Low-rank Constrained Super-Resolution for Mixed-Resolution Multiview Video

Shao-Ping Lu, *Member, IEEE,* Sen-Mao Li, Rong Wang, Gauthier Lafruit, Ming-Ming Cheng, *Senior Member, IEEE,* and Adrian Munteanu, *Member, IEEE*

***Abstract*—Multiview video allows for simultaneously presenting dynamic imaging from multiple viewpoints, enabling a broad range of immersive applications. This paper proposes a novel super-resolution (SR) approach to mixed-resolution (MR) multiview video, whereby the low-resolution (LR) videos produced by MR camera setups are up-sampled based on the neighboring HR videos. Our solution analyzes the statistical correlation of different resolutions between multiple views, and introduces a low-rank prior based SR optimization framework using local linear embedding and weighted nuclear norm minimization. The target HR patch is reconstructed by learning texture details from the neighboring HR camera views using local linear embedding. A low-rank constrained patch optimization solution is introduced to effectively restrain visual artifacts and the ADMM framework is used to solve the resulting optimization problem. Comprehensive experiments including objective and subjective test metrics demonstrate that the proposed method outperforms the state-of-the-art SR methods for MR multiview video.**

***Index Terms*—Multiview video, mixed-resolution, super-resolution, low-rank, ADMM optimization.**

I. Introduction

Multiview videos are simultaneously acquired from different viewpoints, which allows for presenting dynamic visual content from multiple viewing angles at the same time. This functionality renders multiview video as a powerful representation format enabling a variety of immersive applications including free viewpoint video, 3D gaming, interactive teleconferencing among others. Various increasingly popular high-resolution (HR) 3D displays, aim at producing highly realistic viewing perception and immersive visual experience, but they require up to hundreds of HR multiview videos to operate.

One of the most critical challenges in multiview video acquisition and display systems is to acquire, process and transmit the massive amounts of multiview video data. The data rates produced by such HR acquisition setups are currently prohibitive for broad scale deployment, so a solution to this problem is required. In this context, developing practical approaches to capture low-resolution (LR) multiview videos followed by up-sampling them to HR has received extensive attention [1]–[3]. Such setups are however often impaired by the lack of fine details in the upsampled videos. The alternative is given by mixed-resolution (MR) multiview video capturing setups that combine HR and LR video acquisition. Such systems incorporate multiple HR cameras among which LR cameras are arranged to capture LR videos. These novel setups produce both HR and LR video streams which enable to significantly decrease the data transmission and storage requirements compared to full HR video acquisition.

For MR multiview videos, because the display needs to process and display videos with a uniform HR resolution, reconstructing the LR videos at the same resolution as that of the HR videos captured by the HR cameras becomes an important issue. This can be seen as a special kind of super-resolution (SR) problem, for which the following two problems should be considered: 1) *how to reconstruct the missing texture details from the other input HR camera views*; and 2) *how to restrain visual artifacts since the reference view images are not exactly the same as the targeted one*. The state-of-the-art SR approaches cannot effectively solve the above-mentioned problems. On one hand, existing multiview video SR techniques can neither effectively learn from the HR patches of the neighboring HR cameras [3], nor restrain noise [1], [2] from the great redundancy of textured patches. On the other hand, most of image (or video) based SR approaches [4]–[6] do not function well in our MR multiview system, due to the lack of analysis of the spatio-temporal correlation of videos captured from different cameras.

This paper introduces a low-rank constrained SR approach to MR multiview video. We consider that the matched patches of all involved views construct a manifold structure in a high-dimensional space, and introduce a local linear embedding (LLE) based texture reconstruction optimization to learn the HR patches from the neighboring camera views. Furthermore, we construct a low-rank constrained global optimization scheme to restrain visual artifacts, to overcome the imperfect matching of patches and the consequent fitting errors of such patches to the target patches. Our low-rank constrained SR approach is especially effective at exploiting the redundant information of patches of all cameras when performing SR in MR multiview video; extensive experimental results will demonstrate the effectiveness of our approach.

In summary, our main technical contributions are as follows:

- We introduce a low-rank representation based SR frame-

S-P. Lu, S-M. Li, R. Wang and M-M. Cheng are with TKLNDST, CS, Nankai University, China. The first two authors contributed equally (corresponding to email: slu@nankai.edu.cn).

G. Lafruit is with LISA department, Universite Libre de Bruxelles (email: gauthier.lafruit@ulb.ac.be).

A. Munteanu is with Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB) (email: acmuntea@etrovub.be).

Manuscript received xx xx, 20xx; revised xx xx, 20xx and xx xx 20xx; accepted xx xxxx; date of current version March xx, 20xx. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hairong Qi.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2020.3042064

work for MR multiview video. Our solution analyzes the redundancy of the matched spatio-temporal patches, and effectively reconstructs the fine textures using low-rank based optimization.
- We solve the SR problem with a well-defined initialized reconstruction using local linear embedding, which effectively learns the high-frequency details from other reference cameras.
- We design a multiview SR evaluation metric accounting for the global gradient distribution in multiview images.

II. Related Work

The image and video SR problem is a fundamental topic in many research communities (detailed survey work can be found in [7]–[9]). Here we briefly review some previous papers that are related to our work.

The very early SR methods include some well-known linear or non-linear filters, such as Nearest-Neighbor, Bilinear, Bicubic, Lanczos and their variants. Kopf *et al.* [4] introduce joint bilateral upsampling to further consider each pixel's spatial coordinates and color intensities. These methods easily suffer from various artifacts (the most common being blurriness), due to their simple filtering kernels or weighting factors.

Besides these relatively simple filtering methods, there are a great number of *single image based SR approaches* injecting various prior knowledge to generate the solution, including regularization constraints, such as gradient statistics, local texture similarity, geometry prior, etc. For instance, in [10] two piece-wise continuous functions are used to approximate the gradient density distribution of images in an iterative deconvolution of the upsampling. Gao *et al.* [11] use histograms of oriented gradients of LR patches to find good neighbors as well as reconstruction weights. A parametric gradient profile model of the image structures is presented in [12]. In [6] a directional standard-deviation-based weights selection is formulated to model the local geometric duality and the non-local similarity of images. To suppress the effect of pseudo-edges, Yuan *et al.* [13] introduce an adaptive total variation model to favor the piece-wise constant property of flat regions. In [14] affine transformations are supported to detect local shape variants of the same LR image, and the internal patch search space is greatly extended.

Many SR methods take *multiple external images or consecutive video frames* into account. In order to handle motion blur, noise, outliers and other effects when upsampling from a set of LR images, Farsiu *et al.* [15] introduce a L_1 norm minimization and robust regularization strategy for data fusion. In [16] a mixed model of L_1 and L_2 norms is created to detect multi-orientation and multi-order variations of multiple frames in a LR video. The method proposed in [17] approximates each pixel of the video sequence with a 3D local Taylor series, and computes the local motion of each pixel when performing SR. Adam *et al.* [18] further consider to improve the recognition of small moving objects, and a sub-pixel precise object boundary model is constructed to solve the intensity of the target object. This class of SR methods exploits also the temporal information compared to single-image based techniques, yielding better reconstruction results. However, in multiview video based SR problem, more video channels should be processed to utilize the correlation between the camera views.

In the last decade, many SR methods that rely on *sparse representation techniques* have been proposed. One of the pioneering works in this direction is presented in [5], where the coefficients of the sparse representation of LR image patches are used to generate the HR textures. Schulter *et al.* [19] propose to map the LR and HR patches using random forests. In [20] a set of auto-regressive models are used to learn the dataset of *example image patches*. The method in [21] proposes to learn a parametric sparse prior of HR images from the training set and the input LR image. There are also some schemes that operate in the Fourier domain [22], Wavelet domain [23], [24], or directly process the compressed data stream [25]. In general, it is difficult for the above-mentioned SR methods to deal with complex textures in natural images, which renders the SR problem as ill-posed when using a single image as input.

