Video Super-Resolution via Dynamic Local Filter Network

Yang Zhou, Xiaohong Liu, Lei Chen, *Student Member, IEEE*, Jiying Zhao, *Member, IEEE*

***Abstract*—Conventional Convolutional Neural Network (CNN) based video super-resolution (VSR) methods heavily depend on explicit motion compensation. Input frames are warped according to flow-like information to eliminate inter-frame differences. These methods have to make a trade-off between the distraction caused by spatio-temporal inconsistency and the pixel-wise detail damage caused by compensation. In this paper, we propose a novel video super-resolution method based on dynamic local filter network. Unlike traditional VSR techniques, our method implicitly performs motion estimation, compensation and fusion simultaneously via local convolutions with dynamically generated filter kernels. An optional autoencoder based refinement module is also proposed to sharpen edges and remove artifacts. The experimental results demonstrate that our method outperforms the best existing VSR algorithm by 0.53 dB in terms of PSNR, and provides superior visual quality.**

***Index Terms*—video super-resolution, locally-connected network, dynamic filter.**

I. Introduction

IMAGE super-resolution (SR) refers to the process that recovers a high-resolution (HR) image from one or a sequence of low-resolution (LR) images. It has been a long-standing fundamental research topic in image processing field. And it is widely applied to medical imaging, satellite imaging, surveillance fields and facilitation for image/video enhancement and text/object recongnition. In single image super-resolution (SISR), the HR image is supposed to be estimated from a single LR input, where the inherent similarities have to be exploited to recover the lost high-frequency details. In video super-resolution (VSR), different observations of a scene are available in the form of multiple LR input frames, therefore the explicit redundancy in common can be used to construct the HR image. However, how to combine and assemble the information from multiple frames have become a challenge, which limits extensive explorations of VSR.

The rapid development of neural network techniques in recent years has provided new possibilities for solving VSR problem and significant improvements in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). Convolutional neural network (CNN) was used for VSR in [1] [2] [3] and enhanced by custom operation layers [4] [5] or recurrent neural network (RNN) [6]. The VSRnet proposed by Kappeler *et al.* [1] uses optical flow technique to estimate the motion information among the input frames. The frames are then warped according to the motion information and fed to an SRCNN-inspired network [7] for super-resolving. The Video Efficient Sub-Pixel Convolutional Neural network (VESPCN) proposed by Caballero *et al.* [2] uses a motion compensation module, which is inspired by the Spatial Transform Network (STN) [8], to replace the optical flow calculating procedure. An ESPCN [9] based module is used for super-resolving. Tao *et al.* [4] improved the VESPCN with a Sub-Pixel Motion Compensation (SPMC) layer, which combines motion compensation and sub-pixel upsampling into one operation and an additional detail fusion module is appended to interpolate previous synthesized HR image. Makansi *et al.* [5] proposed a network combining with an improved VSRnet, which introduces a more advanced optical flow calculation procedure and an integrated compensation-upsampling operation. Their work is considered to offer the state-of-the-art VSR performance and will be denoted as Joint Upsampling and Backward Warping (JUBW) in this paper. Video frame interpolation problem is similar to VSR. Both of them need to analyze input video and exploit its spatio-temporal information. Furthermore, video frame interpolation can also be regarded as super-resolution on temporal dimension, hence Btz *et al.* [10] proposed a three dimensional approach for VSR. Our method is inspired by recent works from Niklaus *et al.* [11] [12], who leveraged dynamic filter network [13] to generate kernels for motion estimation and frame interpolation.

The rest parts of this paper are organized as follows. Section II introduces the motivation and contributions of our work. Section III details the architecture of the proposed VSR neural network. Section IV illustrates the experimental results and Section V concludes this paper.

II. Motivation and Contributions

Most of recent VSR works address the inter-frame inconsistency issue by a procedure consisting of three stages: 1) Estimate inter-frame motion to achieve flow-like information; 2) Compensate additional frames to reference frame by warping pixels according to their flow-like data, which makes all frames have similar contents; 3) Generate an HR image by fusing the compensated LR images. The content of the HR image is supposed be a super-resolved version of reference frame.

