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

Ankit Chadha
Stanford
ankitrc@stanford.edu

Mohamed Masoud
Stanford
masoud@stanford.edu

Abstract

Video super-resolution (VSR) aims to infer a high-resolution (HR) video sequence from multiple low-resolution (LR) frames. VSR helps with many important applications such as new scene generation, anomaly detection and recovery of information lost during video compression. For single-image super-resolution (SISR), adversarial training has been highly successful, yielding realistic and highly detailed results. The Recurrent Back-Projection Network (RBPN) uses a recurrent framework that treats each frame as a separate spatial source of information [1]. Additionally, RBPN combines the frames' spatial information in a recurrent iterative refinement framework inspired by the idea of back-projection to produce temporally coherent multiple-image super-resolution (MISR). The recurrent generative framework integrates spatial and temporal contexts from continuous video frames using a recurrent encoder-decoder module that fuses multi-frame information with the SR version of the target frame [1]. 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. Our quantitative and visual analyses highlight the enhanced capability of the RBPN generator to be able to learn high frequency details both spatially and temporally.

1. Introduction

The goal of video super-resolution (VSR) is to generate a high resolution (HR) video from its corresponding low-resolution (LR) video. For single image super-resolution (SISR), adversarial training achieves major improvements in terms of perceptual quality. A straightforward way to perform VSR is to perform SISR on each video frame individually. However, the temporal coherence of the super-resolved video would be significantly impacted with a high probability that consecutive frames are not connected naturally, resulting in the flickering artifact.

In MISR, the missing details in target frame are available in the spatial contextual features of the target frame, as well as the temporal details in the neighboring frames. For the temporal details to be extracted, all the frames in the neighboring window must be spatially aligned implicitly or explicitly. The frames can be explicitly aligned using motion cues between the frames with alignment modules [10]. Implicit alignment can be achieved using convolutional networks, and the aligned frames are then fed into recurrent neural networks (RNNs) [12]) in temporal order. The RNN based methods are highly impacted by frames that have both slow and fast motion in the foreground [9].

Our proposed method aims to integrate the benefits of RBPN and the Spatio-Temporal recurrent discriminator. The RBPN overcomes the shortcomings of the other VSR methods including the significant change between the consecutive frames by iteratively super-resolivng each frame using the spatial context features along with temporal details in a recurrent fashion. The recurrent spatio-temporal discriminator enhances the high frequency details extracted from spatial contextual and temporal details. The details on the model's components are described in the methods and architecture section below. Our main contributions can be summarized as:

- Introducing a spatio-temporal recurrent discriminator that uses the SRGAN to enhance the high frequency details to capture spatial and temporal information.
- Incorporating adversarial training for 4x upsampled frame and flow information using PWCNet. This novel addition aims to add more emphasis on the temporal coherency by training the spatio-temporal discriminator to differentiate between the flow details generated by the generator and ground truth flow.

We report experiments to evaluate our VSR on Vid4 [5] and vimeo90k [13] which is a dataset containing various types of motion. The 4x (4-times) upsampled super-resolved output of the models are evaluated in terms of the

peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM).

2. Related Work

The goal of VSR is to super resolve the spatially realistic characteristics of a video frame while preserving temporal information to ensure a natural change over time. Many deep learning based VSR methods that exploit motion information in videos have been proposed. The most traditional approach to retain temporal information in VSR is by concatenating the frames as in [16, 10]. VSR DUF [16] proposed a mechanism to construct an adaptive motion estimation network that uses filters and residual images to handle various types of motion. However, this approach fails to represent the multiple motion regimes on a sequence because the input frames are concatenated [1].

The RNN approach was first introduced by [12]. The RNN model used a bidirectional connection between video frames with three types of convolutions: a feed forward convolution for spatial dependency, a recurrent convolution for long-term temporal dependency, and a conditional convolution for long-term contextual information. While RNN models improve the performance by utilizing temporal smoothness between neighbor frames, they face the challenge of jointly capturing the significant motion changes observed in neighboring frames.

The Recurrent Back-Projection Network (RBPN)[1], inspired by the idea of back-projection in multiple-image super-resolution [2], considers each context neighboring frame as a separate source of temporal information. These separate temporal details are iteratively refined and combined. Furthermore, rather than explicitly aligning frames, the RBPN explicitly represents estimated inter-frame motion with respect to the target, thus allowing for better performance in multiple motion regimes.

