Spatio-Temporal Recurrent Adversarial Back Projection Network for Video Super Resolution

Ankit Chadha & M. Masoud: [ankitrc,masoud]@stanford.edu

Introduction

Problem: Video super-resolution (VSR) aims to infer a high-resolution video sequence from multiple low-resolution frames. VSR helps with many important applications such as new scene generation and anomaly detection. The Recurrent Back-Projection Network (RBPN) uses a recurrent framework that treats each frame as a separate spatial source of information and combines this information in a recurrent iterative refinement framework inspired by the idea of back-projection to produce temporally coherent multi-image super-resolution. In this project, we propose a novel architecture for a spatio-temporal adversarial recurrent discriminator to achieve photorealistic and temporally coherent super resolved frames with a more sophisticated objective function to fine-tune spatial and temporal features of RBPN.

Model

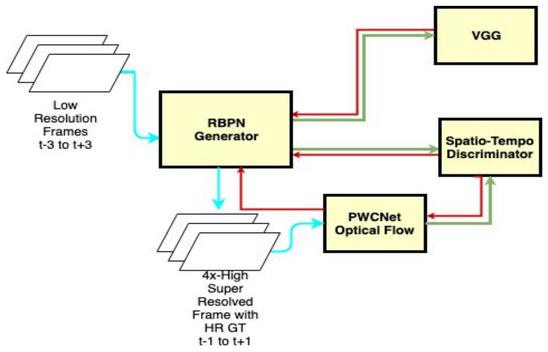
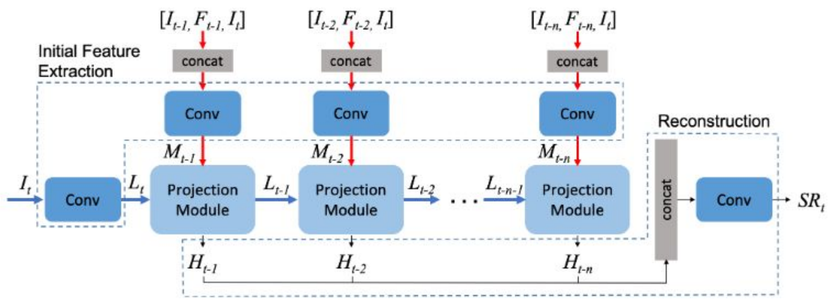


Figure 1: Overall Model Architecture.

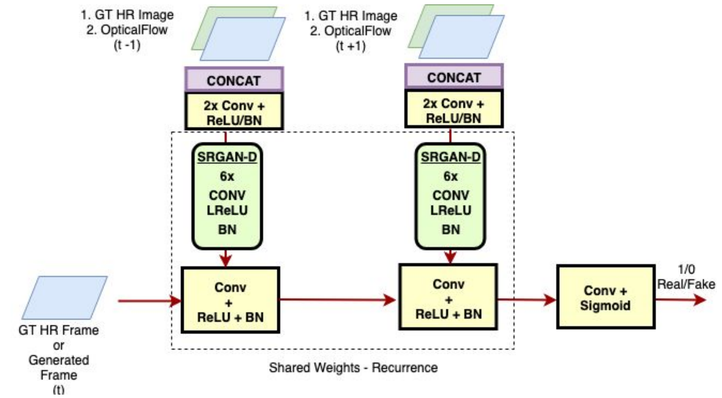
Our temporally coherent VSR architecture consists of four main components: a recurrent back projection network (RBPN) generator, a flow estimation network (PWCNet), VGGNet for perceptual quality and a spatio-temporal recurrent discriminator.



RBPN Generator Network (Figure 2). RBPN is the current state of the art method for in-plane VSR (4x). This network can be subdivided into three stages:

1. Two types of feature tensors are extracted from the input data:
   a. L is extracted directly from the current frame via a convolution layer
   b. M is extracted from the concatenation of the current frame, previous frame, and the Pyflow precomputed flow map.
2. The projection module is constructed from an encoder and decoder. Each projection module takes in the feature tensors and outputs the high resolution feature tensor H, produced by the encoder, and the input frame feature tensor for the subsequent projection module, which is produced by the decoder.
3. The final stage completes the super-resolution by reconstructing the SR image from the convolved concatenation of all of the high resolution feature tensors produced in the second stage.

Model Cont.



Spatio-Temporal Recurrent Discriminator (Figure 3:) The discriminator takes the SR or HR frame as an input alongside the neighboring HR ground truth frames and their optical flow with respect to the target frame (t). The task of the discriminator is to predict the generated SR image as fake (0) and the ground truth image as real (1).

