We start by selecting contiguous sequences of 8 high resolution images from SINTEL. For each image except for the last, we iteratively apply the provided motion vectors to each image until all of them are aligned with the last image. As the dataset does not supply backward motion vectors, all forward warping is done using splatting. This introduces additional blur but should not be an issue for the SR network. We then apply different but consistent small translation to each image except the last to simulate small sub-pixel camera movements. The last image of the 8 serves as the ground truth. The eight images are then downsampled with a $\times 4$ factor. This gives us 8 aligned low resolution images and the corresponding ground truth image.

Figure 1: One example from the generated dataset. Occlusions are most noticeable for earlier frames.

2.2 Model Architecture

The proposed network consists of two parts, feature extraction and image reconstruction. Feature extraction is done at the low resolution spatial size, and then enlarged to the high resolution spatial size using a pixel-shuffle layer[4]. These enlarged features are then passed through the reconstruction network, which outputs the high resolution image. The last layer of the feature extraction network and the image reconstruction network are densely connected[1] to all previous layers of their respective parts for better performance and gradient flow. These dense connections do not add a significant amount of parameters. All intermediate convolutional layers layers use 8 kernels of size 3×3, with CReLU as activation. CReLU has been shown to improve gradient flow and feature reuse[3]. The last layer of the feature extraction network uses 128 kernels and the last layer of the image reconstruction network uses 1 kernel. Zero padding is used to keep all feature maps at a constant spatial size. An overview of the MISR (multiple image super resolution) network architecture is shown below. Our final network has $m = 3$ and $n = 3$.

We also created a SISR (single image super resolution) network for performance comparison. Everything is identical except that the input features consists only of one low resolution image.

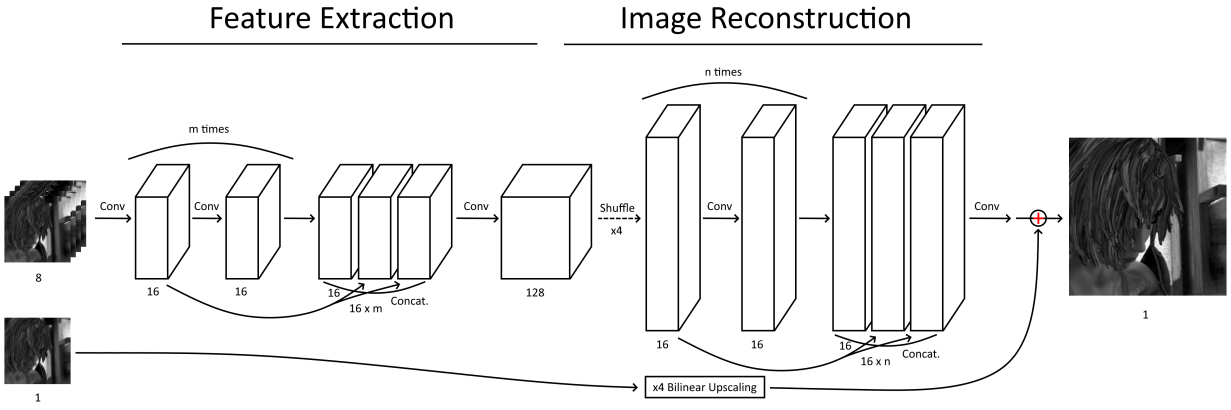


Figure 2: Proposed x4 MISR network architecture.

2.3 Training

For the objective function, we minimize the mean squared error $\frac{1}{2}||y - f(x)||^2$ over the training set. Training is carried out using mini-batch gradient descent (mini-batches of 16) with backpropagation. We used the Adam optimizer, with a learning rate $\alpha = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Weight decay and gradient clipping was not used.

