iSeeBetter: Spatio-Temporal Video Super Resolution using Recurrent-Generative Back-Projection Networks

https://www.youtube.com/watch?v=2HC0wdeQRiM

Aman Chadha
amanc@stanford.edu

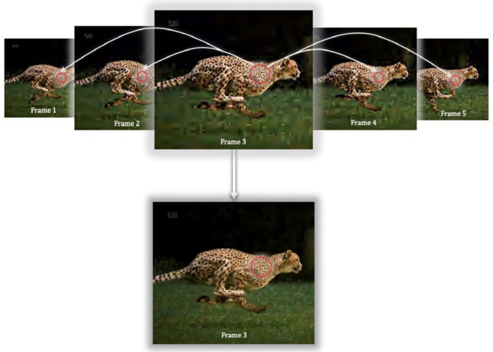
Introduction

Motivation

- ✓ Applying Single Image Super Resolution (SISR) successively to each video frame leads to lack of temporal coherency
- ✓ Video Super Resolution (VSR) models based on CNNs outperform traditional approaches in terms of PSNR
- ✓ However, CNNs lose finer texture details when super-resolving at large upscaling factors

Underlying Principles

- ✓ Use data from adjacent frames along with the input frame
- ✓ Use GANs for a competitive advantage compared to CNNs



Approach

- ✓ iSeeBetter: spatio-temporal VSR
- ✓ Uses recurrent-generative back-projection networks
- ✓ Extracts spatial and temporal information from current + neighboring frames
- ✓ Improves the "naturality" of the output while eliminating artifacts, using super-resolution GAN discriminator
- ✓ Uses a four-fold (adversarial, perceptual, MSE and TV) loss function that focuses on perceptual quality

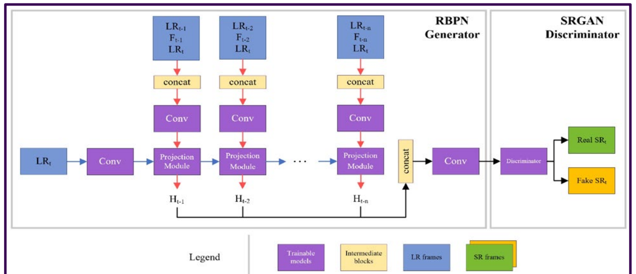
Datasets

Approach

- ✓ Amalgamated diverse datasets with differing video lengths, resolutions, motion sequences and number of clips
- ✓ Generated LR frame for each HR input by down-sampling
- ✓ Training/validation/test split was 80/10/10

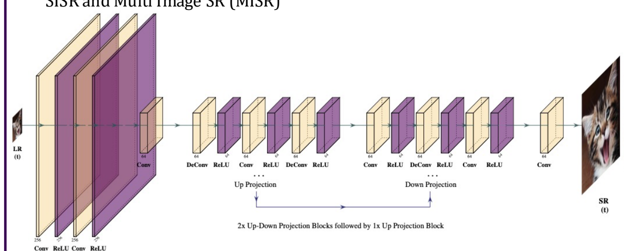
Dataset	Resolution	# of clips	# of frames/clip	# of frames
Vimeo90K	448 × 256	39,000	7	91,701
SPMCS	240 × 135	30	31	930
Vid4	(720 × 576/480 × 3)	4	41, 34, 49, 47	684
Augmented	(960 × 720)	7,000	110	77,000
Total	-	**46,034**	-	**170,315**

Model Architecture



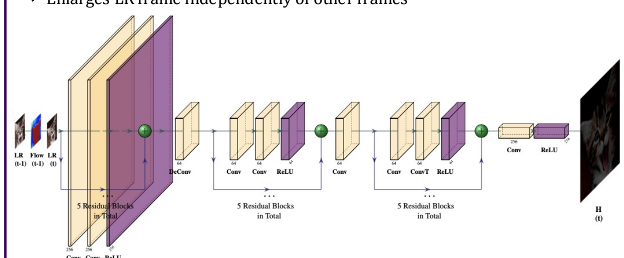
Building Blocks

- ✓ Uses RBPN [2] as generator and SRGAN [1] as discriminator
- ✓ RBPN has two approaches that extract missing details from different sources: SISR and Multi Image SR (MISR)