For single image super resolution (SISR), prior findings highlight that adversarial training significantly helps in obtaining realistic high-frequency details and significantly improves the perceptual quality [11]. TecoGAN [3] and TempoGAN [4] introduced adversarial training for video super resolution and volumetric applications respectively. TecoGAN [3] introduced adversarial training using the spatio-temporal adversarial loss and the recurrent generator structure. Additionally, [3] proposed Ping-Pong loss that aims to remove recurrent artifacts and long-term drifting without reducing perceptual quality.

In this project, we propose to integrate the benefits of RBPN and adversarial training inspired by [3]. We introduce a spatio-temporal recurrent discriminator (with SRGAN pretrained layers) that differentiate between the spatial and temporal details generated by the RBPN generator and ground truth. Additionally, we introduce a novel temporal loss to the discriminator using the neighboring frames' flow information generated by PWCNet.

3. Data

The dataset used in this project is vimeo90k by TOFlow [13]. We are using the septuplet dataset which has 91,701 7-frame sequences which cover a variety of motion and detail in the videos. We augmented the data by flipping, rotating, and randomly cropping the input frames to extend the distribution and variance of the dataset and reduce overfitting. Each augmentation was applied to the whole septuplet to avoid discrepancies between frames. The choice of this dataset was inspired by RBPN [1], which we choose as the model generator. The proposed model uses as an input 7 LR frames including the target LR frame for SISR. The network is trained on patches rather than full frames. These patches are 64x64 random crops from the original video frames. The pre-trained network does not perform any explicit normalization to the data. The data was arranged into 70-15-15 split for train, validation and test sets. Additionally, following most of the video super resolution literature comparisons, we conducted further test experiments over the Vid4 dataset. That allows to further test the generalizability of our model on other datasets and to have a fair comparison with SoTA methods.

4. Methods and Architecture

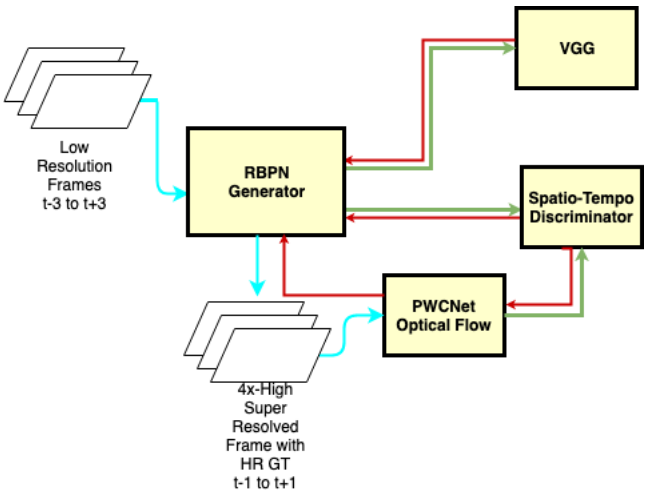

Figure 1. Overall Training Architecture

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.

The RBPN generator is used to generate high-resolution video frames from low-resolution inputs. The PWCNet

flow estimation network learns the temporal motion estimation between super-resolved frames to aid the spatio-temporal discriminator. We used an incremental approach for our adversarial training with and without PWCNet optical flow. First, we trained the generator as discussed in section 4.1 to fool the discriminator. Second, the generator and the pretrained PWCNet flow estimator are trained together to fool the discriminator.

The recurrent spatio-temporal discriminator takes into account spatial as well as temporal aspects and penalizes incoherent temporal discontinuities without excessively smoothing the image content. The adversarial training requires the generator to generate high-frequency details that are coherent with neighboring frames.

An additional level of temporal information we consider in our model is the optical flow estimated by the PWCNet. The discriminator learns directly from the *explicit* temporal information embedded in the flow images. Therefore, we propose independent flow generator and discriminator adversarial losses that put more emphasis on the temporal details and force the generator to produce temporally realistic details.

4.1. RBPN Generator

RBPN is the current state of the art method for in-plane VSR (4x) (Figure 2). This network can be subdivided into three stages. First, two types of feature tensors are extracted from the input data. Feature tensor L is extracted directly from the current frame via a convolution layer, while feature tensor M is extracted from the concatenation of the current frame, previous frame, and the Pyflow precomputed flow map. The second stage encompasses the series of projection modules. The projection module is constructed from an encoder and decoder. Each projection module takes in the two types of 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. 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.

