iSeeBetter: A Novel Approach to Video Super-Resolution using Adaptive Frame Recurrence and Generative Adversarial Networks

Aman Chadha
System Performance and Architecture
Apple Inc.
amanc@stanford.edu

***Abstract*—Recently, learning-based models have enhanced the performance of Single-Image Super-Resolution (SISR). However, applying SISR successively to each video frame leads to lack of temporal consistency. On the other hand, VSR models based on convolutional neural networks outperform traditional approaches in terms of image quality metrics such as Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). While optimizing mean squared reconstruction error during training improves PSNR and SSIM, these metrics may not capture fine details in the image leading to misrepresentation of perceptual quality. We propose an Adaptive Frame Recurrent Video Super Resolution (AFRVSR) scheme that seeks to improve temporal consistency by utilizing information from multiple similar adjacent frames (both future LR frames and previous SR estimates), in addition to the current frame. Further, to improve the "naturality" of the reconstructed image while eliminating artifacts seen with traditional algorithms, we combine the output of the AFRVSR algorithm with a Super-Resolution Generative Adversarial Network (SRGAN). The proposed idea thus not only considers spatial information in the current frame but also temporal information in the adjacent frames thereby offering superior reconstruction fidelity. Once our implementation is complete, we plan to show results on publicly available datasets that demonstrate that the proposed algorithms surpass current state-of-the-art performance in both accuracy and efficiency.**

Keywords*—*super-resolution, frame recurrence, generative adversarial networks, video upscaling, optical flow, convolutional neural networks.

I. Introduction

Traditional Video Super-Resolution (VSR) methods upscale based on a single degradation model (usually bicubic interpolation), followed by reconstruction. This is sub-optimal and adds computational complexity [1]. Further, the ability of mean square error, which these studies utilize to capture high texture details based on pixel-wise frame differences, is very limited causing the resulting video frames to be too smooth [2].

Recently, learning-based models have enhanced the performance of Single-Image Super-Resolution (SISR). However, applying SISR independently to each video frame leads to lack of temporal consistency. While high-frequency details need to be reconstructed exclusively from spatial statistics in the case of SISR, temporal relationships inherent in videos can be exploited to improve reconstruction for VSR. It is therefore imperative to combine the information from as many low resolution (LR) frames as possible to reach the best video super-resolution results [5].

VSR models based on Convolutional Neural Networks (CNNs) outperform traditional approaches in terms of widely used image reconstruction metrics such as Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). The perceptual image quality of resulting super-resolved image is principally dependent on choice of a loss function, which is optimized during model training. Recent work is largely based on optimizing mean squared reconstruction error. Such loss improves PSNR and SSIM, however, these metrics may not capture fine details in the image leading to misrepresentation of perpetual quality [9].

II. Proposed idea

A. Method

Rather than applying super-resolution to each frame independently, our approach involves utilizing adjacent frame similarity to identify additional frames similar to the input frame that can be fed to the algorithm for reconstruction, along with the input frame. To this end, we train a network that utilizes optical flow-based methods to estimate future frames using similar adjacent frames.

To mitigate the issue of lack of finer texture details when super-resolving at large upscaling factors which is seen with CNNs, our approach utilizes Generative Adversarial Networks (GANs). Per [6], we use adversarial loss along with content loss (which focuses on perceptual similarity instead of similarity in pixel space) to limit model "fantasy", and thus improve the naturality associated with the reconstructed image using a generator-discriminator model [9]. The generator-discriminator architecture pushes the model to generate more realistic and appealing frames while eliminating artifacts seen with traditional algorithms.

The novelty in our approach is that we have proposed an adaptive version of FRVSR, which combines information from multiple adjacent frames (both future LR images and previous SR estimates) along with the current frame, instead of just the last frame and the current frame as proposed in [5]. The proposed idea thus combines the virtues of the Adaptive Frame Recurrent Video Super Resolution (AFRVSR) technology with Super-Resolution Generative Adversarial Network (SRGAN) proposed in [6] to reconstruct high-definition videos with superior temporal consistency and fidelity.

