Frame-Recurrent Video Super-Resolution

Mehdi S. M. Sajjadi [1,2] Raviteja Vemulapalli [2] Matthew Brown [2]

[1] Max Planck Institute for Intelligent Systems [2] Google

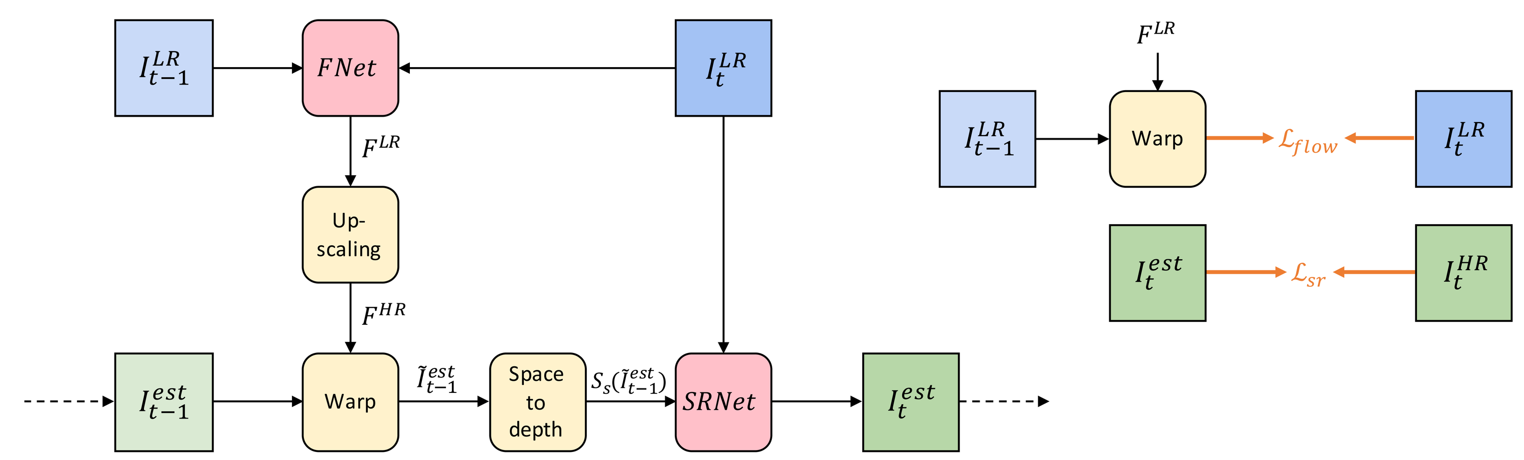
msajjadi.com

Overview

- **Common approach - Consider video SR task as a multi-frame SR task:**
  - Process several surrounding input frames to generate a single output frame.
  - Apply this to the entire video in a sliding window fashion.
- **Weaknesses of the multi-frame SR approach:**
  - Each output video frame is produced independently **limiting the system's ability to produce temporally consistent results.**
  - Each input video frame is processed multiple times **increasing the computational cost**.
- **Proposed frame-recurrent video super-resolution (FRVSR) framework**
  - achieves a **significant boost in image quality** despite being **more efficient**.
  - produces **more temporally consistent** output videos.
  - is fully convolutional and **end-to-end trainable from scratch**.
- **Illustration of the results for 4x upscaling:** The input image lacks significant amount of details which are restored by the proposed recurrent approach using information from the past.

