End-to-End Trainable Video Super-Resolution Based on a New Mechanism for Implicit Motion Estimation and Compensation

Xiaohong Liu†, Lingshi Kong†, Yang Zhou‡, Jiying Zhao‡, Jun Chen†

† McMaster University, Ontario, Canada, ‡ University of Ottawa, Ontario, Canada

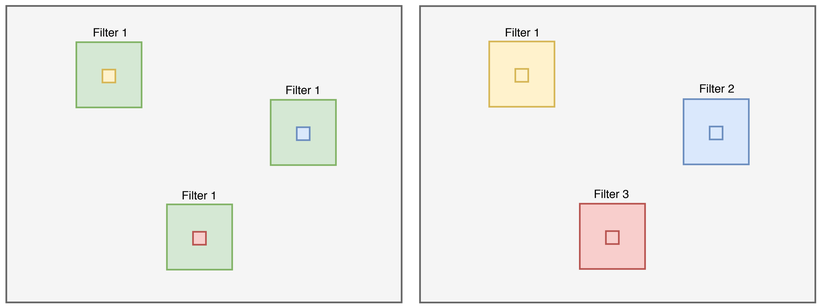
WACV 2020

Introduction

Video super-resolution aims at generating a high-resolution video from its low-resolution counterpart. With the rapid rise of deep learning, many recently proposed video super-resolution methods use convolutional neural networks in conjunction with explicit motion compensation to capitalize on statistical dependencies within and across low-resolution frames. Two common issues of such methods are noteworthy. Firstly, the quality of the final reconstructed HR video is often very sensitive to the accuracy of motion estimation. Secondly, the warp grid needed for motion compensation, which is specified by the two flow maps delineating pixel displacements in horizontal and vertical directions, tends to introduce additional errors and jeopardize the temporal consistency across video frames. To address these issues, we propose a novel dynamic local filter network to perform implicit motion estimation and compensation by employing, via locally connected layers, sample-specific and position-specific dynamic local filters that are tailored to the target pixels. We also propose a global refinement network based on ResBlock and autoencoder structures to exploit non-local correlations and enhance the spatial consistency of super-resolved frames. The experimental results demonstrate that the proposed method outperforms the state-of-the-art and validate its strength in terms of local transformation handling, temporal consistency as well as edge sharpness.

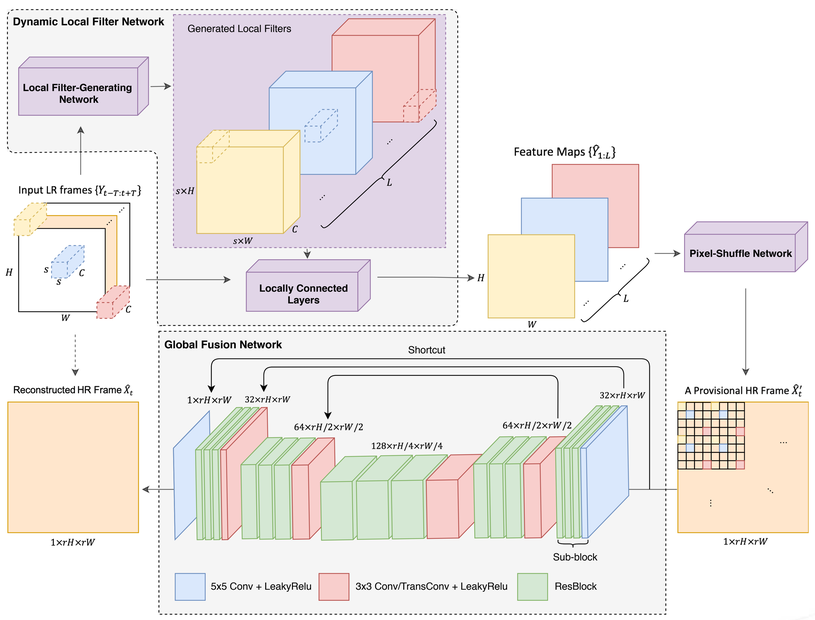
Proposed Method

In standard CNN layers, the filters (shown as green color) applied on three different pixel positions (shown as yellow, blue and red colors) have the same weights. In comparison, for LC layers, three different filters (shown as yellow, blue and red colors) are applied on the relevant pixel positions with unshared weights that are generated locally according to the imposed pixels.