Recent works [4] [5] tried to merge parts of the above stages together to reduce the amount of duplicated computation. However, the compensation operations are still performed by manipulating pixels in spatial domain explicitly. The manipulation is performed according to the motion estimation, which is usually achieved by optical flow or affine-transform based techniques. For optical flow based methods, the final performance largely depends on the accuracy of flow data which is commonly considered computation-expensive and error-prone. Moreover, when the flow data involves non-integer coordinates, the compensated pixels have to be re-sampled. The resampling operation usually averages pixels

Y. Zhou, X. Liu, L. Chen and J. Zhao are with the School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward Ave., Ottawa, ON K1N 6N5 Canada (e-mail: yzhou152@uottawa.ca; xliu151@uottawa.ca; lchen148@uottawa.ca; jzhao@uottawa.ca)

978-1-7281-1295-4/18/$31.00 ©2018 IEEE

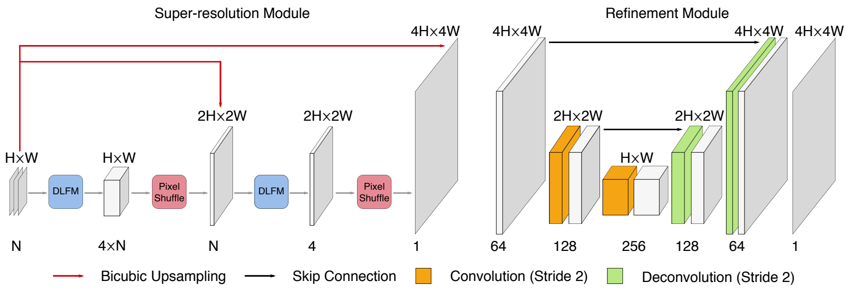

Fig. 1. Network architecture.

aggressively and makes edges blurry. Therefore details may not be preserved well. Although the STN based techniques avoid above disadvantages, only a limited number of motion patterns can be expressed by affine transformation.

The novelty of our proposed method is that, unlike previous works, we use a dynamic local filter network for implicit motion estimation, compensation and image fusion. The contribution is that our method outperforms the best existing VSR algorithm [5] by 0.53 dB in terms of PSNR, and provides superior visual quality.

III. Proposed Method

In this section, we describe the proposed network architecture which consists of two components: a super-resolution module and a refinement module, as illustrated in Fig. 1. The input LR images are fed into the SR module to synthesize an HR image. Then the HR image is processed by the refinement module to obtain the final result. We first introduce the dynamic local filter module (DLFM) which fuses multiple input frames to integrate their spatial differences. Then we describe the super-resolution module using the DLFM for super-resolving. We also incorporate the tailored detail fusion network [4] [5] as a refinement module, with which the SR results tend to have sharper edges and higher PSNR/SSIM values.

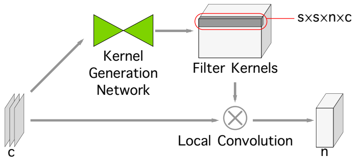

Fig. 2. The structure of dynamic local filter module.

A. Dynamic Local Filter Module

The dynamic local filter module (DLFM) consists of two parts: a kernel generation network (KGN) which produces filter kernels based on the input images, and the generated filters are applied to the input images via local convolutions, as illustrated in Fig. 2.

The KGN takes an input $I \in \mathbb{R}^{c \times h \times w}$, where c, h and w are number of channels, height and width of the input image I respectively, and generates filter kernels F_θ parameterized by $\theta \in \mathbb{R}^{s \times s \times n \times c \times h \times w}$, where s is the kernel size of the generated filters, and n is the number of output channels. Theoretically, this network can be any differentiable architecture. In this work, we propose an autoencoder [14] based KGN inspired by the work in [12]. All the convolution layers use a kernel size of 3×3. The average pooling layers are with a kernel size of 2×2. Rectified Linear Units (ReLU) are used for all convolution layers except the last one. Other parameters are noted in Fig. 3.