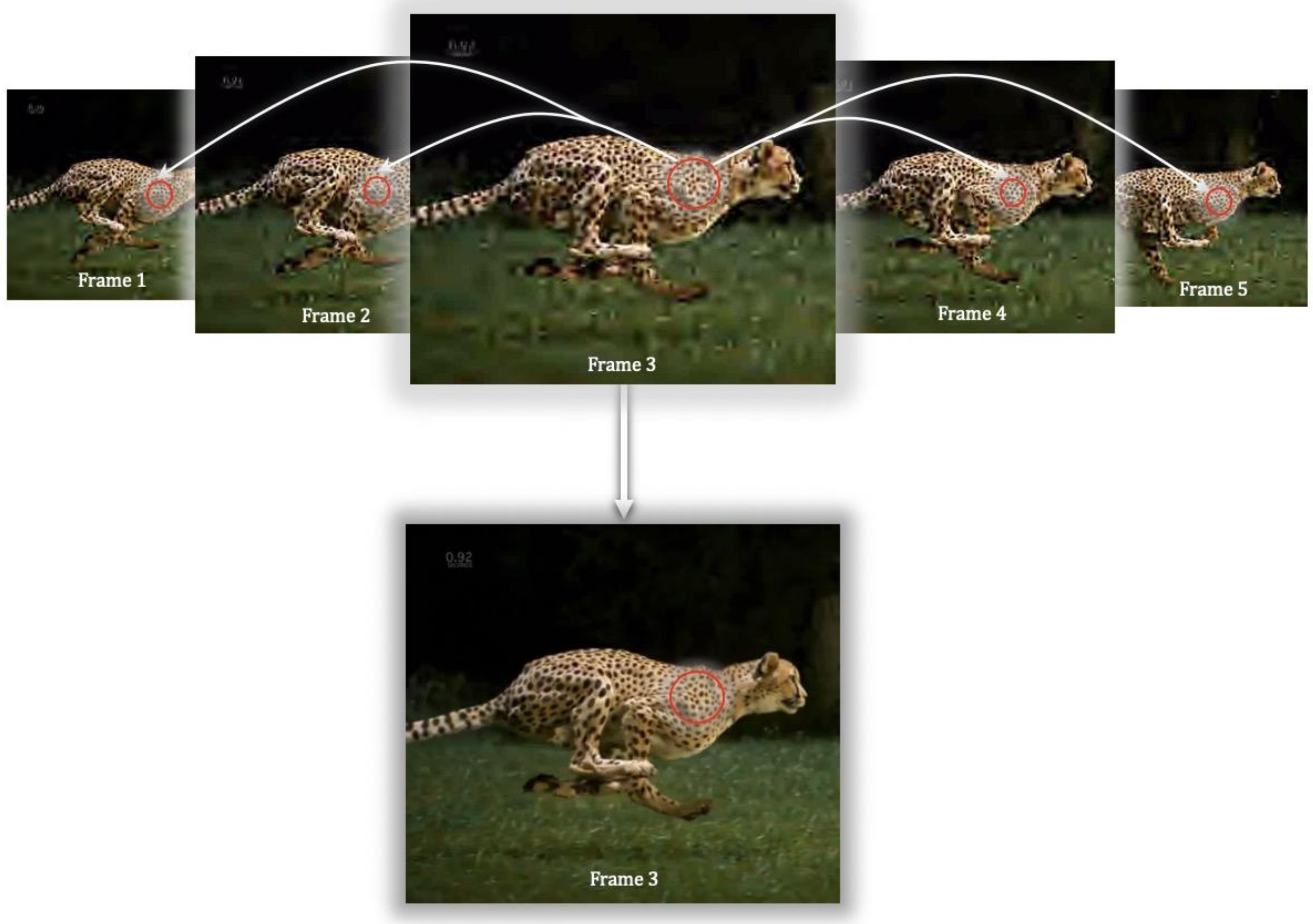


Figure 1: Adjacent frame similarity

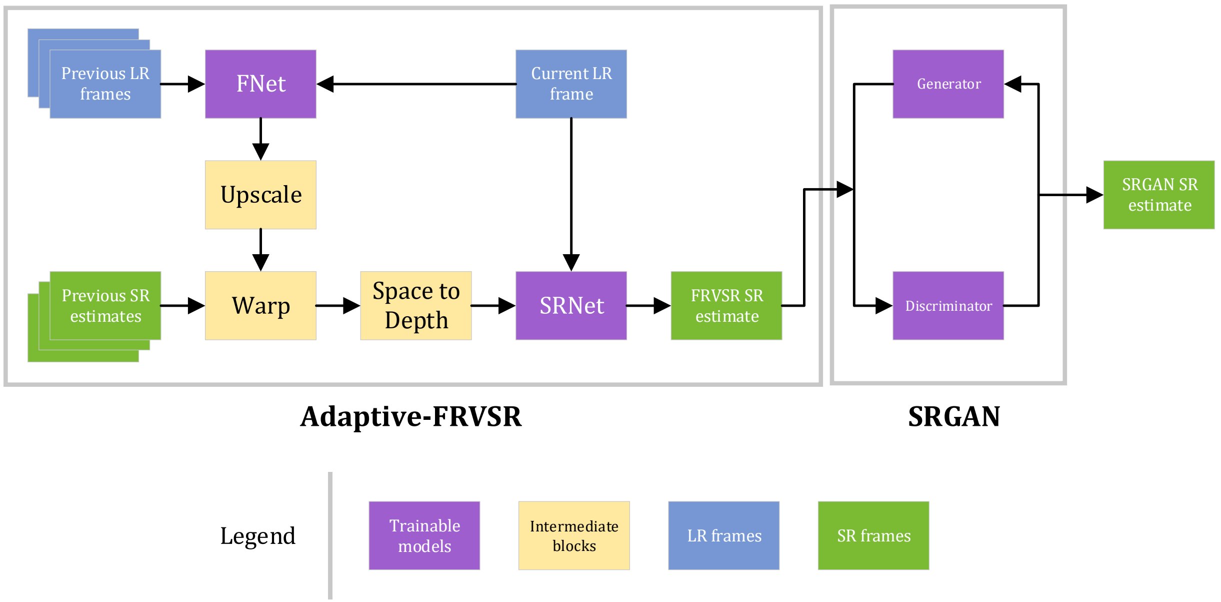


Figure 2: Network architecture

B. Implementation

The proposed implementation consists of an Adaptive-FRVSR and SRGAN. The implementation which is currently functional is Baseline-FRVSR and SRGAN. We have implemented FRVSR using the implementation from [7] as reference. The FRVSR implementation available at [7] lacks important elements such as a top-level flow that performs testing on a sample using the trained model and reports back performance metrics. Moreover, it lacks documentation in the form of code comments. We are thus re-writing most of the implementation. For our SRGAN implementation, we are utilizing the one available at [6].

Note that the current Baseline-FRVSR implementation, as proposed in [5], utilizes the last SR estimate frame and the current LR frame for estimating the current SR frame. Our proposed Adaptive-FRVSR implementation utilizes adjacent frames (future LR frames and previous SR estimates) in addition to the current frame to improve VSR performance.

The code for the implementation is available at `https://github.com/amanchadha/iSeeBetter`. Samples are also available in the repository.

C. Loss functions

Our loss-functions are based on FRVSR and SRGAN with relevant modifications to utilize adjacent frames.