Recently, Convolutional Neural Networks (CNNs) have been extensively applied to the SR problem (see e.g. the survey in [9], [26]). Among them, one of the pioneering works in this direction is presented in [27], where a specific SR Convolutional Neural Network (SRCNN) is proposed to learn the LR-to-HR nonlinear mapping function. Many subsequent methods based on this solution have been recently proposed, such as using the sparse priors [28] or generative adversarial networks (GAN) [29]. Tai *et al.* [30] introduce Deep Recursive Residual Network (DRRN) to simplify training of very deep networks. In [31] a residual dense network is presented to learn the features of the original image from all the convolutional layers. In [32] a deep Laplacian Pyramid Super-Resolution Network is proposed to share parameters both across and within pyramid levels.

There are several SR works [1]–[3] that focus on *MR based multiview video*. In [33] a 2D piece-wise auto-regressive strategy is used to interpolate each target pixel. Li *et al.* [34] employ a kernel regression model to analyze the local image structure, and use non-local means to exploit the similarity between different views. Because the lighting would be different among camera views [35], Jin *et al.* [2] propose to synthesize the virtual view, and compensate the luminance difference between views. However, view synthesis is still a very challenging task [36], [37]. The method in [3] focuses on recovering the blurred views using the associated depth maps. In most of multiview video acquisition systems, the depth information is recovered from its LR version [38], which easily suffers from reconstruction errors. Richter *et al.* [1] correct the projection errors introduced by inaccurate depth information when synthesizing the high-frequency content. Interestingly, some researchers also seek to enhance the resolution of light field [39], [40] and the recovered depth information [41]. Different from these methods, our approach does not rely on the inaccurate depth information which might dramatically result in reconstruction errors of the HR image. We focus on exploiting the correlation between different camera views in the MR system, and use the low-rank constraints of patches

to guide the target texture reconstruction process.

III. Proposed Approach

A. Problem Formulation

In many image SR solutions, it is assumed that the LR image is a degraded version of the HR image, and such degradation procedure is mixed by blurring, down-sampling and noise interference (see examples in [5], [10], [21], [42]). Formally, the LR image $I^{LR} \in \mathbb{R}^{N_l}$ is obtained by the degradation of the corresponding HR image $I^{HR} \in \mathbb{R}^{N_h}$, and such process can be expressed as

$$I^{LR} = DBI^{HR} + v, \tag{1}$$

where $B \in \mathbb{R}^{N_h \times N_h}$ is the blurring filter, $D \in \mathbb{R}^{N_l \times N_h}$ is a downsampling operator, and v is additive noise. In the above representation, N_l and N_h are the sizes of the one column of the LR and HR images, respectively. Recovering the desired HR image from the observed LR one is a typical *inverse problem*. Let we denote the reconstructed HR image as $\hat{I}^{HR} \in \mathbb{R}^{N_h}$, such inverse problem can be expressed as the minimization of the following formulation:

$$\hat{I}^{HR} = \arg\min_{I^{HR}} (\left\| I^{LR} - DBI^{HR} \right\|_2^2 + \eta \mathcal{R}(I^{HR})). \tag{2}$$

Here $\mathcal{R}(I^{HR})$ is a regularization term to model the prior knowledge of the HR image, and the parameter η is used to balance the fidelity term and regularization term.

Our work shares the above-mentioned SR formulation, and thus we follow the well-used assumptions that v is a zero-mean Gaussian noise and B is a known Gaussian blurring operator. In the following we introduce how to construct our SR framework and numerically solve it for MR multiview video.

B. System Overview

Different from single image or video SR methods [43], [44] that only consider either the spatial or the temporal correspondence within one view, exploiting the joint spatio-temporal correlation of videos captured from different cameras is of particular importance in MR multiview systems. Hence, we need to effectively learn HR textures from both neighboring camera views (plus the reconstructed previous frame) and regress visual artifacts due to imperfect content matching of patches.

Suppose I is a frame in MR multiview video. At time t the input LR image of the target camera c^{tar} is $I^{LR}_{(t,c^{tar})}$. Our purpose is to reconstruct its desired HR version $I^{HR}_{(t,c^{tar})}$ by making use of the previously generated frame $I^{HR}_{(t-1,c^{tar})}$ and the HR frames from its neighboring camera views, such as $I^{HR}_{(t-1,c_1)}$, $I^{HR}_{(t-1,c_2)}$, $I^{HR}_{(t,c_1)}$, $I^{HR}_{(t,c_2)}$ and so on. We define $\hat{I}^{HR}_{(t,c^{tar})}$ as the reconstructed HR image of $I^{LR}_{(t,c^{tar})}$. More details on the introduced notations are given in Tab. I.

Our proposed framework is shown in Fig. 1. To initialize the system well, we firstly get the approximate HR version of the LR video frame $I^{LR}_{(t,c^{tar})}$, where the low-frequency components of the image are coarsely preserved. This can be easily achieved using the Bicubic upsampling filter, such as in [5], [10]. Then, we use the LLE optimization framework to learn the spatio-temporal HR details from the reconstructed previous HR image and other HR frames of the neighboring views. After that, we construct a low-rank regularization method to effectively perform artifact regression. Finally, we use an iterative optimization method to get a converged solution $\hat{I}^{HR}_{(t,c^{tar})}$, which is the t-th output frame and also is the reference HR frame for the reconstruction of the next frame.

TABLE I
Introduced notations.

Notation	Implication
c^{tar}	Target HR camera view (with LR input video)
C	$C = \{c_i \mid i = 1, 2, ...\}$, is an aggregate of reference HR camera views
$I^{LR}_{(t,c^{tar})}$	LR image from camera c^{tar} at time t
$I^{HR}_{(t,c^{tar})}$	Target HR image of $I^{LR}_{(t,c^{tar})}$
$\hat{I}^{HR}_{(t,c^{tar})}$	Reconstructed HR image of $I^{LR}_{(t,c^{tar})}$
q_i	The i-th $d_M \times d_M$ patch from $\hat{I}^{HR}_{(t,c^{tar})}$ for learning HR details
g_l	The l-th $d_L \times d_L$ patch from $\hat{I}^{HR}_{(t,c^{tar})}$
$g_{(l,k)}$	The k-th $d_L \times d_L$ nearest neighboring patch of g_l from $\hat{I}^{HR}_{(t,c^{tar})}$
$I^{HR}_{LF(t,c^{tar})}$	Low-frequency version of $I^{HR}_{(t,c^{tar})}$
$\hat{I}^{HR}_{LF(t,c^{tar})}$	Reconstructed low-frequency version of $\hat{I}^{HR}_{(t,c^{tar})}$
p_i	The i-th $d_M \times d_M$ patch from $\hat{I}^{HR}_{LF(t,c^{tar})}$
$I^{HR}_{(t,c_i)}$	Reference HR image from view c_i of HR camera at time t, where $c_i \in C$
$q_{(i,j)}$	The j-th $d_M \times d_M$ nearest neighboring patch of q_i from $I^{HR}_{(t,c_i)}$
$I^{HR}_{LF(t,c_i)}$	low-frequency version of $I^{HR}_{(t,c_i)}$
$p_{(i,j)}$	The j-th $d_M \times d_M$ nearest neighboring patch of p_i from $I^{HR}_{LF(t,c_i)}$
W	Correlation matrix between p_i and $p_{(i,j)}$

C. Learning spatio-temporal HR details from the neighboring HR views

In order to learn the HR details from the neighboring HR views, our solution (i) takes the texture patches of multiview videos to form a manifold structure in a high-dimensional feature space, and (ii) expresses each patch as a linear combination of a few nearest neighbors in the feature space. This representation could be optimally solved by the LLE framework [45]. On the other hand, the low-frequency visual textures are the most important components in the image domain, and such components are coarsely preserved when images are degraded with Gaussian blurring and/or down-sampling. We assume that the patch-level manifold representation structure of multiview videos should be preserved under the above-mentioned degradation. Therefore, once we