The local convolution is a translation-variant convolution and also known as locally-connected network. It applies different filter kernels to corresponding patches according to its position in the image. Suppose I_n is the n-th image in the input sequence. For position (i,j), $P_n^{(i,j)}$ is defined as a patch of size $(2K+1, 2K+1)$ centered at (i,j) in I_n, which is

$$P_n^{(i,j)} = \begin{bmatrix} I_n^{(i-K,j-K)} & \cdots & I_n^{(i+K,j-K)} \\ & \ddots & \\ \vdots & I_n^{(i,j)} & \vdots \\ & & \ddots \\ I_n^{(i-K,j+K)} & \cdots & I_n^{(i+K,j+K)} \end{bmatrix}. \quad (1)$$

A set of unique kernels $\theta_n^{(i,j)}$ is generated for $P_n^{(i,j)}$, which is

$$\theta_n^{(i,j)} = \begin{bmatrix} w_{n,(i,j)}^{(i-K,j-K)} & \cdots & w_{n,(i,j)}^{(i+K,j-K)} \\ & \ddots & \\ \vdots & w_{n,(i,j)}^{(i,j)} & \vdots \\ & & \ddots \\ w_{n,(i,j)}^{(i-K,j+K)} & \cdots & w_{n,(i,j)}^{(i+K,j+K)} \end{bmatrix}. \quad (2)$$

A local convolution operation $F_c(\theta_n, I_n)$ is defined as

$$F_c(\theta_n, I_n) = \begin{bmatrix} \theta_n^{(0,0)} \odot P_n^{(0,0)} & \cdots & \theta_n^{(W-1,0)} \odot P_n^{(W-1,0)} \\ \vdots & \ddots & \vdots \\ \theta_n^{(0,H-1)} \odot P_n^{(0,H-1)} & \cdots & \theta_n^{(i,j)} \odot P_n^{(W-1,H-1)} \end{bmatrix}, \quad (3)$$

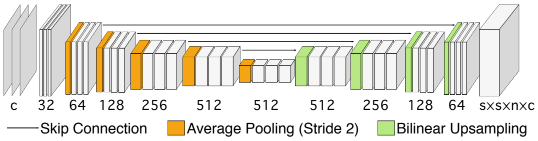

Fig. 3. Kernel generation network.

where W, H are width and height of I_n. The operator $\odot$ is the summation of element-wise product of two matrices, which is defined as

$$A \odot B = \sum_{i=0}^{M} \sum_{j=0}^{N} A_{ij} \circ B_{ij}, \tag{4}$$

where A and B are two $M \times N$ matrices, and $\circ$ is the element-wise product.

In the proposed method, for the local convolution result of input image I_n,

$$Y_n = F_c(\theta_n, I_n), \tag{5}$$

the following relation is supposed to be satisfied.

$$Y_{ref} \simeq Y_n, \tag{6}$$

where ref is the index of the reference frame. Spatial differences among I_n are compensated by the local convolution.

B. Super-Resolution Module

With DLFM, common features can be extracted from input frames without distractions caused by inter-frame differences. A DLFM with an appended pixel-shuffling layer [9] should have the ability to perform VSR. However, since the KGN generated filters have fixed kernel size, the maximum motion magnitude that DLFM can adapt to is upper-limited. Although this problem can be solved by increasing filter kernel size, the network would become impractical because of the rapid increasing computation cost.

Larger filter kernel size can be achieved by using separable convolution in [12]. However, compared with video interpolation, most of computations for SR are performed in LR domain. The regions with $s \times s$ pixels in LR domain correspond to $\alpha s \times \alpha s$ pixels patches in HR domain, where α is the upsampling scale factor. Meanwhile, since the separable convolution aims to reduce the number of parameters of a 2D filter by using the inner-product of two 1D filters, it is inevitable to lose representation flexibility and make performance sub-optimal. These properties make separable convolution unsuitable for SR.

In our proposed method, multiple DLFMs are cascaded to increase the receptive field. Each of them uses a relatively small kernel size, which is 3×3, for efficiency. And inspired by [15], [16] and [17], pixel-shuffle layer with ReLU are appended to each DLFM to form a progressive SR procedure. We also introduce skip connections [18] which add the bicubic-upsampled reference frame to the outputs of pixel-shuffle layers. We denote this network as the super-resolution module. Fig. 1 illustrates a progressive $4\times$ SR module which consists of two $2\times$ SR stages.