low-resolution input | our result | ground truth

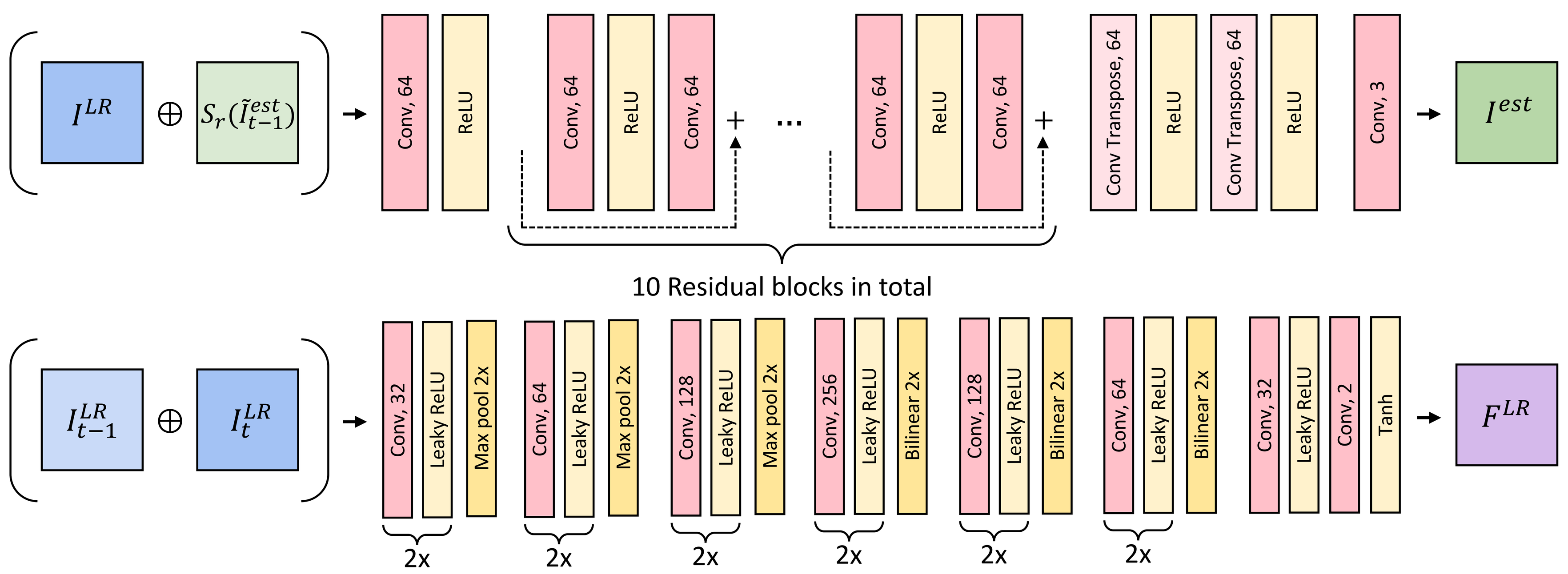
FRVSR Framework



Overview of the proposed FRVSR framework (left) and the loss functions used for training (right).

- **Inference steps:**
  - Compute the flow F^{LR} in LR-space using FNet.
  - Upscale the LR flow F^{LR} to HR flow F^{HR} (bilinear interpolation used in our implementation).
  - Warp the HR estimate I_{t-1}^{est} of the previous frame onto the current frame using F^{HR}.
  - Map the warped previous output $\tilde{I}_{t-1}^{\text{est}}$ to LR-space using the space-to-depth transformation.
  - Feed the previous output frame (after warping and mapping to LR space) and the current LR input frame I_t^{LR} to the super-resolution network SRNet which outputs an estimate for the current HR frame.
- **Training:** Both FNet and SRNet are **trained from scratch in an end-to-end fashion** by unrolling a fixed number of recurrent steps.
- **Loss functions for training:**
  - **The super-resolution loss** $\mathcal{L}_{\text{sr}} = ||I_t^{\text{est}} - I_t^{\text{HR}}||_2^2$ on the HR output encourages the network to produce video frames that are similar to the groundtruth.
  - **The flow loss** $\mathcal{L}_{\text{flow}} = ||\mathrm{WP}(I_{t-1}^{\text{LR}}, F^{\text{LR}}) - I_t^{\text{LR}}||_2^2$ on the warped previous LR frame aids the training of FNet.
- **Dataset:** Our training dataset consists of 256x256 image patches extracted from 40 high-resolution videos (720p, 1080p and 4K) downloaded from vimeo.com.