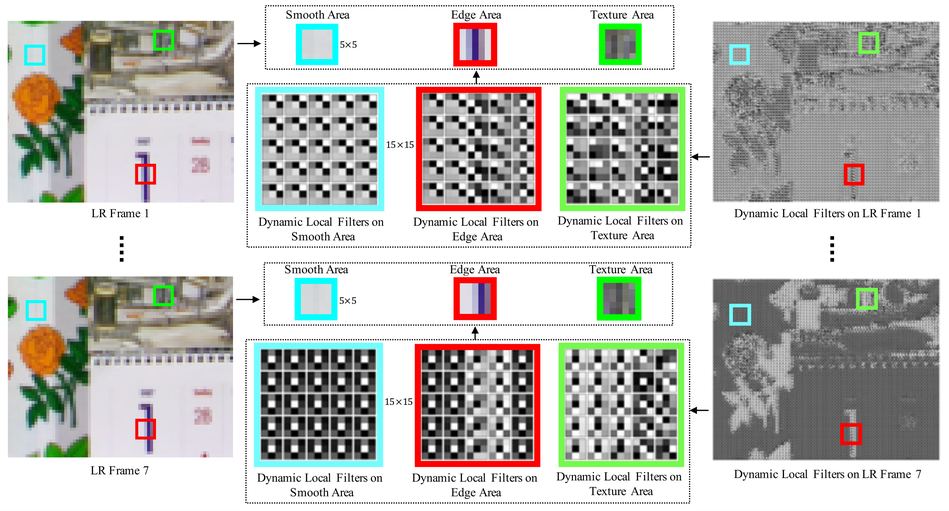


(a) Standard CNN layers (b) Locally connected layers

The Proposed locally-connected VSR:



Visualization of Dynamical Local Filters



The formulation of Dynamic Local Filter Network:

$$\hat{Y}_l^{(i,j)} = \sum_{m=i-d}^{i+d} \sum_{n=j-d}^{j+d} \sum_{k=t-T}^{t+T} \Theta_{i,j,l}^{(m-i+d+1,\,n-j+d+1,\,k-t+T+1)} \cdot Y_k^{(m,n)}$$