Results

	Baseline (RBPN)	Ours	Ours-w-flow
Scene	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Foliage	24.21/0.76	23.93/0.76	24.08/0.75
City	25.36/0.77	**25.46/0.78**	**25.54/0.77**
Walk	28.60/0.89	28.40/0.89	28.45/0.89
Calendar	21.74/0.77	21.62/**0.78**	21.71/**0.78**

Baseline (RBPN)	Ours	Ours-w-flow
PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
35.52/0.94	35.30/**0.95**	35.41/**0.95**

Table 1 & 2 : Our results on Vid4 and vimeo90k datasets compared to the baseline RBPN. Our network does marginally better than the baseline on the City scene (both PSNR and SSIM) in Vid4 and marginally better on Calendar for SSIM metric. For vimeo90k test set, our model performs incrementally better for SSIM. Our model tends to do marginally better than the baseline RBPN on SSIM than on PSNR. Since SSIM accounts for texture and spatial details more than PSNR, better SSIM clearly suggest that the adversarial training with our proposed spatio-temporal discriminator helps achieve better high-frequency details in both spatial as well as temporal relationships.

Conclusion

1- Capturing different motion regimes is a major challenge in producing visually and temporally consistent results is.2- Our model introduce a spatio-temporal recurrent discriminator to produce temporally and spatially consistent results between frames. 3- Adversarial training adds another level of complexity to the training process and requires extended training time, computational resource and careful tuning. 4- Initial results appear promising and our model does incrementally better than the baseline.

Loss

Generator L1

$$\mathcal{L}_G = \|G_t - GT_t\|$$

Perceptual VGG L2

$$\mathcal{L}_P = \|\Phi_{VGG}(G_t) - \Phi_{VGG}(GT_t)\|_2$$

Spatio-Tempoal Disc. Adv Loss

$$\mathcal{L}_{Ds,t} = -log(1 - D(G_t, PyF(GT_t)) - log(D(GT_{t,}, PyF(GT_t))$$
$$\mathcal{L}_{Gs,t} = -log(D(G_t))$$

Spatio-Tempoal Flow Disc. Adv Loss

$$\mathcal{L}_{Df,t} = -log(1 - D(F(G_{t-k}, G_t)) - log(D(F(GT_{t-k}, GT_t))$$
$$\mathcal{L}_{Gf,t} = -log(F(G_{t-k}))$$

Discussion

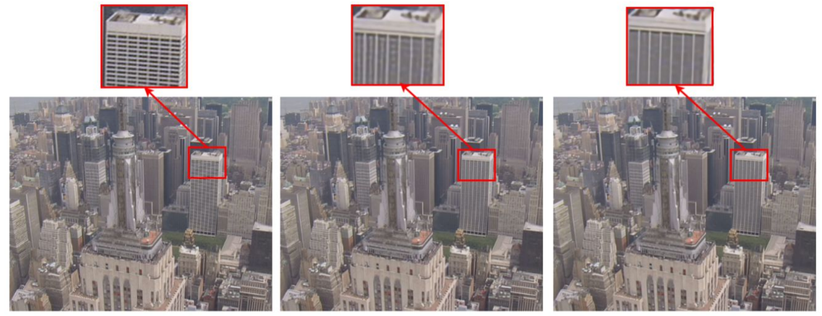
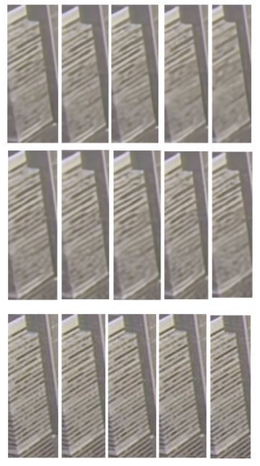


Figure 4: City Scene: Left to right: Ground Truth, Our results, Baseline - our model shows better high frequency details compared to the baseline



Successes: Our network improves over baseline:

1. More high frequency details in the buildings and the edges and separations are more apparent
2. Looking at the neighboring ground truth frame we observed that our training is capturing those edges from nearby frames.
3. Our adversarial model and the spatio-temporal discriminator helped the generator to maintain the sharper edges which leads to natural temporal progressions to closely match the desired ground truth behavior over time.

Areas for Further Improvement:

1. Some edges seem tilted over flat surfaces
2. Compared to the ground truth, we can see that the SR image still has a lot of details missing, such as in the edges around windows in the cropped part of the image

Figure 5: City Scene - Temporal profile of part of the city scene - Red Arrow shows time. Top to bottom: Baseline, Our results, Ground Truth.

References

[1] M. Haris, G. Shakhnarovich, N. Ukita. Recurrent Back-Projection Network for Video Super-Resolution, Conference on Computer Vision and Pattern Recognition, (2019)
[2] M. Chu, Y. Xie, L. L. Taixé, N. Thuerey, TecoGAN: A Temporally Coherent GANs for Video Super-Resolution, Computer Vision and Pattern Recognition, (2019).
[3] D. Sun, X. Yang, M. Liu, Jan Kautz, PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume, International Computer Vision and Pattern Recognition, (2018).