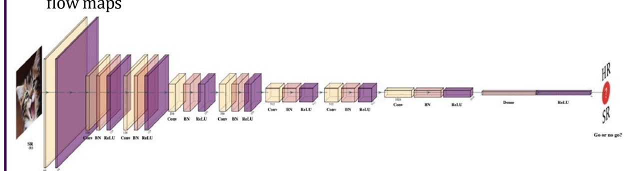
SISR Architecture

- ✓ Enlarges LR frame independently of other frames



MISR Architecture

- ✓ Computes residual features from a pair of input-to-neighbor frames and flow maps



Discriminator Architecture

- ✓ Trained to differentiate between SR images and original photo-realistic images

Loss Functions

MSE Loss

- ✓ MSE improves PSNR/SSIM but these metrics may not capture fine details in the image
- ✓ Experimentally, it was found that even manually distorted images still had an MSE comparable to the original image

Four-fold Loss

- ✓ Adversarial loss: focuses on perceptual similarity to limit model "fantasy"
- ✓ Perceptual loss: relies on features extracted from a pre-trained network
- ✓ MSE loss: pixel-wise error between the SR output and the HR source
- ✓ TV loss: de-noising function

$$Loss_{G_{\theta_G}}(t) = \begin{aligned} &\alpha \times MSE\left(I_t^{est}, I_t^{HR}\right) \\ &-\beta \times log\left(D_{\theta_D}\left(I^{est}\right)\right) \\ &+\gamma \times PercepLoss\left(I_t^{est}, I_t^{HR}\right) \\ &+\delta \times TV\,Loss\left(I_t^{est}, I_t^{HR}\right) \end{aligned}$$

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

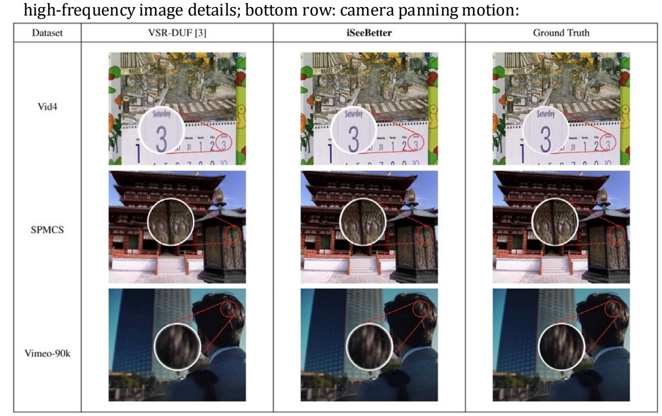
Results

Results

- ✓ PSNR/SSIM evaluation of state-of-the-art VSR systems for 4× upsampling:

Dataset	Clip Name	VSR-DUF [3]	RBPN/6-PF [2]	**iSeeBetter**
Vid4	Calendar	24.09/0.813	23.99/0.807	**24.13/0.817**
	City	28.26/0.833	27.73/0.803	**28.34/0.841**
	Foliage	26.38/0.771	26.22/0.757	**26.27/0.773**
	Walk	30.50/0.912	**30.70/0.909**	30.68/0.908
Vimeo90K	Fast Motion	37.49/0.949	40.03/0.960	**40.17/0.971**
Average		27.31/0.832	27.12/0.818	**27.36/0.835**

- ✓ Top row: fine-grained textual features that help with readability; middle row: intricate high-frequency image details; bottom row: camera panning motion:



Conclusion

- ✓ iSeeBetter offers superior VSR fidelity and surpasses state-of-the-art performance for majority of test sequences by combining spatial and temporal information
- ✓ Four-fold loss function helps emphasize perceptual quality

References

[1] C. Ledig, et al., "Photo-realistic single image super-resolution using a generative adversarial network," CVPR 2017, pp. 4681–4690.

[2] M. Haris, et al., "Recurrent back-projection network for video super-resolution," CVPR 2019, pp. 3897–3906.

[3] Y. Jo, et al., "Deep video super-resolution network using dynamicupsampling filters without explicit motion compensation," CVPR 2018, pp. 3224–3232.