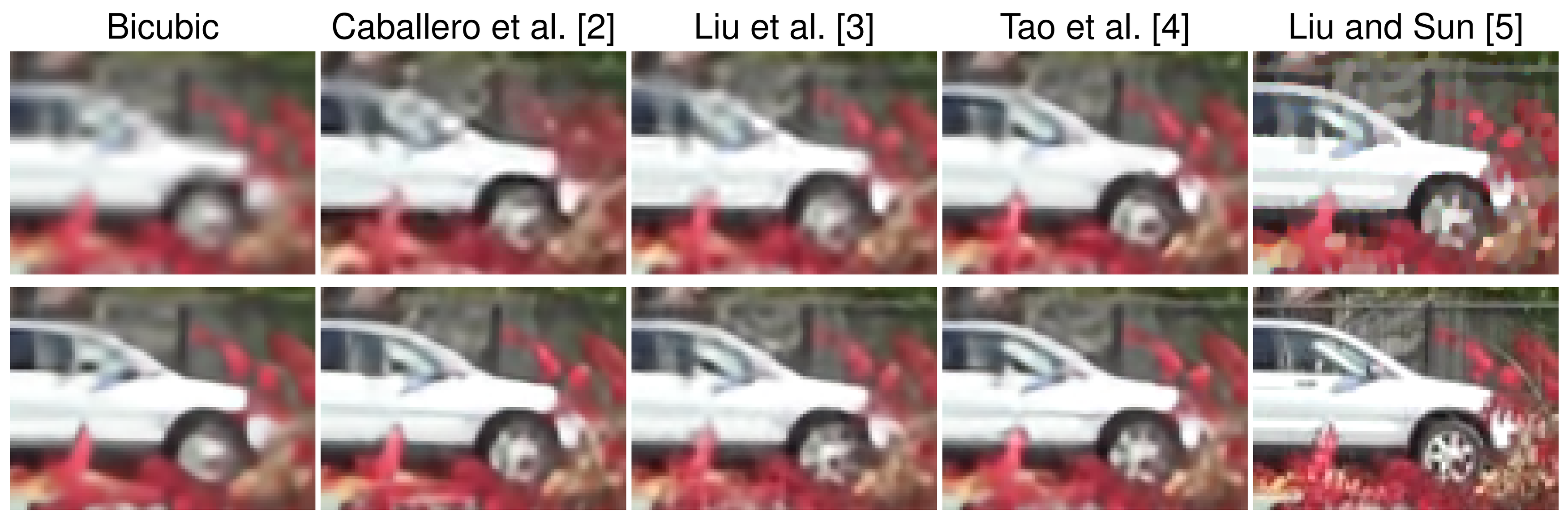
Network architecture



Architectures for super-resolution SRNet (top) and optical flow estimation FNet (bottom).

- We use fully convolutional neural networks operating in LR space for both super-resolution and optical flow estimation.
- For **super-resolution**, we use the **architecture of EnhanceNet [1]**, which consists of several residual blocks followed by upsampling layers.
- For **optical flow estimation**, we use the **standard encoder-decoder** style architecture to increase the receptive field.

Comparison with baselines and prior art on Vid4 dataset [5]