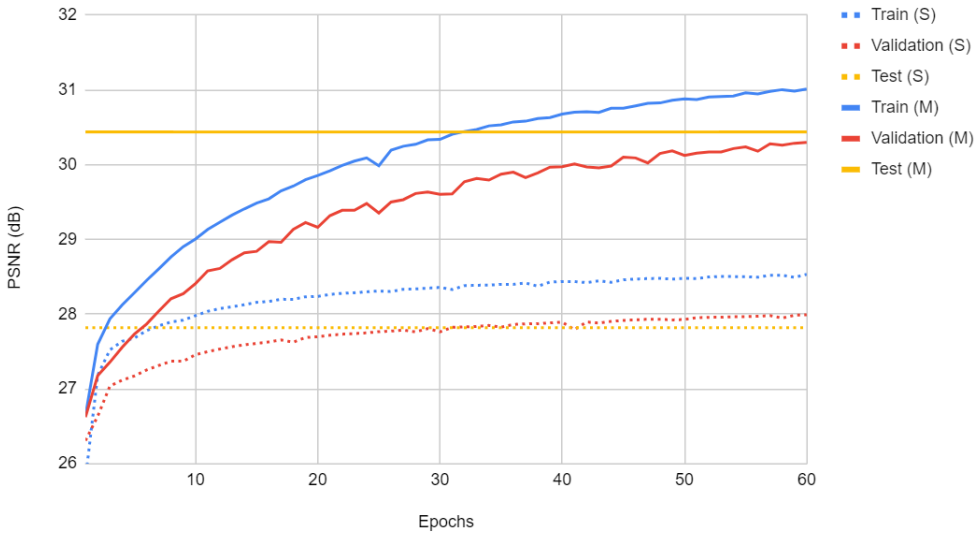


Figure 3: x4 SR PSNR during and after training. M represents multiple images and S means single image. The test result was computed at the end of training using the test set.

3 Analysis

Our MISR model performs well compared to a similar model which only a single image is given as input. The model recovers much more details and has a much higher PSNR compared to bicubic and the SISR model. However, objects that suffer from large amounts of occlusions are not super-resolved as well as objects that are visible across many frames.

Additionally, we have to note that this MISR model has access to perfect motion vector data from SINTEL, which might not be the case in real life datasets. Due to a lack of time, we have not explored the effects of using a lower quality optical flow estimation, but we expect the results to be degraded significantly as the alignment of the images will not be sufficient enough unless corrected by another optical flow estimation network. But this rather simple method still has uses for some applications. For instance, in computer graphics, where we can compute and have perfect optical flow information.

A deeper end-to-end network will perform better at the task of super-resolution, up to a limit, as the quantity of information is limited within a quantized low resolution picture. Our network is rather small, and it performs better by increasing the number of layers and features per layer, but due to the limited amount of memory available on Google Colab, we could not increase our network size any further.

4 Experimental Results

Using input images of size 256×109, the neural network was trained for 60 epochs, with mini-batches of 16. The dataset used contained 880 pairs of potential training instances, split between training (780), validation (50) and test (50) sets. The training set was augmented by flipping the images across the x-axis, giving us a total of 1560 training instances.

The MISR network used has a total of 17.6k parameters, and the SISR network used has a total of 17.1k parameters. The difference comes only from the first convolutional layer, and we expect the internal expressivity of both networks to be similar. The final PSNR obtained on the test set is 30.44 dB for the MISR model, and 27.82 dB for the SISR model. Figure 3 confirms the improvement of PSNR over time during training, and no overfitting is visible as the validation loss does not increase. Visual inspection of the output confirms that the network successfully learned to perform super-resolution on low resolution images, as shown in Figure 4. The total training time on Google Colab using a NVIDIA Tesla P4 GPU was 1 hours and 11 minutes for a single model.