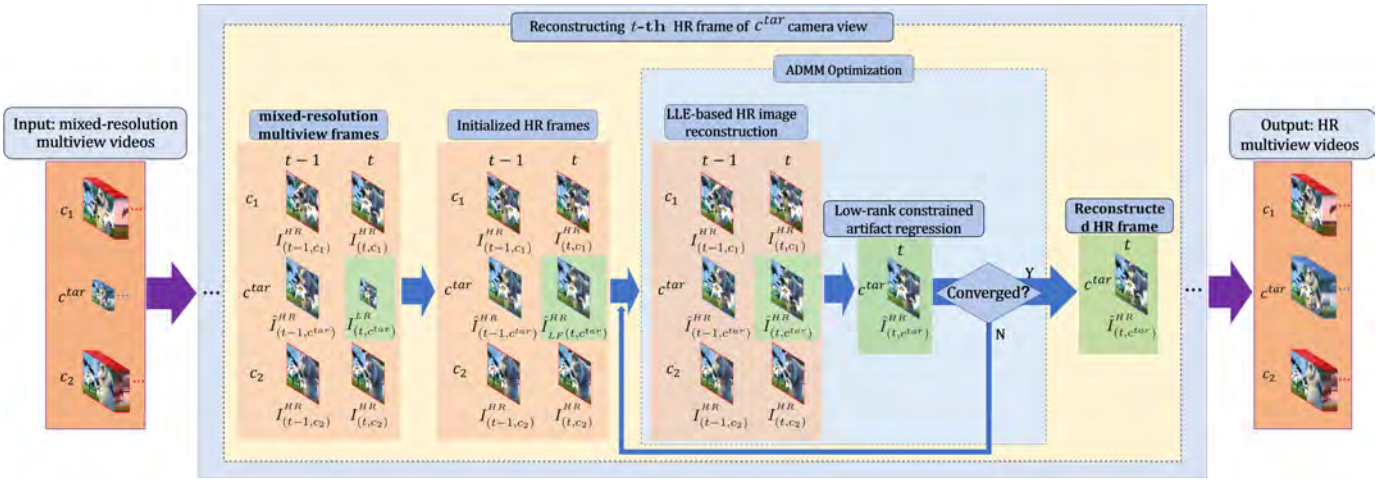


Fig. 1. System overview. We take the mixed-resolution multiview videos as input and generate all HR videos. To reconstruct the t-th HR frame $\hat{I}^{HR}_{(t,c^{tar})}$, we learn the textures by representing each patch using some HR images from the neighboring views and the reconstructed previous HR frame of the same view. The low-rank constrained regularization of the image patch is also introduced to regress spatio-temporal artifacts.

obtain the LLE-based representation parameters of all low-frequency patches of the target HR image, we can take them to reconstruct the target all-frequency HR image with the HR patches of the neighboring HR images.

We now consider reconstructing $\hat{I}^{HR}_{(t,c^{tar})}$ on the image patch level. As mentioned before, we can utilize the HR images from the neighboring HR views and the reconstructed previous HR image $I^{HR}_{(t-1,c^{tar})}$ in the same view; we denote such HR images as $I^{HR}_{(\cdot,\cdot)}$, what we call *reference images*, and we thus define $I^{HR}_{LF(\cdot,\cdot)}$ as their corresponding low-frequency images.

For each patch p_i with the size of $d_M \times d_M$ in the initialized low-frequency image $\hat{I}^{HR}_{LF(t,c^{tar})}$, its most similar patch from the j-th reference image is $p_{(i,j)}$. Here the similarity between two patches is computed by using the ℓ_2 color distance of their corresponding RGB pixels. We then let $i = 1, ..., M$, and M is the number of patches when the image $\hat{I}^{HR}_{LF(t,c^{tar})}$ is separated with the patch size $d_M \times d_M$. Once the pixels of patch $p_{(i,j)}$ are reordered into one column, we build a reference matrix $\mathcal{P}_i = [p_{(i,1)}, p_{(i,2)}, ..., p_{(i,n)}]$, where n is the number of reference images involved in the system. Therefore, with an error term Δ, the patch p_i can be linearly represented by its most similar patches $p_{(i,j)}$ from the n reference images as:

$$p_i = \sum_{j=1}^{n} \omega_i^j p_{(i,j)} + \Delta, \; s.t. \sum_{j=1}^{n} \omega_i^j = 1, \tag{3}$$

where ω_i^j is the weight factor of the corresponding patch from the reference image j. Furthermore, we define a local similarity weight vector $\omega_i = [\omega_i^1, \omega_i^2, ..., \omega_i^n]^\top \in \mathbb{R}^{n\times 1}$, so the expected correlation between p_i and $\mathcal{P}_i$ can be calculated by solving the following constrained least square problem:

$$\arg\min_{\omega_i} \|p_i - \mathcal{P}_i\omega_i\|_2^2, s.t. \mathbf{A}^\top \omega_i = 1, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{n\times 1}$ is a vector of all 1. Therefore, the reconstruction weights of all patches of low-frequency image $\hat{I}^{HR}_{LF(t,c^{tar})}$ can be represented using the matrix $W \in \mathbb{R}^{(nM)\times M}$, and its elements are defined as

$$W(M(j-1)+i, i) = \begin{cases} \omega_i^j, & when\ p_{(i,j)}\ is\ the\ similar\ patch\ of\ p_i \\ 0, & otherwise. \end{cases} \tag{5}$$

Fig. 2 shows one example on how we represent the target HR patch with the above-mentioned optimization. In this figure, the top-right and bottom-right respectively show two represented patches with the best and worst matching errors in the image. Now we have computed the weights of each low-frequency patch with its nearest neighbors in the manifold structure, and such weights can be further utilized in the HR texture reconstruction.

D. Low-rank Constrained Artifact Regression

For multiview videos, it is challenging to avoid visual artifacts when learning texture details from similar patches among different views under complex lighting and motion conditions. On the other hand, low-rank representation based modeling of similar patches has been successfully used in many image reconstruction tasks [21], [46]. Inspired by these exciting works, we propose a low-rank regularization method to effectively avoid reconstruction artifacts.

Formally, for each $d_L \times d_L$ patch g_l in the image $I^{HR}_{(t,c^{tar})}$, we find its K most similar patches in the same image using the KNN algorithm, and let $g_{(l,k)} \in \mathbb{R}^{(d_L)^2}$ be the k-th most similar patch of g_l, where $1 \le k \le K$. Let $T_{(l,k)} \in \mathbb{R}^{(d_L)^2 \times N_h}$ be a patch selection matrix, which is used to represent each patch $g_{(l,k)}$ within the image $I^{HR}_{(t,c^{tar})}$ as:

$$g_{(l,k)} = T_{(l,k)} I^{HR}_{(t,c^{tar})}. \tag{6}$$

As we use the one column version of the image to be computed, only if the i-th pixel in $g_{(l,k)}$ is selected from the j-th pixel in $I^{HR}_{(t,c^{tar})}$, the corresponding element $T_{(l,k)ij}$ is 1, otherwise it is 0. Obviously, $T_{(l,k)}$ is a binary matrix, and $(T_{(l,k)})^\top T_{(l,k)}$ is a diagonal matrix. We aggregate all similar