4.2. Optical Flow

RBPN uses a static optical flow generation tool called PyFlow [14], Pyflow is a python wrapper of Coarse2Fine Optical Flow implementation of FAIR's [15] *Learning Features by Watching Objects Move* paper. [15] is an unsupervised feature learning approach of optical flow and object motion estimation. Pyflow is used to estimate the optical flow of target and each context frame. The flow is estimated per batch of data.

In our adversarial spatio-temporal discriminator, we use a deep CNN model to produce optical flow, called PWCNet [6]. PWCNet uses pyramidal processing, warping, and the use of a cost volume to estimate optical flow between two images. PWCNet is used to provide the discriminator with the temporal information embedded in (two channels) optical flow images between the upsampled (4x) target frame and two neighboring context frames. In our implementation, we use PWCNet instead of Pyflow python tool for easier training. We fine-tune the pretrained PWCNet through the flow forward and backward passes as show in Figure 1. Figure 3 shows visualization of sample optical flow (two channels) images for both Pyflow and PWCNet. Both Pyflow and PWCNet produce similar results for HR images.

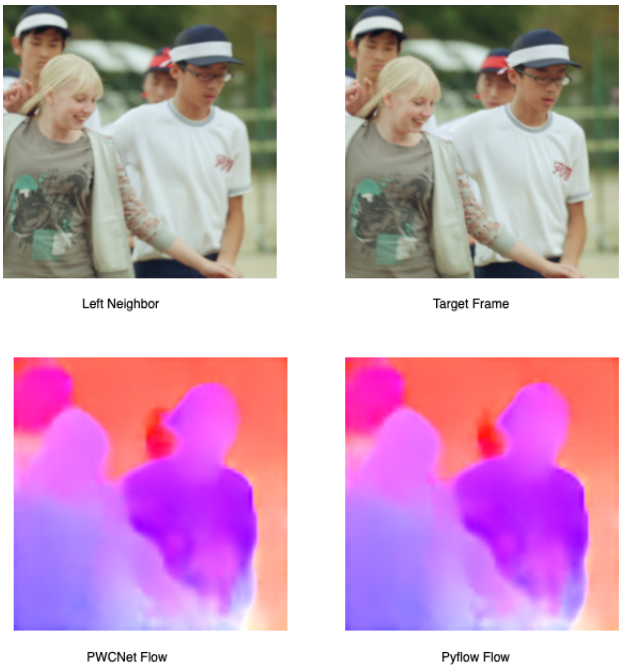

Figure 3. PWCNet vs Pyflow Optical Flow images

4.3. Perceptual Loss - VGGNet

Pre-trained VGGNet is used to employ a in-plane L2 Perceptual Loss [7] to help the generator produce visually pleasing results. The Perceptual Loss has been shown to help training converge faster and also to optimize structural details using features extracted from the pre-trained VGG Network.

4.4. Spatio-Temporal Recurrent Discriminator

As show in Figure 4, inspired by TecoGAN [3] and TempoGAN [4], we propose a novel architecture for a spatio-temporal discriminator that will learn a loss function for both spatially and temporally consistent frames from the ground truth target and ground truth neighbors. The goal of a generative adversarial network is to minimize the KL

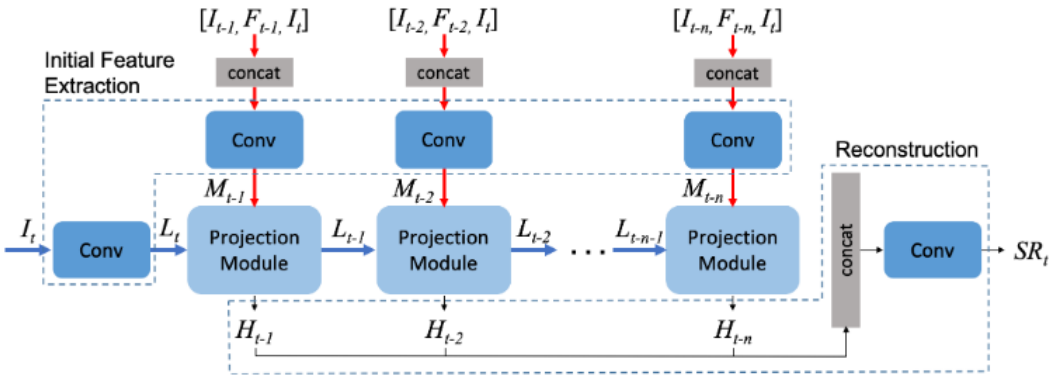

Figure 2. RBPN Generator

divergence between the input and ground truth distributions. As such, the advantage of an adversarial network is that the final distribution of the super-resolved results is likely more similar to the ground truth distribution than that of the super-resolved results of the generator alone. This fact implies that the adversarial super-resolved images are visually closer to the ground truth images, and thus are more realistic.