C. Refinement Module

While the SR module has the capability to generate satisfactory results, we observed that an additional refinement stage is still beneficial. We propose a refinement module, inspired by the tailored detail fusion network [4] [5]. The module adapts an autoencoder-style [14] architecture with skip connections. The first and last layer use a kernel size of 5×5. The other convolution layers have a kernel size of 3×3. The deconvolution layers are with a kernel size of 4×4. ReLUs are used for each layer. Other parameters are illustrated in Fig. 1. Experimental results show that the refinement can achieve considerable improvements for the final results which will be described in the next section.

IV. Experimental Results

In the experiments, we collected 3022 720p video clips online as training dataset and normalized the pixel values to $[-1.0, 1.0]$. The super-resolution scale factor is fixed to 4. We use Adam [19] solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and various learning rates. $\mathcal{L}_1$ and $\mathcal{L}_2$ loss functions are incorporated to measure the differences between super-resolved image I_{sr} and corresponding ground truth I_{gt}. The mini-batch size is fixed to 1 due to hardware limitations. The training procedure are divided to 3 stages. Firstly, only the SR module is trained with $\mathcal{L}_1$ loss and learning rate 0.0001 for about 600k iterations. Then the parameters of the SR module are freezed. The refinement module is attached and trained with $\mathcal{L}_1$ loss and learning rate 0.0001 for about 120k iterations. Finally, the whole network is trained jointly with learning rate 0.00001 for 480k iterations using $\mathcal{L}_2$ loss function.

We chose the widely-used VID4 [20] as the testing dataset. We compared our method with an SISR method, SRCNN [7] and five recent VSR methods: BayesSR [20] , VSRNet [1], VESPCN [2], SPMC [4] and JUBW [5]. The BayesSR is a traditional VSR algorithm. The other five methods are based on neural network. The JUBW [5] is a state-of-the art algorithm and it provides the best VSR performance so far in terms of PSNR and visual quality. In our experiments, all the measurements are performed on the luminance channel. For visual comfort, the chrominance channels of images are upsampled by bicubic interpolation. In Table I, our method uses 3 consecutive frames. The quantitative and visual comparisons are illustrated in Table I and Fig. 4. The experimental result of SRCNN is from [2]. The experimental results of BayesSR, VSRNet, VESPCN, SPMC are from [4]. The experimental results of JUBW result is from [5]. Our method outperforms the JUBW by 0.53 dB in terms of PSNR. Due to the absence of SSIM metric in the original paper on JUBW, we did not compare our method with JUBW in terms of SSIM. From the close-up images, we see that the texture details and object edges of the original video are better recovered by the proposed method.

For VSR problem, the multiple input frames can be viewed as various observations of the ground truth. A proper VSR method should be able to generate superior results when extra observations are provided. To prove that our proposed method can exploit information from additional input frames,

TABLE I
QUANTITATIVE COMPARISON WITH OTHER VSR METHODS

Method	SRCNN	BayesSR	VSRNet	VESPCN	SPMC	JUBW	Proposed (without Refinement)	Proposed (with Refinement)
PSNR / SSIM	24.68 / 0.72	24.42 / 0.72	22.81 / 0.65	25.35 / 0.76	25.52 / 0.76	25.85 / -	26.04 / 0.80	**26.38 / 0.81**

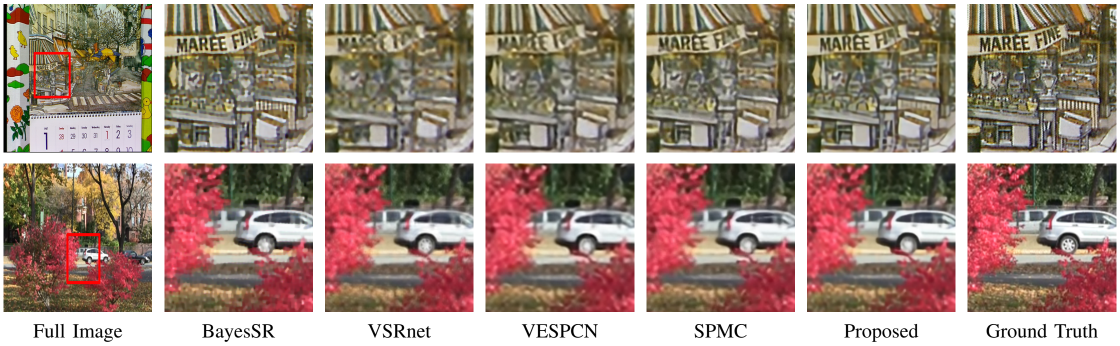
Full Image BayesSR VSRnet VESPCN SPMC Proposed Ground Truth

Fig. 4. Visual comparison with previous VSR methods.