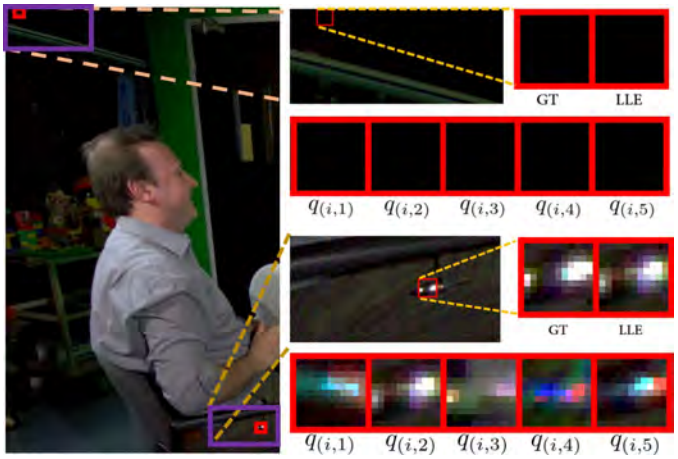


Fig. 2. Our LLE-based HR image patch representation of the target view using 5 most similar patches selected from the reference images. GT: the ground truth; LLE: the LLE-based optimization representation; $q_{(i,1)}$ to $q_{(i,5)}$ are the most similar patches from $I^{HR}_{(t-1,c_1)}$, $\hat{I}^{HR}_{(t-1,c^{tar})}$, $I^{HR}_{(t-1,c_2)}$, $I^{HR}_{(t,c_1)}$ and $I^{HR}_{(t,c_2)}$, respectively. The top-right and bottom-right respectively show two represented patches with the best and worst matching errors in the image.

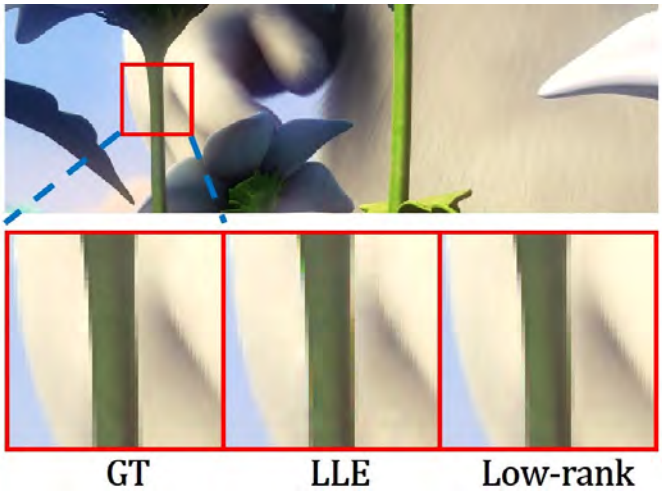


Fig. 3. Our low-rank constrained artifact regression can effectively eliminate visual artifacts introduced by patch representation of reference patches. LLE: our spatio-temporal HR texture representation using reference patches; Low-rank: our low-rank constrained regression result.

patches $g_{(l,k)}$ of g_l as $G_l = [g_{(l,1)}, ..., g_{(l,K)}]$, impose it to have a low-rank property, and model the matrix G_l using the Weighted Nuclear Norm (WNN) introduced in [46]. Typically, it is expected that

$$\|G_l\|_{*,\delta} = \sum_j \delta_j \sigma_j, \tag{7}$$

where σ_j is the j-th singular value of G_l, and $\delta_j \geq 0$ is its corresponding weight. Here the larger singular values capture the most important low-frequency information, while the smaller singular values typically encode high-frequency information [47]. Therefore, as presented in [21], [46], the weight δ_j can be introduced to make sure that the components corresponding to larger singular values have less shrinkage:

$$\delta_j = \lambda/(\sigma_j + \varepsilon). \tag{8}$$

Here we introduce a positive constant λ to scale the singular values. Different from other methods (e.g. denoising), in our solution it is difficult to obtain an analytic solution for λ, and we turn to empirically find a relatively good value (see more details in the experiments section). ε is a small positive constant to avoid division by zero.

Then, the solution of the low-rank constrained patch recovery can be optimally obtained by the Weighted Singular Value Thresholding (W-SVT) solver:

$$\mathcal{S}_\delta(G_l) = U(\Sigma - Diag(\delta))_+ V^\top. \tag{9}$$

In this equation, $U\Sigma V^\top$ is the singular value decomposition (SVD) of G_l. Let $\Sigma^{'} = (\Sigma - Diag(\delta))_+$, this is the matrix of soft-thresholded singular values such that

$$\Sigma^{'}_{jj} = max\{\Sigma_{jj} - \delta_j, 0\}. \tag{10}$$

Consequently, with this optimal solution we can get the artifact regression result of the reconstructed image. Fig. 3 shows one example on how to eliminate the visual artifacts introduced by inaccurate patch representation with reference patches.

E. Joint Optimization

We have separately introduced means to: 1) learn texture details from the reference HR frames using LLE-based reconstruction, and 2) perform low-rank approximation based patch artifact regression. Combining these two ideas, we can further formulate the proposed HR image reconstruction task as a global optimization problem:

$$\begin{aligned} \hat{I}^{HR}_{(t,c^{tar})} = \arg\min_{I^{HR}_{(t,c^{tar})}} & \left\| \hat{I}^{HR}_{LF(t,c^{tar})} - BI^{HR}_{(t,c^{tar})} \right\|_2^2 \\ & + \alpha \left\| Q - VW \right\|_2^2 + \beta \sum_{l=1}^{L} \left\| G_l \right\|_{*,\delta}, \\ & s.t.\ q_i = S_i I^{HR}_{(t,c^{tar})},\ i = 1, ..., M; \\ & g_{(l,k)} = T_{(l,k)} I^{HR}_{(t,c^{tar})};\ l = 1, ..., L, k = 1, ..., K. \end{aligned} \tag{11}$$

In detail, the image $I^{HR}_{(t,c^{tar})}$ is first segmented into patches with the size of $d_M \times d_M$ pixels, and for each patch we convert it to a one-column q_i where the pixels are rearranged vertically. $q_{(i,j)}$ is the patches of the reference HR frames $I^{HR}_{(\cdot,\cdot)}$, for $i = 1, ..., M$, and j-th nearest neighbors patch of q_i. We define $Q = [q_1, ..., q_M]$, $V = [q_{(1,1)}, ..., q_{(M,1)}, q_{(1,2)}, ..., q_{(M,2)}, q_{(1,3)}, ..., q_{(M,3)}, q_{(1,4)}, ..., q_{(M,4)}, q_{(1,5)}, ..., q_{(M,5)}]$. To represent q_i with image $I^{HR}_{(t,c^{tar})}$, we define a binary matrix $S_i \in \mathbb{R}^{(d_M)^2 \times N_h}$ to select the desired patch which function is similar to $T_{(j,k)}$. Combining Q and $q_i = S_i I^{HR}_{(t,c^{tar})}$, we can get the following formulation

$$\begin{aligned} Q &= [S_i I^{HR}_{(t,c^{tar})}, ..., S_M I^{HR}_{(t,c^{tar})}] \\ &= [S_i, ..., S_M] I^{HR}_{(t,c^{tar})} \triangleq \widetilde{S} I^{HR}_{(t,c^{tar})}. \end{aligned} \tag{12}$$

With this definition, Eq. 11 can then be reformulated to the following optimization problem:

$$\hat{I}^{HR}_{(t,c^{tar})} = \underset{I^{HR}_{(t,c^{tar})}}{\arg\min} \left\| \hat{I}^{HR}_{LF(t,c^{tar})} - BI^{HR}_{(t,c^{tar})} \right\|_2^2 + \alpha \left\| \widetilde{S} I^{HR}_{(t,c^{tar})} - VW \right\|_2^2 + \beta \sum_{l=1}^{L} \|G_l\|_{*,\delta}, \quad s.t.\ g_{(l,k)} = T_{(l,k)} I^{HR}_{(t,c^{tar})};\ l = 1, ..., L;\ k = 1, ..., K. \quad (13)$$