The discriminator takes the generated SR or the ground truth 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). The discriminator also takes the PWCNet generated flow images of the generator SR frame and the context neighboring frames to explicitly predict if the flow of the SR image and its neighbors as fake and the ground truth flow images as real.

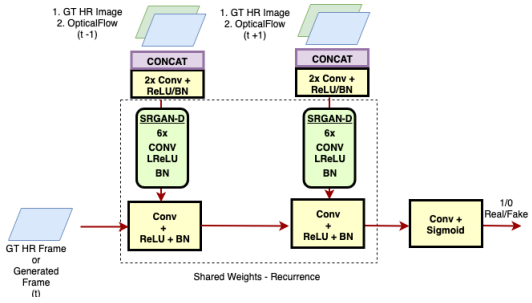

Figure 4. Spatio-Temporal Recurrent Discriminator

4.5. Training Objective

The red arrows in Figure 1 indicate backpropogation flow. Our training objective can be summarized with the following loss functions. Our model uses the discriminator to adversarially train/fine-tune both the generator and PWC-Net.

- Generator L1 Loss:

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

- Perceptual VGG L2 Loss:

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

- Spatio-Temporal Discriminator Adversarial 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-Temporal Discriminator flow Adversarial Loss:

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

- s - Spatial component
- t - Temporal component
- f - flow component
- PyF - Pyflow Optical Flow
- F - PWCNet Optical Flow
- D - Discriminator
- t - time
- k - index over time -1,+1
- Φ_{VGG} - VGG Network Features map

5. Experimental Results

In order to evaluate our model, we conducted visual and quantitative analyses. Quantitatively, we evaluate our methods in terms of peak signal-to-noise ratio (PSNR) and Structural similarity index (SSIM). PSNR is a pixel-wise accuracy metric that represents a measure of the peak error. It is inversely proportional with the mean square error between the true and generated images. MSE and PSNR are not highly indicative of perceived similarity. SSIM aims to address this shortcoming by taking texture into account. We trained the RBPN (generator) and the recurrent spatio-temporal discriminator initialized with pretrained weights. We trained for a total of 14 epochs. Tables 1 and 2 summarize the results of the experiments on both Vid4 and vimeo90k datasets. 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.

It is interesting to note that 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. That is consistent with the visual analysis in section 5.2.

	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**

Table 1. Evaluation on Vid4 Dataset

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 2. Evaluation on Vimeo Dataset

5.1. Hyperparameter Tuning

Hyperparameter tuning was performed with our cross validation set (15% of the data). Hyperparameter tuning is especially challenging for the task of VSR since it is very computationally intensive due to the 7 frame input and our three frame spatio-temporal discriminator. Adam was used as an optimizer for the Generator and the Discriminator. The following hyperparameters were chosen for our experiments.

- **Batch Size:** we were only able to fit a batch size of 2 on the compute available.
- **Generator Learning Rate:** Since we had to scale down the batch size we decided to make sure the fine-tuning done over the RBPN network was scaled down by a factor of 10, for a learning rate of 1e-5.
- **Discriminator Learning Rate:** Our recurrent discriminator uses SRGAN's pretrained layers so we decided to start off with the default value of 1e-4. However, to cope up with a largely pretrained generator for the task we decided to boost the discriminator's learning rate to 1e-3.

5.2. Results Discussion: Spatio-Temporal Discriminator

Table 1 compares PSNR and SSIM metrics with respect to the baseline (default pre-trained RBPN). Even though our model and the baseline produce marginally similar PSNR and SSIM , the adversarial training produces visually sharper results and adds more detail with the addition of high frequency components to the frames both spatially and from temporally contiguous frames. Our network does marginally better than the baseline on the City scene (both PSNR and SSIM) in Vid4 and marginally better on Calendar with respect to the SSIM.

A visual analysis was done to compare our network and the baseline RBPN outputs.