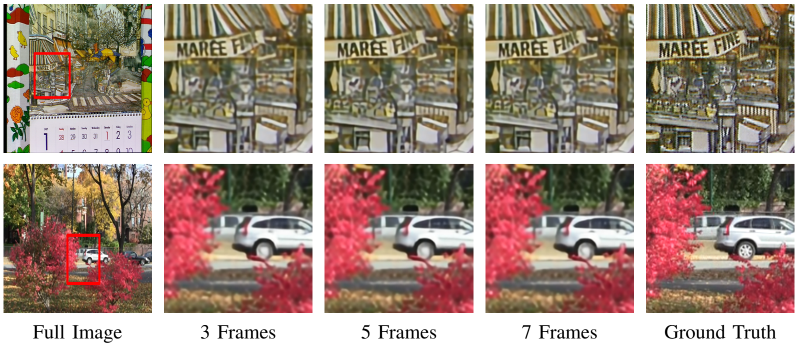
Full Image 3 Frames 5 Frames 7 Frames Ground Truth

Fig. 5. Visual comparison with inputs of various lengths.

TABLE II
QUANTITATIVE COMPARISON OF VARIOUS INPUT LENGTHS AND THE EFFECT OF REFINEMENT IN PSNR/SSIM.

Input Length	With Refinement	Without Refinement
3 Frames	26.38 / 0.81	26.04 / 0.80
5 Frames	26.51 / 0.82	26.17 / 0.81
7 Frames	26.52 / 0.82	26.20 / 0.81

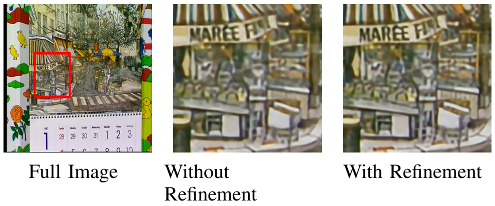
Full Image Without Refinement With Refinement

Fig. 6. Visual comparison of the effect of refinement with 3 frames input.

we evaluate our network on 3, 5 and 7 consecutive frames. Experimental results show that more consecutive inputs do lead to results with sharper edges as well as higher PSNR/SSIM values, as illustrated in Fig. 5 and Table II.

Unlike the SPMC and JUBW, the refinement module in our method is not used for interpolation. The synthesized HR images without refinements are still complete, which means the refinement module in this work is optional. Table II and Fig. 6 indicate that the refinement is beneficial to the final results. However, the stand-alone SR module can still outperform other VSR methods in terms of PSNR/SSIM values, as illustrated in Table I.

Our experiment platform is equipped with a single NVIDIA GTX 1080Ti graphical adapter. The program took 4.727 seconds to process all the 147 frames in VID4 with 3 frames as inputs.

V. CONCLUSION

In this paper, we proposed a novel video super-resolution framework which consists of a dynamic local filter based video SR module and an autoencoder based refinement module. The experimental results demonstrated that our proposed framework outperforms the existing best-performing method by 0.53 dB in terms of PSNR and provides superior visual quality. As for future work, we are trying to further improve the model efficiency and explore more applications for the DLFM.

REFERENCES

[1] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, June 2016.

[2] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 531–539.

[4] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] O. Makansi, E. Ilg, and T. Brox, "End-to-end learning of video super-resolution with motion compensation," in *German Conference on Pattern Recognition (GCPR)*, 2017.

[6] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[8] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025.

[9] W. Shi, J. Caballero, F. Huszr, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1874–1883.

[10] M. Btz, F. Brand, A. Eichenseer, and A. Kaup, "Motion compensated frame rate up-conversion using 3d frequency selective extrapolation and a multi-layer consistency check," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 1452–1456.

[11] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] ——, "Video frame interpolation via adaptive separable convolution," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 667–675.

[14] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2802–2810.

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

[16] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *2015 International Conference on Learning Representations (ICLR)*, 2015.

[20] C. Liu and D. Sun, "A bayesian approach to adaptive video super resolution," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 209–216.