For FRVSR, we use two loss terms to train our model, as show in Fig. 2. The loss L_{sr} is applied to the output of SRNet and is backpropagated through both SRNet and FNet:

$$L_{sr} = \sum_{t=k1}^{k2} \left\| I_t^{est} - I_t^{HR} \right\|_2^2 + \sum_{t=l1}^{l2} \left\| I_t^{est} - I_t^{LR} \right\|_2^2 \qquad (1)$$

Note that the value of *k1* and *l1* are the previous estimated SR frame and the next LR frame respectively. *k2* and *l2* are determined by an image similarity statistical measure such as SSIM. If the similarity between the adjacent frames is beyond a certain high threshold (say, 95%) threshold, we qualify these frames as usable. based on a. We expect image similarity to reduce as we go beyond a particular value of n in either the forward or backward direction in time. Our technique thus not only takes into account spatial information in the current frame but also temporal information in the adjacent frames thereby offering superior reconstruction fidelity.

Since we do not have a ground truth optical flow for our video dataset, we calculate the spatial mean squared error on the warped LR input frames leading to the auxiliary loss term L_{flow} to aid FNet during training.

$$L_{flow} = \sum_{t=k1}^{k2} \left\| WP(I_{t-1}^{LR},\, F_t^{LR}) - I_t^{LR} \right\|_2^2 \qquad (2)$$

The values of *k1* and *k2* are obtained similarly as above. The total loss used for training is $L = L_{sr} + L_{flow}$.