This optimization problem is difficult to directly solve due to the WNN term. We turn to use an iterative optimization method based on the Alternating Direction Method of Multipliers algorithm (ADMM) [48]. Then, the constraint is converted to the cost function by an augmented Lagrangian formulation, and the total optimization function becomes:

$$\hat{I}^{HR}_{(t,c^{tar})} = \underset{I^{HR}_{(t,c^{tar}))}}{\arg\min} \left\| \hat{I}^{HR}_{LF(t,c^{tar})} - BI^{HR}_{(t,c^{tar})} \right\|_2^2 + \alpha \left\| \widetilde{S} I^{HR}_{(t,c^{tar})} - VW \right\|_2^2 + \beta \sum_{i=1}^{L} \|G_i\|_{*,\delta} + \gamma \sum_{l=1}^{L} \sum_{k=1}^{K} \left\| g_{(l,k)} - T_{(l,k)} I^{HR}_{(t,c^{tar})} + e_{(l,k)} \right\|_2^2. \quad (14)$$

Here $e_{(l,k)}$ is the weighting parameter, and α, β and γ are three Lagrange multipliers to balance the corresponding components. As mentioned in [48], α, β and γ might affect the performance of the algorithm. However, most of existing ADMM methods are not overly sensitive to those parameters, and slightly tuning these parameters is required when they are initialized with small positive values. As in standard ADMM methods, assuming all other parameters are fixed, we update each variable G_l, $\hat{I}^{HR}_{(t,c^{tar})}$ and $e_{(l,k)}$ iteratively, until the convergence of the total function is reached. Now we present the details on how to solve our optimization function with the ADMM pipeline.

Updating $\hat{I}^{HR}_{(t,c^{tar})}$: Assuming all other variables and parameters are fixed, the reconstructed image $\hat{I}^{HR}_{(t,c^{tar})}$ is updated by solving the following problem:

$$\underset{I^{HR}_{(t,c^{tar})}}{\arg\min} \left\| \hat{I}^{HR}_{LF(t,c^{tar})} - BI^{HR}_{(t,c^{tar})} \right\|_2^2 + \alpha \left\| \widetilde{S} I^{HR}_{(t,c^{tar})} - VW \right\|_2^2 + \gamma \sum_{l=1}^{L} \sum_{k=1}^{K} \left\| g_{(l,k)} - T_{(l,k)} I^{HR}_{(t,c^{tar})} + e_{(l,k)} \right\|_2^2. \quad (15)$$

This is a typical unconstrained quadratic program issue. Suppose $\widetilde{T} = \sum_l \sum_k (T_{(l,k)})^\top T_{(l,k)}$ and $\widetilde{g} = \sum_l \sum_k (T_{(l,k)})^\top (g_{(l,k)} + e_{(l,k)})$, the optimal value of Eq. 15 can be obtained with the following equation:

$$\hat{I}^{HR}_{(t,c^{tar})} = [B^\top B + \alpha(\widetilde{S})^\top \widetilde{S} - \gamma \widetilde{T}]^{-1} (B^\top \hat{I}^{HR}_{LF(t,c^{tar})} + \alpha \widetilde{S} VW + \gamma \widetilde{g}). \quad (16)$$

Updating $\hat{I}^{HR}_{LF(t,c^{tar})}$: Once we get the value of $\hat{I}^{HR}_{(t,c^{tar})}$, the reconstructed low-frequency version of the target HR image is updated by:

$$\hat{I}^{HR}_{LF(t,c^{tar})} = B\hat{I}^{HR}_{(t,c^{tar})}. \quad (17)$$

Updating G_l : Let us denote:

$$\widetilde{G_l} = [(T_{(l,1)} \hat{I}^{HR}_{(t,c^{tar})} - e_{(l,1)}), ..., (T_{(l,K)} \hat{I}^{HR}_{(t,c^{tar})} - e_{(l,K)})], \quad (18)$$

the task of updating the patch matrices $G_l, l = 1, ..., L$ can then be expressed as follows:

$$\underset{G_l}{\arg\min}\, \beta \|G_l\|_{*,\delta} + \gamma \left\| G_l - \widetilde{G_l} \right\|_F^2. \quad (19)$$

Following the methodology described in Sec. III-D, this issue can further be fixed with the W-SVT solver [46]:

$$G_l = U_l(\Sigma_l - \frac{\beta}{2\gamma} Diag(\delta))_+ V_l^\top. \quad (20)$$

Note that in this equation $U_l \Sigma_l V_l^\top$ is the SVD of $\widetilde{G_l}$.

Updating $e_{(l,k)}$: Lastly, this weighting parameter can be updated with the following standard ADMM method:

$$e_{(l,k)} := e_{(l,k)} + [g_{(l,k)} - T_{(l,k)} \hat{I}^{HR}_{(t,c^{tar})}], l = 1, ..., L, k = 1, ..., K. \quad (21)$$

Therefore, the proposed algorithm yields the reconstructed HR image $\hat{I}^{HR}_{(t,c^{tar})}$. The whole algorithm including the above-mentioned updating is also summarized in Algorithm 1.

Algorithm 1 HR reconstruction for MR multiview videos

Input: The LR image $I^{LR}_{(t,c^{tar})}$, the reference HR images $\hat{I}^{HR}_{(t-1,c^{tar})}$, $I^{HR}_{(t-1,c_1)}$, $I^{HR}_{(t-1,c_2)}$, $I^{HR}_{(t,c_1)}$ and $I^{HR}_{(t,c_2)}$;
Output: The reconstructed HR image $\hat{I}^{HR}_{(t,c^{tar})}$;

1: **Pre-processing:** Get the initialized image $\hat{I}^{HR}_{LF(t,c^{tar})}$;
2: **Optimization:**
3: **while** not converged **do**
4: Update $\hat{I}^{HR}_{(t,c^{tar})}$ via Eq. 16;
5: Update $\hat{I}^{HR}_{LF(t,c^{tar})}$ via Eq. 17;
6: Update G_l via Eq. 20;
7: Update $e_{(l,k)}$ via Eq. 21;
8: **end while**
9: **return the reconstructed HR image** $\hat{I}^{HR}_{(t,c^{tar})}$.

IV. Experimental Results

We implemented the proposed SR approach on a desktop with Microsoft Windows10 operating system, Intel (R) Core (TM) i7-7700 CPU, 16G Memory. We use C++ programming language and OpenCV Library to implement our SR approach. Our experiments are applied on various well-known multiview videos, which are captured by well-calibrated multiview camera rigs, or rendered based on computer-generated multiview scenes. For these typical multiview videos, the target LR video can always take the neighbouring HR videos as references. The proposed approach is also compared with some state-of-the-art SR methods, including single- and multi-image based methods. Some latest SR methods based on CNNs are also adapted such that they can be evaluated on MR multiview videos. Finally, we also introduce a new SR evaluation model for MR multiview video.

A. Parameters Selection

In our solution the value of parameter d_L is 15 in the low-rank regularization which is the same as in [46]. Besides

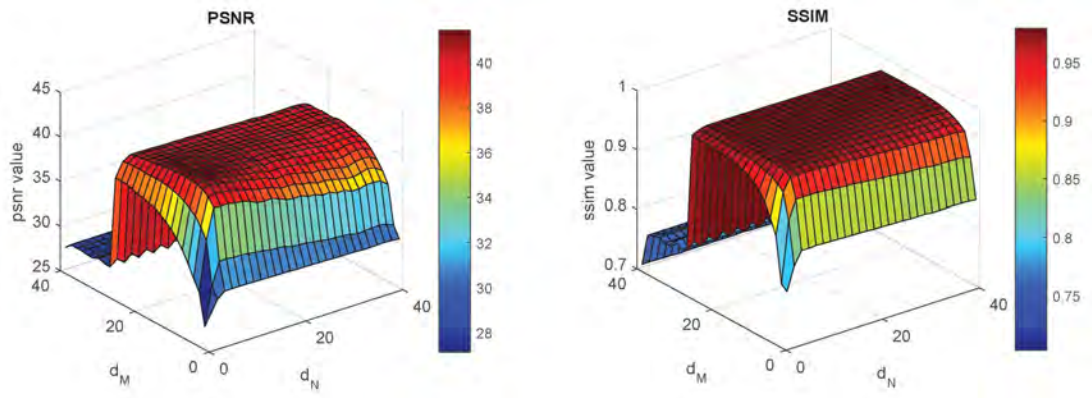


Fig. 4. Parameters d_N and d_M selection of the proposed method.