- **Wave Scene** Figure 6 presents a comparison of a snippet from the test set of vimeo90k dataset. Our network seems to be capturing high frequency texture information of the water waves more than the baseline. As shown, even though the water texture seems visible, there is still room for improvement to produce more detail as compared to ground truth image.
- **City Scene** Figure 5 is the City scene from the Vid4 dataset, our network seems to be adding more detail to the buildings, the edges and separations are more apparent than the baseline. Some of the edges seem tilted over relatively flat surfaces. Looking at the neighboring ground truth frame we observed that our training is capturing those edges from nearby frames. 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. This presents various opportunities further improvement for VSR.
- **Temporal Profile** Figure 7 shows the progression of a part of the city scene over 5 time steps. As show in the figure while the RBPN in earlier time steps captured

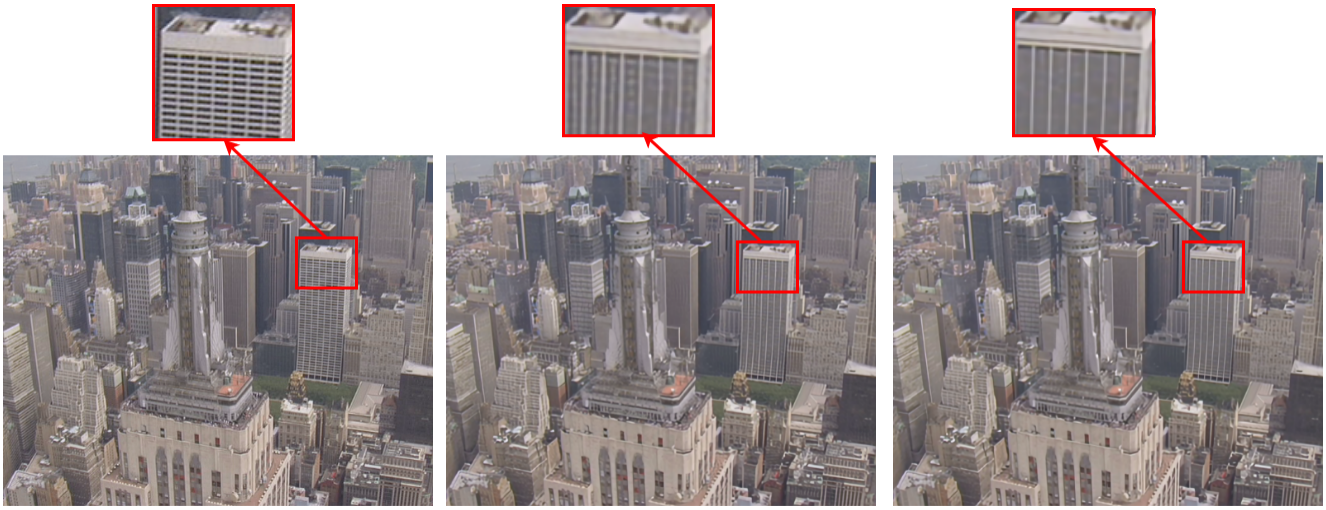

Figure 5. City Scene Results: our model shows better high frequency details compared to the baseline

Figure 6. Waves Scene Results: our model shows better high frequency details compared to the baseline

some of the high frequency details in the building, it fails to maintain the details in later time steps. 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.

5.3. Spatio Temporal Discriminator with Flow

To evaluate our model with the additional adversarial temporal flow details, we incorporated the PWCNet network as described in Section 4. Due to the computational budget limitation, we were only able to fit a batch of 1 set of video frames. Within the timeline, we were able to train the model for 3 more epochs with the trained weights of the discriminator as the starting point. The (ours-w-flow) results as shown in tables 1 and 2 are similar (slightly better) to the adversarial training with no explicit temporal flow adversarial loss/training (no PWCNet and no flow layers in the discriminator)as expected. We expect the temporal coherence would be enhanced with training our full architecture (with PWCNet flow) much longer than 3 epochs.

6. Conclusion and Future Work

Video Super Resolution presents a unique challenge in the area of computer vision. One major challenge in producing visually and temporally consistent results is capturing different motion regimes. We present our take on the task with spatio-temporal discriminator to produce temporally and spatially consistent results between frames. Adversarial training adds another level of complexity to the training process and requires extended training time, computational resource and careful tuning.