For SRGAN, we define our loss function based on [3] to be:

$$\begin{aligned} Loss_{G_{\theta_G}}(t) = \; & \alpha \times MSE\left(I_t^{est},\, I_t^{HR}\right) \\ & - \beta \times log\left(D_{\theta_D}\left(I^{est}\right)\right) \\ & + \gamma \times PercepLoss\left(I_t^{est},\, I_t^{HR}\right) \\ & + \delta \times TVLoss\left(Iest,\, IHR\right) \\ & + \in \times MSE\left(\tilde{I}_{t-1}^{LR},\, I_t^{LR}\right) \end{aligned} \qquad (3)$$

$$Loss_{D_{\theta_D}}(t) = 1 - D_{\theta_D}(I_t^{HR}) + D_{\theta_D}(I_t^{est}) \qquad (4)$$

The total loss of a sample is the average of all frames.

$$\begin{aligned} Loss_{G_{\theta_G}} &= Avg(Loss_{G_{\theta_G}}(t)) \\ Loss_{D_{\theta_D}} &= Avg(Loss_{D_{\theta_D}}(t)) \end{aligned} \qquad (5)$$

Once we have achieved functional convergence, as a stretch goal for the project, we would like to improve the performance of the algorithm by profiling and optimizing the runtime of the algorithm. The end goal would be to make the algorithm run at real-time during playback, thereby enabling VSR on-the-fly. This would require us to be able to perform frame generation within a couple of milliseconds to subsume processing within the 16.67ms intervals between successive VSYNCs (for smooth 60FPS playback).

D. Next steps

Planned work includes: (i) Adaptive-FRVSR implementation, (ii) training the current model for more iterations, (iii) data augmentation using other similar datasets, (iv) reporting performance on standard datasets to compare with currently available VSR techniques, (v) code refactoring, (vi) documentation in the form of code comments and (vii) establishing additional performance metrics.

III. Experimental Results

A. Dataset

To evaluate the proposed model, we used the Vimeo90K dataset collected in the TOFlow project of MIT CSAIL [8] which contains around 90,000 7-frame HR sequences with a fixed resolution, extracted from 39K video clips from Vimeo.com. When training our models, we generated the corresponding LR frame for each HR input frame by performing 4x down-sampling. To extend our dataset further, we have also built a video-to-frames tool to collect more data from YouTube.

B. Training platform

To train the model, we used the Amazon EC2 P2.XLarge instance which provides 16 NVIDIA K80 GPUs, 64 vCPUs and 732 GB of host memory.

C. Results

Note that the below results are with the baseline implementation of FRVSR and SRGAN. The baseline FRVSR scheme only looks at the prior SR estimate and the current LR frame, while the adaptive scheme utilizes multiple similar adjacent frames along with the current LR frame.

TABLE I. TEMPORAL PROFILES

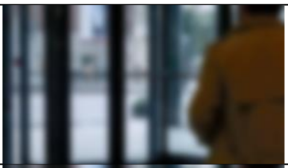
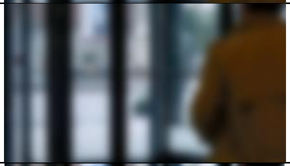
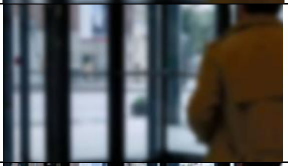
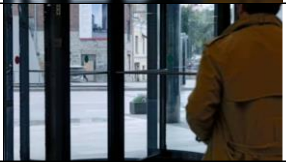
SRGAN	
FRVSR	
iSeeBetter (using the baseline FRVSR scheme)	
Ground Truth	

TABLE II. COMPARISON OF TEMPORAL PROFILE PSNR FOR THE ABOVE EXAMPLE

Model	**Temporal Profile PSNR (dB)**
SRGAN	22.44
FRVSR	19.63
iSeeBetter (using the baseline FRVSR scheme)	**23.56**

From the above results, we can see that the baseline implementation yields a 5% improvement over SRGAN. We expect to see bigger improvements once we implement the adaptive FRVSR scheme.

REFERENCES

[1] Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] M. Cheng, N. Lin, K. Hwang, and J. Jeng, “Fast video super-resolution using artificial neural networks,” in Proc. 8th Int. Symp. Commun. Syst. Netw. Digital Signal Process. (CSNDSP), 2012, pp. 1–4.

[3] López-Tapia, Santiago, et al. "A Single Video Super-Resolution GAN for Multiple Downsampling Operators based on Pseudo-Inverse Image Formation Models." arXiv preprint arXiv:1907.01399 (2019).

[4] Gopan, K. Gopika and G A Sathish Kumar. “Video Super Resolution with Generative Adversarial Network.” 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (2018): 1489-1493.

[5] Sajjadi, Mehdi SM, Raviteja Vemulapalli, and Matthew Brown. "Frame-recurrent video super-resolution." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6626-6634. 2018.

[6] Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken 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.

[7] FR-SRGAN, Report for MIT 6.819 Advances in Computer Vision, Nov 2018.

[8] http://toflow.csail.mit.edu

[9] Boris Kovalenko, “Super resolution with Generative Adversarial Networks”, CS 231n Project Report, 2017.