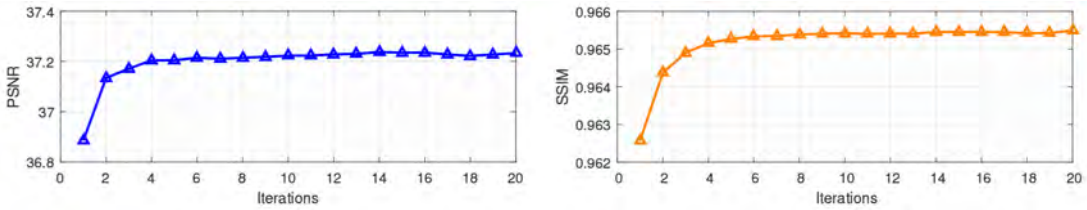


Fig. 5. Convergence curves of PSNR and SSIM of our optimization (for the *BBB* sequence under SR factor of 3).

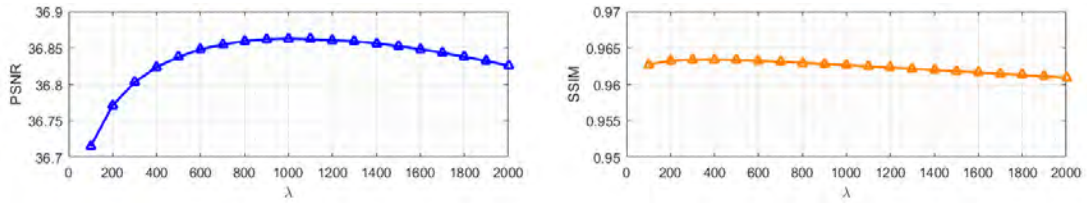


Fig. 6. Parameter λ selection (for the *BBB* sequence under SR factor of 3).

that, we jointly consider choosing two optimal parameters d_N and d_M, which are used to define the patch size in pre-processing and learning HR details, respectively. Fig. 4 shows the reconstruction quality metrics of the Peak-Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index (SSIM) on the *Big Buck Bunny (BBB)* video sequence [49]. Based on this test, we empirically choose $d_N = 5$ and $d_M = 11$ respectively in our solution. As shown in Fig. 5, the PSNR and SSIM values can quickly converge under several iterations. Therefore, we set the iteration number as 4 in all our experiments.

The parameter λ, which is used to scale the singular values, is set to 1000. As shown in Fig. 6, with this value the proposed solution yields a good performance in terms of PSNR and SSIM metrics. Since the proposed solution is solved using a typical ADMM framework, the parameters that balance all involved terms in the final optimization problem are experimentally tuned; based on these experiments, we empirically set α, β and γ as 1, 1, and 0.5, respectively.

In most of our experiments we take five nearest neighbors when representing each HR patch using our LLE based optimization. This is because that the reference frames consist of four HR images from the neighboring views and one reconstructed previous HR image of the same viewpoint, and the most similar patch of each reference frame to the target patch is selected to comprise the latter's nearest neighbors. The exception is reconstructing the first HR frame of the target video stream, where only four HR frames of the neighboring views will be involved. Besides that, some more complex extensions will be presented in Sec. IV-D.

B. Comparison with Other SR Methods

Our framework has been compared with many other SR methods on a variety of multiview images and videos, both for synthetic and real image datasets. We used the following synthetic multiview images for test: *BBB* [49], *Shark*, *Dancer* and *MicroWorld* [55]. The cameras #5, #6, and #7 have been chosen as left, central, and right camera views for *BBB*, and the cameras #1, #2, and #3 have been chosen as left, central, and right camera views for *Shark*, *Dancer* and *MicroWorld*. As for the real image dataset, we used cameras #1, #2 and #3 of multiview videos *Knuffelgooi* and *Ballroom* [56] respectively, as their own left, central, and right camera views. All central camera views are down-sampled as input LR videos. In the following experiments, the best three calculation values are with red, blue and green colors in the corresponding tables.

TABLE II
COMPARISON OF PSNR AND SSIM METRICS WITH DIFFERENT SR FACTORS.

		BBB	*Shark*	*Knuffelgooi*	*Ballroom*	*Undo_Dancer*	*MicroWorld*
Bicubic	2	32.280 / 0.932	36.269 / 0.961	34.072 / 0.859	34.288 / 0.937	30.267 / 0.868	30.503 / 0.837
	4	27.744 / 0.835	31.372 / 0.903	30.992 / 0.756	27.785 / 0.791	26.769 / 0.757	26.154 / 0.616
	8	24.199 / 0.733	28.050 / 0.848	28.303 / 0.674	23.176 / 0.609	24.092 / 0.641	23.405 / 0.453
ScSR [5]	2	34.615 / 0.951	38.316 / 0.972	34.804 / 0.871	36.562 / 0.955	32.057 / 0.901	32.016 / 0.884
	4	28.671 / 0.853	32.191 / 0.912	31.389 / 0.764	29.034 / 0.821	27.466 / 0.757	26.734 / 0.654
	8	24.843 / 0.741	28.579 / 0.852	28.754 / 0.679	23.922 / 0.631	24.568 / 0.648	23.720 / 0.468
Kim [50]	2	35.047 / 0.954	38.607 / 0.973	34.844 / 0.870	36.741 / 0.956	32.066 / 0.900	32.165 / 0.887
	4	29.308 / 0.865	32.461 / 0.916	31.724 / 0.769	29.347 / 0.824	27.897 / 0.761	26.778 / 0.654
	8	—	—	—	—	—	—
SelfExSR [14]	2	34.945 / 0.956	38.731 / 0.974	35.294 / 0.887	37.141 / 0.961	32.269 / 0.907	32.316 / 0.889
	4	29.796 / 0.876	32.938 / 0.923	32.167 / 0.786	29.813 / 0.837	28.309 / 0.775	27.118 / 0.668
	8	25.576 / 0.770	28.960 / 0.863	29.336 / 0.695	24.229 / 0.638	25.193 / 0.668	23.932 / 0.476
SRCNN [27]	2	34.812 / 0.954	38.454 / 0.973	34.919 / 0.875	36.917 / 0.960	32.235 / 0.903	32.069 / 0.885
	4	29.640 / 0.868	32.586 / 0.916	31.875 / 0.772	29.888 / 0.834	28.313 / 0.769	26.796 / 0.652
	8	—	—	—	—	—	—
VDSR [51]	2	35.453 / 0.959	39.013 / 0.976	35.099 / 0.878	36.941 / 0.960	32.402 / 0.908	32.448 / 0.894
	4	30.184 / 0.880	32.999 / 0.923	32.163 / 0.780	30.164 / 0.844	28.557 / 0.779	27.003 / 0.667
	8	—	—	—	—	—	—
SRResNet [29]	2	35.655 / 0.960	39.316 / 0.976	35.481 / 0.889	37.376 / 0.963	32.529 / 0.910	32.774 / 0.899
	4	30.574 / 0.889	33.522 / 0.929	32.448 / 0.791	30.533 / 0.854	28.674 / 0.786	27.489 / 0.686
	8	26.199 / 0.790	29.523 / 0.873	29.959 / 0.710	25.020 / 0.676	25.844 / 0.687	24.244 / 0.491
EDSR [52]	2	35.820 / 0.961	39.458 / 0.9768	35.542 / 0.889	37.434 / 0.963	32.587 / 0.911	32.885 / 0.901
	4	30.892 / 0.894	33.789 / 0.932	32.603 / 0.795	30.662 / 0.858	28.735 / 0.789	27.626 / 0.693
	8	—	—	—	—	—	—
RCAN [53]	2	35.900 / 0.962	39.505 / 0.9769	35.524 / 0.890	37.444 / 0.964	32.625 / 0.912	32.959 / 0.903
	4	31.080 / 0.896	33.868 / 0.933	32.633 / 0.796	30.879 / 0.863	28.879 / 0.791	27.677 / 0.696
	8	26.704 / 0.804	29.659 / 0.876	30.112 / 0.714	25.481 / 0.701	26.035 / 0.694	24.342 / 0.497
SRGAN [29]	2	35.007 / 0.950	38.465 / 0.968	34.553 / 0.860	36.372 / 0.949	31.755 / 0.888	31.830 / 0.876
	4	29.742 / 0.866	32.560 / 0.906	31.540 / 0.752	29.394 / 0.812	27.505 / 0.721	26.452 / 0.636
	8	24.505 / 0.727	27.714 / 0.827	28.439 / 0.654	23.125 / 0.582	24.478 / 0.620	22.779 / 0.417
ESRGAN [54]	2	35.149 / 0.953	38.278 / 0.970	34.528 / 0.860	36.719 / 0.954	31.888 / 0.896	32.023 / 0.883
	4	30.203 / 0.876	33.090 / 0.918	31.608 / 0.750	29.883 / 0.827	27.904 / 0.755	26.814 / 0.654
	8	24.005 / 0.735	25.912 / 0.822	27.576 / 0.651	21.995 / 0.544	23.207 / 0.601	22.210 / 0.420
SRNTT-ℓ_2 [44]	2	—	—	—	—	—	—
	4	31.472 / 0.909	33.940 / 0.947	31.980 / 0.786	31.553 / 0.884	30.597 / 0.855	28.197 / 0.767
	8	25.930 / 0.790	27.463 / 0.854	28.475 / 0.590	25.234 / 0.692	25.399 / 0.687	24.075 / 0.500
SRNTT [44]	2	—	—	—	—	—	—
	4	30.543 / 0.887	33.208 / 0.926	31.114 / 0.687	30.541 / 0.851	30.153 / 0.838	27.219 / 0.737
	8	25.185 / 0.760	26.245 / 0.781	26.175 / 0.514	23.928 / 0.627	24.678 / 0.641	22.480 / 0.438
CrossNet [43]	2	—	—	—	—	—	—
	4	35.191 / 0.960	38.979 / 0.982	31.038 / 0.782	33.285 / 0.932	34.683 / 0.943	32.768 / 0.933
	8	30.048 / 0.921	36.575 / 0.982	25.094 / 0.648	24.209 / 0.832	32.193 / 0.921	28.874 / 0.896
Ours	2	39.322 / 0.976	45.750 / 0.987	35.564 / 0.883	39.299 / 0.967	40.461 / 0.976	36.951 / 0.961
	4	37.024 / 0.964	43.490 / 0.983	33.242 / 0.806	34.469 / 0.918	37.320 / 0.963	33.844 / 0.936
	8	33.691 / 0.943	29.691 / 0.909	31.286 / 0.749	29.901 / 0.873	30.261 / 0.878	32.515 / 0.906