Spatial radius, Temporal radius

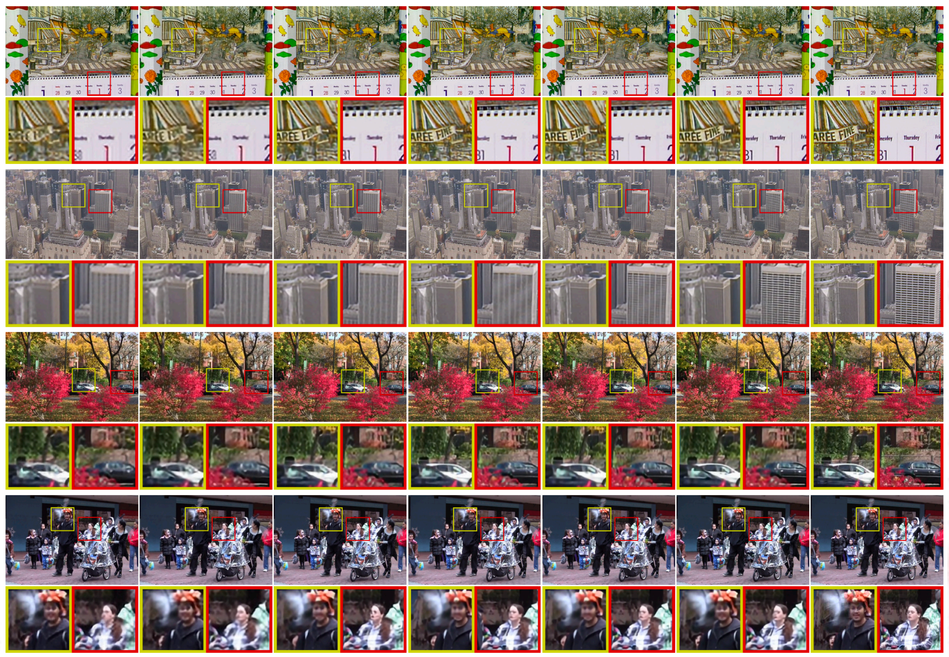
The produced lth feature map

Dynamic local filters

The kth input LR frame

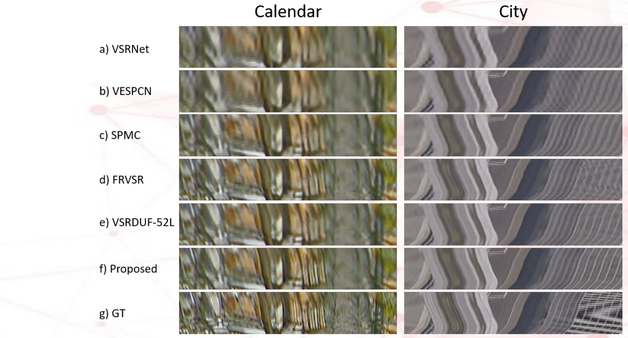
Experimental Results

Qualitative comparisons on Vid4 dataset:



(a) VSRNet [18] (b) VESPCN [1] (c) SPMC [33] (d) FRVSR [31] (e) DUF-52L [17] (f) Proposed (g) GT

Analysis of Temporal Consistency:



Quantitative comparisons and ablation study:

Vid4	Metric	*Bicubic*	*Bayesian* [24]	*VSRNet* [18]	*VESPCN* [1]	$B_{1,2,3}+T$ [25]	*SPMC* [33]	*FRVSR* [31]	*DUF-16L* [17]	*DUF-52L* [17]	*Proposed*
x3	**PSNR**	25.28	25.82	26.79	27.25	-	27.49	-	28.90	-	**29.51**
	SSIM	0.7329	0.8323	0.8098	0.8447	-	0.8400	-	0.8898	-	**0.8964**
x4	**PSNR**	23.79	25.06	24.84	25.35	25.39	25.52	26.69	26.81	27.34	**27.47**
	SSIM	0.6332	0.7466	0.7049	0.7557	0.7490	0.7600	0.8220	0.8145	0.8327	**0.8394**

Table 1: Quantitative comparisons on the Vid4 dataset with $r = 3, 4$.

SPMCS	Metric	*SPMC* [33]	*DUF-16L* [17]	*DUF-52L* [17]	*Proposed*
x3	**PSNR**	32.10	-	-	**33.91**
	SSIM	0.9000	-	-	**0.9358**
x4	**PSNR**	29.89	30.01	30.39	**30.66**
	SSIM	0.8400	0.8355	0.8646	**0.8711**

Table 2: Quantitative comparisons on the SPMCS dataset with $r = 3, 4$.

Vid4	Metric	w/o LC Layers	w/o GRN	w/ U-Net [30]	Our full model
x3	**PSNR**	27.27	28.13	29.20	**29.51**
	SSIM	0.8471	0.8752	0.8896	**0.8964**
x4	**PSNR**	25.45	26.20	27.29	**27.47**
	SSIM	0.7530	0.8134	0.8352	**0.8394**

Table 3: Quantitative comparisons on the Vid4 dataset for different variants of the proposed method with $r = 3, 4$.