Initial results appear promising and our model does incrementally better than the baseline with respect to structural content in the SR image. Pretraining is helpful but

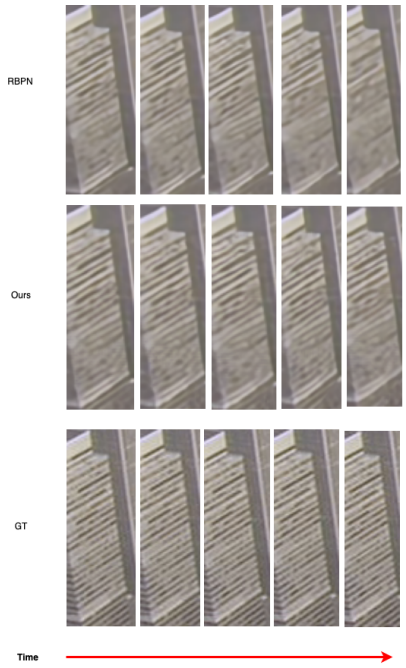

Figure 7. City Scene - Temporal profile of part of the city scene - Time shown on the x-axis

presents it's own set of challenges with respect to a balance in adversarial training where both networks need to be competitive in order to learn from each other. In our case, pretraining for the discriminator was not as conducive as for the generator where we applied weights learned for SISR with SRGAN to a MISR task. Furthermore, the task of VSR is computationally intensive and limits the flexibility of the experiments. In the future, we plan to apply the lessons learned and to extend our approach further with more features like depth. We also plan to further experiment with the proposed discriminator. Given more time and compute resources, we would like to experiment and tune our adversarial setting further to aid sharper and more temporally coherent results with the additional flow loss (full architecture figure 1). We would also like to explore additional applications for our spatio-temporal recurrent discriminator such as video frame interpolation and denoising.

We are currently training both the generator and discriminator from scratch and hope to present more results during the poster session.

7. Contributions and Acknowledgements

Both authors contributed equally to problem research, formulation, design and coding of the model architecture, development environment setup, data preparation and analysis, designing and conducting experiments, hyperparameter tuning, evaluation and visual/error analyses and report writing.

We are thankful to the CS231n Team for all the support and help throughout and to the folks at Google for the Cloud credits that made this work possible.

References

[1] M. Haris, G. Shakhnarovich, N. Ukita. Recurrent Back-Projection Network for Video Super-Resolution, *Accepted: Conference on Computer Vision and Pattern Recognition*, (2019).

[2] M. Haris, G. Shakhnarovich, N. Ukita. Deep Back-Projection Networks For Super-Resolution , *Conference on Computer Vision and Pattern Recognition*, (2018).

[3] M. Chu, Y. Xie, L. L. Taix, N. Thuerey, tempoGAN: A Temporally Coherent, Volumetric GAN for Super-resolution Fluid Flow, textitACM Transaction on Graphics, (2018).

[4] Y. Xie, E. Franz, M. Chu, N. Thuerey, tempoGAN: A Temporally Coherent, Volumetric GAN for Super-resolution Fluid Flow, textitACM Transaction on Graphics, (2018).

[5] C. Liu and D. Sun. A bayesian approach to adaptive video super resolution, *European Conference on Computer Vision*, (2016).

[6] D. Sun, X. Yang, M. Liu, Jan Kautz PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost

Volume, *Proceedings of International Computer Vision and Pattern Recognition*, (2018).

[7] J. Johnson, A. Alahi, F. Li Perceptual Losses for Real-Time Style Transfer and Super-Resolution, *European Conference on Computer Vision*, (2016).

[8] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia. Video super-resolution via deep draft-ensemble learning, *Proceedings of the IEEE International Conference on Computer Vision,*pages 531539, (2014).

[9] F.Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm, *Neural Computation* 12(10):24512471, (2000).

[10] J. Caballero, C. Ledig, A. P Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , (2017).

[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang. Photo-realistic single image super-resolution using a generative adversarial network, *Conference on Computer Vision and Pattern Recognition* , (2017).

[12] Y. Huang, W. Wang, and L. Wang Bidirectional recurrent convolutional networks for multi-frame superresolution, *Advances in Neural Information Processing Systems* pages 235243, (2015).

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

[14] https://github.com/pathak22/pyflow

[15] D. Pathak, R. Girshick, P. Dollr, T. Darrell, B. Hariharan, Learning Features by Watching Objects Move, *Conference on Computer Vision and Pattern Recognition* , (2017).

[16] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 32243232, (2018)