Our approach has been compared with some state-of-the-art SR methods, including single-image SR methods (e.g. ScSR [5], Kim [50], SelfExSR [14], SRCNN [27], VDSR [51], SRResNet [29], EDSR [52], RCAN [53], SRGAN [29] and ESRGAN [54]) and multi-image SR methods (e.g. Richter [1], CrossNet [43] and SRNTT [44]). In order to be as fair as possible, some methods such as CrossNet [43] have been retrained by taking multiview HR images as given reference images. The method [44] are with two versions: one for minimizing the MSE (named SRNTT-ℓ_2) and another complete version with adversarial loss (named SRNTT).

Firstly, the objective evaluation with PSNR and SSIM metrics on the *BBB*, *Shark*, *Knuffelgooi* and *Ballroom* sequences are shown in Tab. II, where SR factors include 2, 4 and 8 in both spatial dimensions. In most of these experiments, our method achieves the best results. Especially, when the SR factor becomes 8 on the *MicroWord* sequence, our PSNR and SSIM values are +3.755 and +0.01 higher than that of the second best method (i.e. CrossNet [43]).

In order to observe the temporal consistency of the generated results, Fig. 7 presents the results of multiple consecutive frames on the *BBB*, *Shark*, and *Knuffelgooi* sequences. These stable results demonstrate that our method reasonably preserves the temporal consistency, because the previous HR frames of multiple camera views are well modeled in the optimization framework.

Secondly, we provide some examples on visual comparison against other existing methods. Fig. 8 shows the zooming

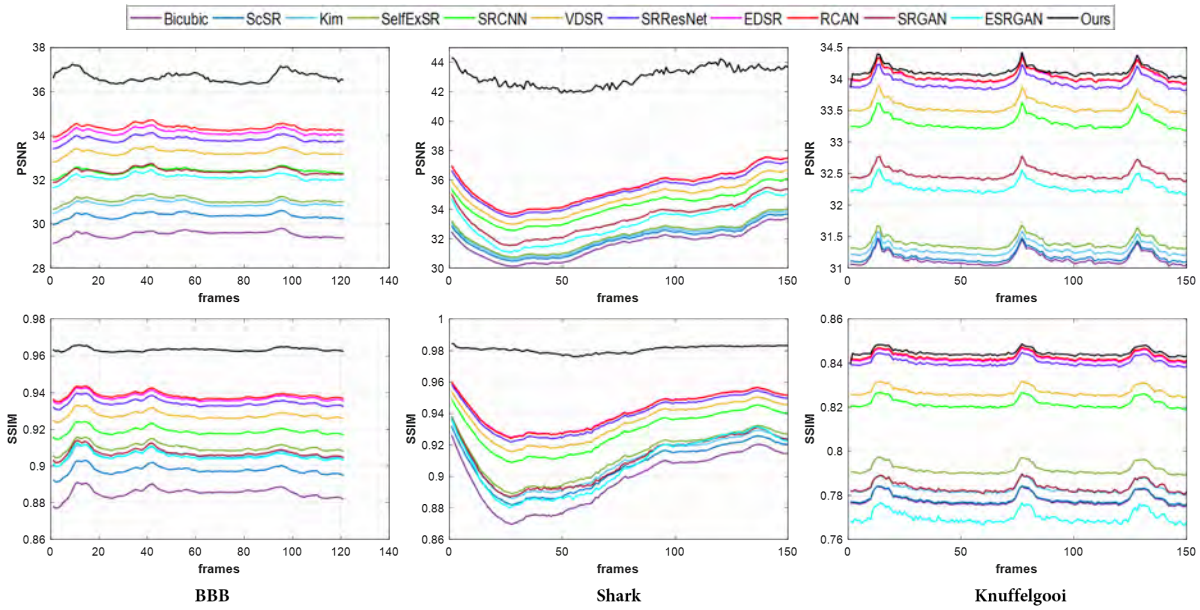


Fig. 7. PSNR and SSIM for *BBB*, *Shark* and *Knuffelgooi* videos, the SR factor is 3.

details of those results generated by different SR methods. For the *BBB* and *Knuffelgooi* the upsampling factors are 4 and 8, respectively, in both spatial dimensions. It is easy to observe that our solution is better at recovering the detailed textures. For instance, in the upper half of Fig. 8, the yellow leaves are well reconstructed using our approach, while the other methods tend to generate over smoothing results or other artifacts. This obvious visual phenomenon also happens on *Knuffelgooi* example (see the bottom half of Fig. 8), where the latter example is captured from the real world. In all of these experiments, our method achieves the highest PSNR and SSIM values (see the bottom of each sub-figures).