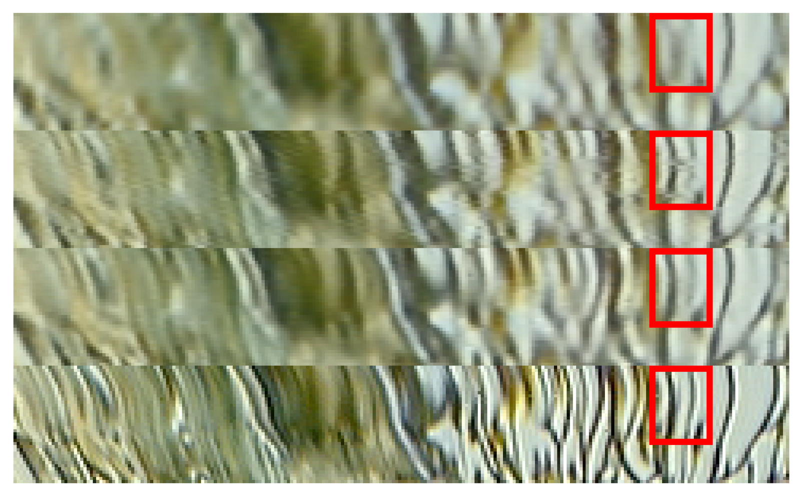


SISR 10-28 | VSR 10-128 | FRVSR 3-64 | FRVSR 10-128 | HR ground truth

Method	Bicubic	RAISR [6]	BRCN [7]	VESPCN [2]	$B_{1,2,3}+T$ [3]	DRVSR [4]	Bayesian [5]	SISR 10-128	VSR 10-128	FRVSR 3-64	FRVSR 10-128
PSNR	23.53	24.24	24.43*	25.35*	25.35	25.87	26.16	24.96	26.25	26.17	**26.69**
SSIM	0.628	0.665	0.662*	0.756*	0.738	0.772	0.815	0.721	0.803	0.798	**0.822**

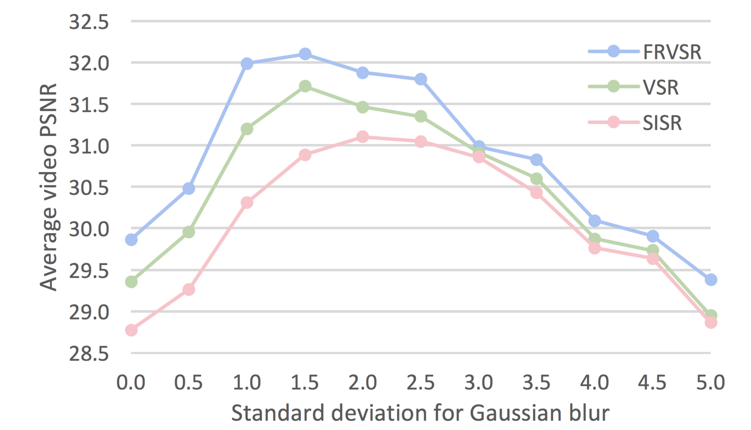
- **Baselines:**
  - **SISR:** Single image super-resolution using SRNet.
  - **VSR:** The previous and next input frames are warped onto the current frame using optical flow, and all the three frames are given as input to SRNet.
- **Results:**
  - Amongst prior art, [5] produces the best results, but their method uses a slow optimization procedure.
  - Our baseline VSR already produces results that are comparable to state-of-the-art.
  - The proposed FRVSR produces significantly better results (both visually and quantitatively) compared to state-of-the-art.
- **Temporal profiles** for a video from Vid4 dataset.
  - The VSR approach produces finer details than SISR, but its output still contains temporal inconsistencies (jitter in red box).
  - Only FRVSR is able to produce temporally consistent results while reproducing fine details.

SISR
VSR
FRVSR
HR



Ablation studies

- We use a dataset of ten 3-5s high-quality 1080p video clips downloaded from YouTube.
- **Effect of blur kernel size:**
  - The blur kernel size used for downsampling has a **significant effect on the performance** of the models.
  - Video SR methods (VSR and FRVSR) benefit from more aliased inputs compared to single image SR.



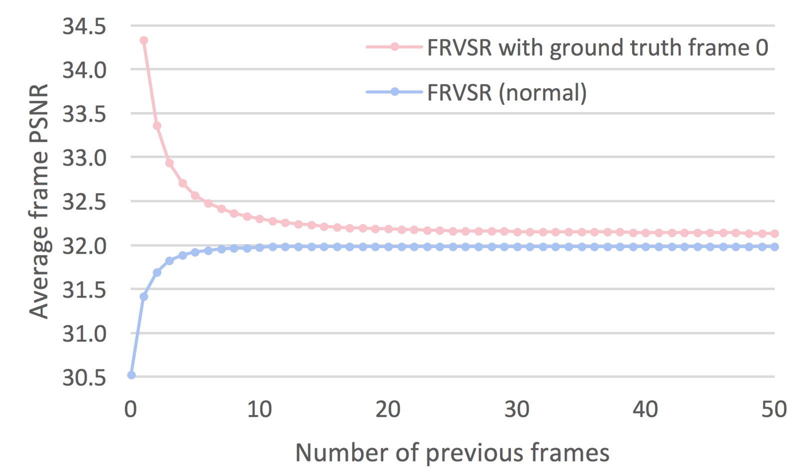
- **Performance under input degradations:**
  - Average PSNR under **Gaussian noise** (left two columns) and **JPEG artifacts** (right two columns).
  - **FRVSR achieves the best results.**