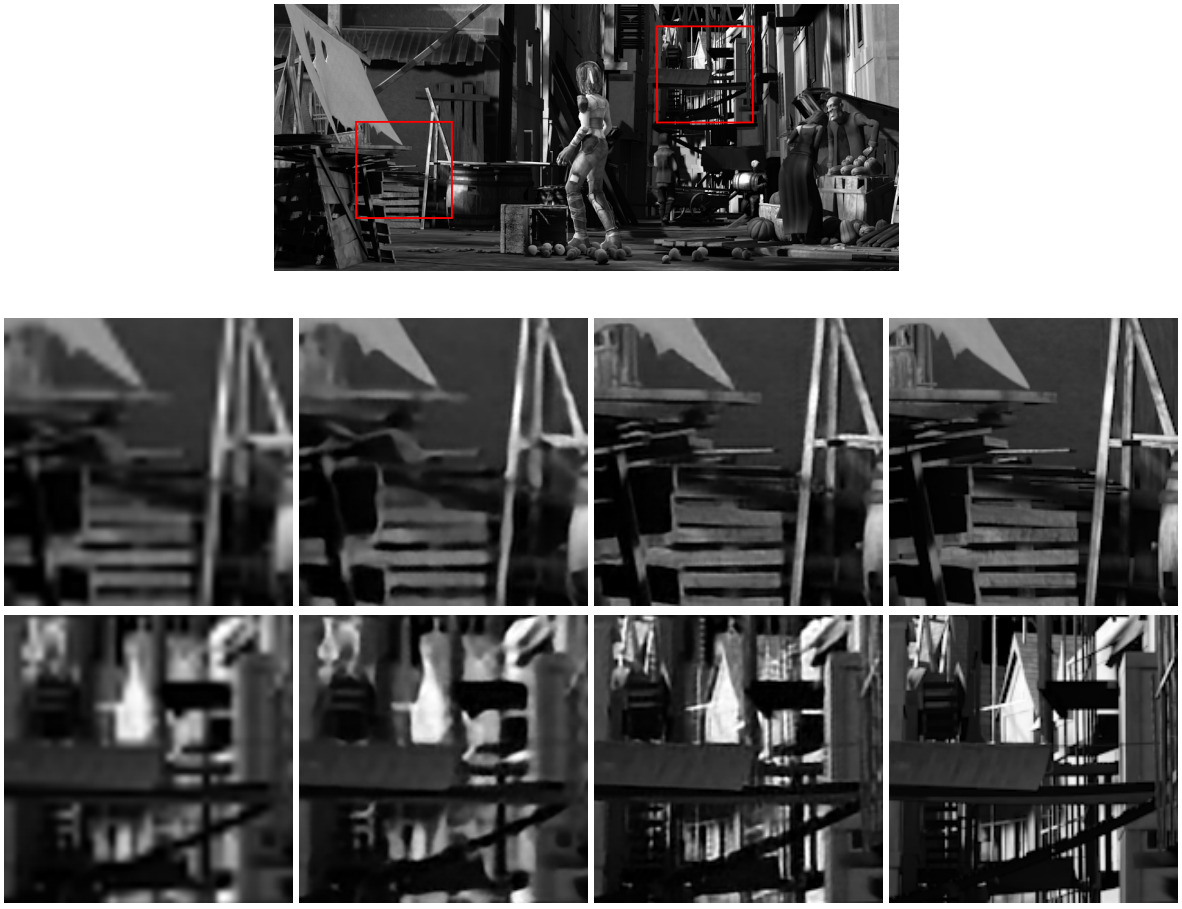


Figure 4: x4 SR result on an image from the test set.

4.1 Failure cases

Objects that are moving a lot cause a large amount of self-occlusions and occlusions to other objects, which significantly degrades the SR result on the affected objects. The degraded result is often not better than the result of a SISR model and is sometimes accompanied by artifacts.

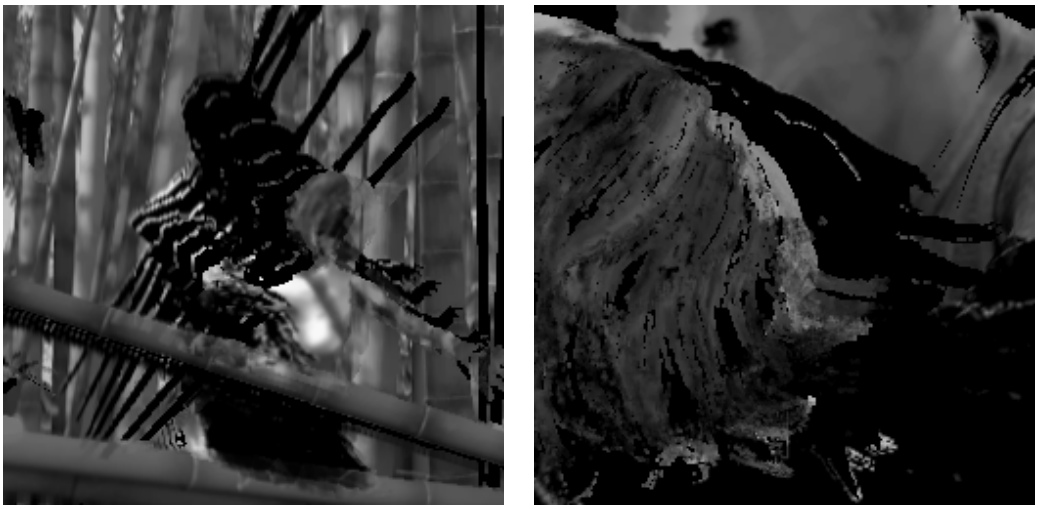


Figure 5: Examples of large amounts of occlusions and self-occlusions on fast moving objects.

LR SISR MISR Ground Truth

Figure 6: Failure case due to occlusion. Note the blurriness and artifacts on and near the moving body on the MISR model. The foliage further away to the right is not affected.

4.2 Test Time Ablation

To explore whether the network is truly using all 8 input features for its prediction and to explore the robustness of the network to input degradation, we set some input features to 0 before passing it to the predictor. As seen in the figure below, the MISR network still surpasses the SISR model when only given 2 images as input, but is worse with 1 image. At 0 images, our model is equivalent to bilinear interpolation, which confirms that our model performs better than bilinear even when inputs are highly degraded.

Due to time constraints, we were only able to test ablation of input features during prediction. Future research can focus on the effects of feature ablation during training, or even training models that can accept a variable amounts of input features.

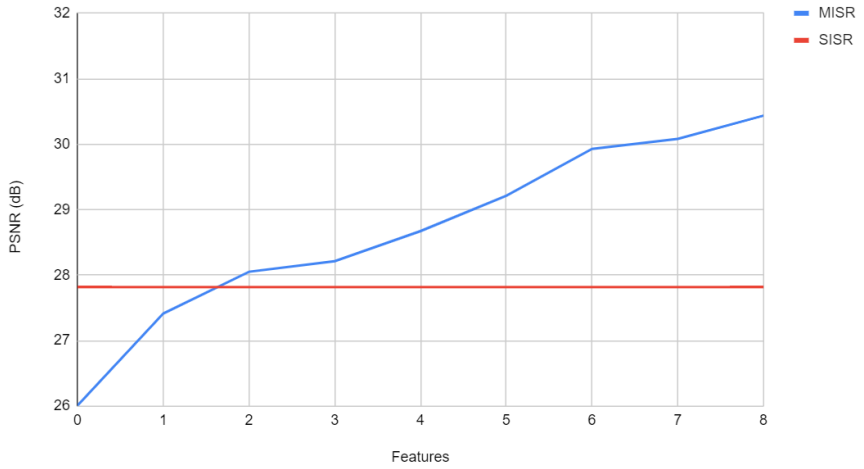
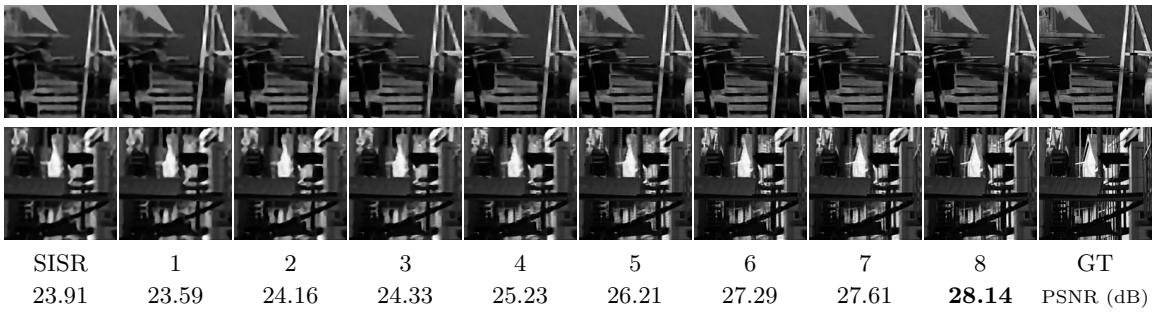


Figure 7: x4 SR PSNR ablation study at test time.



SISR	1	2	3	4	5	6	7	8	GT
23.91	23.59	24.16	24.33	25.23	26.21	27.29	27.61	**28.14**	PSNR (dB)

Figure 8: x4 SR ablation results on an image from the test set.

SISR	3	5	8	GT
23.91	24.33	26.21	**28.14**	PSNR (dB)

Figure 9: Same results as above, magnified for better visualization.

5 Conclusion

In conclusion, we have shown that multiple image super resolution is more efficient than single image super resolution, especially when the network size is small. When multiple low resolution images and high quality optical flow information is available, it is better to use a MISR model.

References

[1] Gao Huang et al. *Densely Connected Convolutional Networks*. IEEE, 2016.

[2] Longguang Wang et al. *Deep Video Super-Resolution using HR Optical Flow Estimation*. IEEE, 2020.

[3] Wenling Shang et al. *Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units*. ICML, 2016.

[4] Wenzhe Shi et al. *Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network*. IEEE, 2016.