model	$\sigma = 0.025$	$\sigma = 0.075$	Q40	Q70
SISR	29.93	28.20	27.94	28.88
VSR	30.36	28.42	28.12	29.07
FRVSR	**30.84**	**28.62**	**28.30**	**29.29**

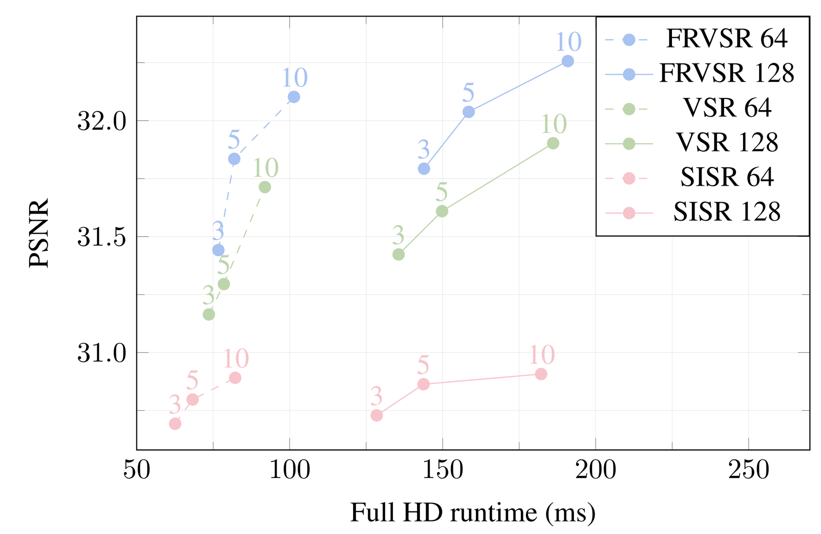
- **Effect of FRVSR training clip length:**
  - The PSNR has started to saturate with a length of 5 and going beyond 10 may not yield significant improvements.

$L = 3$	$L = 5$	$L = 10$
31.60	32.01	32.10

- **Range of information flow -** Performance of FRVSR as a function of the number of previous frames processed:
  - In the normal mode (blue), PSNR increases up to 12 frames, after which it remains stable.
  - When we have access to the first groundtruth HR frame (red), FRVSR propagates high-frequency details across a large number of frames and **performs better than the normal mode even after 50 frames.**



- **Network size -** Runtime vs PSNR for different numbers of convolution filters (64 / 128) and residual blocks (3 / 5 / 10) in SRNet.
  - Inference time is measured for generating a single 1080p frame with 4x upsampling on an Nvidia P100 GPU.
  - FRVSR achieves better results than both SISR and VSR with significantly smaller super-resolution networks and less computation time. For example, FRVSR with 5 residual blocks is **both faster and better** than VSR with 10 residual blocks.



References

[1] M. S. M. Sajjadi et al. EnhanceNet: Single Image Super-Resolution through Automated Texture Synthesis. ICCV 2017.
[2] J. Caballero et al. Realtime video super-resolution with spatio-temporal networks and motion compensation. CVPR 2017.
[3] D. Liu et al. Robust video super-resolution with learned temporal dynamics. CVPR 2017.
[4] X. Tao et al. Detail-revealing deep video super-resolution. ICCV 2017.
[5] C. Liu et al. A Bayesian approach to adaptive video super resolution. CVPR 2011.
[6] Y. Romano et al. RAISR: Rapid and accurate image super resolution. 2016.
[7] Y. Huang et al. Bidirectional recurrent convolutional networks for multi-frame super-resolution. NIPS 2015.