Thirdly, we also compare with some other methods that are specified (or could be easily adapted) for MR multiview video. Typical reference methods include Richter [1], CrossNet [43] and SRNTT [44], where the first method is a depth map based multiview SR solution, and the other methods are using deep learning techniques. We then performed all of these methods on different Middlebury multiview sequences. For these tests, cameras #1, #3, and #5 have been chosen as left, central, and right camera views, respectively, and all input central camera views are with LR images. The upsampling factors have been chosen to 4 in both spatial dimensions. Moreover, the videos are respectively applied on both the Y luminance and RGB color channels, while the results of Richter [1] comes from their article directly. Tab. III summarizes the results. In general, we notice that our method yields the best PSNR values for all examples. This evaluation proves again that our approach is very suitable for MR multiview video.

C. Ablation Study

We firstly evaluate the gains brought by LLE-based HR patch reconstruction in our solution. As listed in the first row of Tab. IV, when directly removing the LLE-based reconstruction component, the decrease in PSNR could be up to -7.920 in the *Undo_Dancer* sequence. In addition, we check whether the LLE-based reconstruction with five most similar reference patches is better compared to that of only one nearest patch. For the latter, the target HR patch is directly duplicated from the most similar patch taken from the five HR reference images. As shown in the "one ref patch" row of Tab. IV, using only one reference patch leads to a substantial decrease of the quality of the final reconstruction. Secondly, we remove the low-rank constrained objective in our solution to assess the impact on the overall performance; the results show that the PSNR of the reconstructed image will reduce with up to -1.318 dB for the *Ballroom* sequence.

Thirdly, we investigate the impact of the reconstructed previous HR frame of the target view in our optimization. The third row of Tab. IV (*i.e.* "no pre-frames") shows the result in this case, revealing that the PSNR values would be quickly decreased, with up to -1.088 dB in the *MicroWorld* sequence.

Finally, we test how much we gain when solving the optimal weight matrix W in Eq. 5 with the low-frequency versions of the reference HR images. As reported in the last row of Tab. IV (i.e. "no LFRef"), compared to that directly with the reference HR images, our solution gains respectively up to $+2.054$ dB and $+0.041$ in PSNR and SSIM respectively.

D. Experiments for More Complex Situations

Till now, all the above-mentioned experiments select three adjacent viewpoints as left, central, and right camera views. Now we will consider more cases with different viewpoints, baselines, or even more complex situations that could probably arise in practice. For example, the left or right view might be missing, or there are more than one view just from the left/right side of the target viewpoint. For these cases, the reference HR frames of the target HR image are changed accordingly in

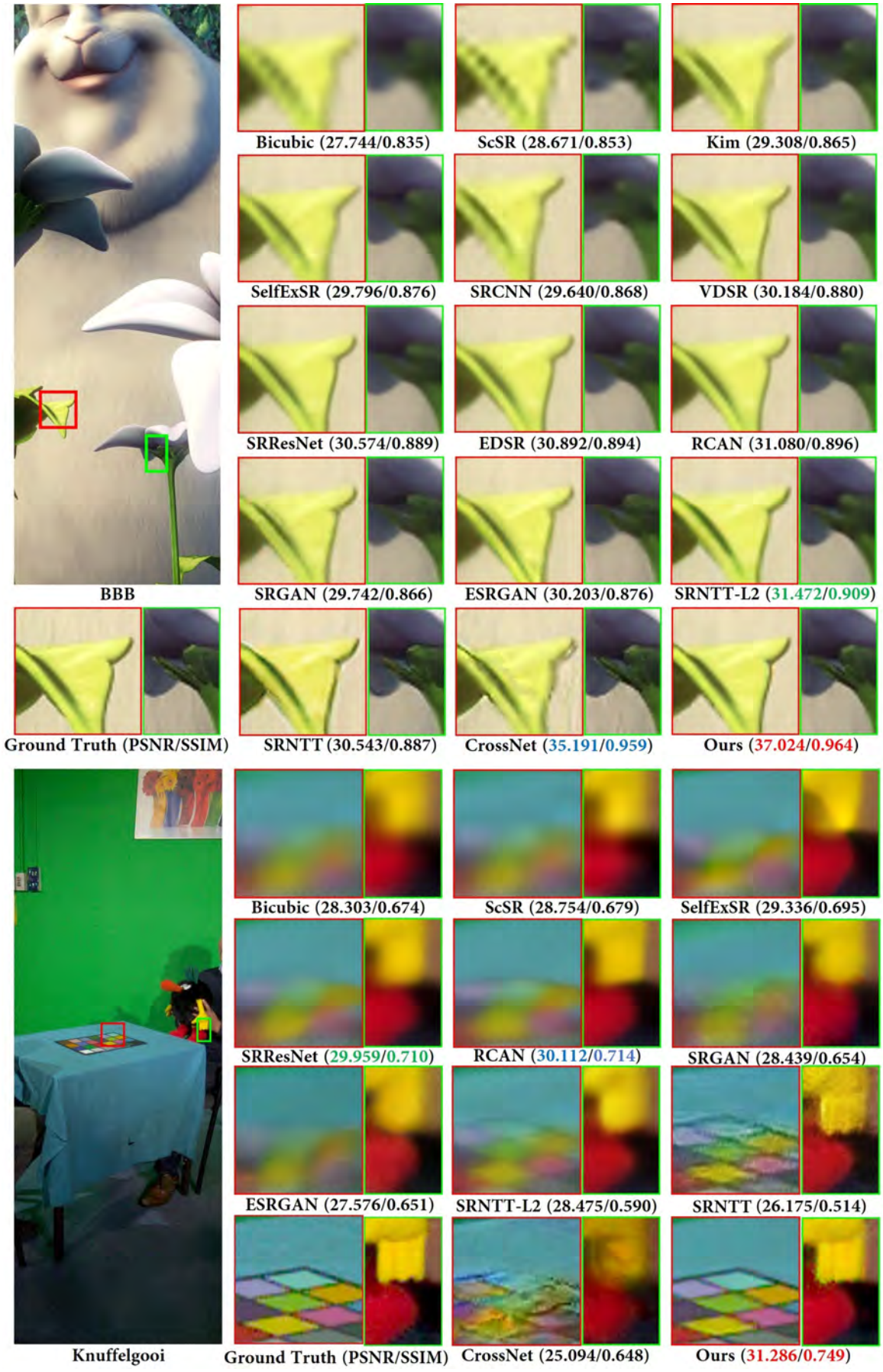


Fig. 8. Visual comparison for the *BBB* and *Knuffelgooi* sequences, the SR factors are 4 and 8, respectively.

TABLE III
PSNR COMPARISON ON MIDDLEBURY DATA SETS WITH SR FACTOR 4.

	Art	*Books*	*Dolls*	*Moebius*	*Aloe*	*Baby1*	*Bowling1*	*Lampshade1*	*Midd1*	*Plastic*
Bicubic	31.70	29.43	30.42	31.85	27.12	31.32	35.25	35.86	30.65	38.06
Kim [50]	32.87	30.34	31.27	32.49	27.45	31.65	35.91	36.67	31.48	39.47
ScSR [5]	32.32	29.95	30.86	32.23	27.43	31.44	35.56	36.08	31.08	38.46
Richter [1]	35.13	34.23	34.90	35.44	33.07	37.23	38.88	39.33	34.90	42.76
CrossNet-RGB [43]	33.54	30.04	31.92	33.91	29.24	32.48	37.45	35.67	32.61	36.25
CrossNet-Y [43]	35.23	31.56	33.51	35.76	30.92	34.81	39.28	38.28	33.97	38.88
SRNTT-RGB [44]	34.19	30.60	32.75	33.21	30.29	32.74	35.62	33.70	32.27	35.95
SRNTT-Y [44]	36.12	32.20	34.58	35.14	32.26	34.84	37.95	37.90	34.07	40.20
SRNTT-ℓ_2-RGB [44]	34.90	31.39	33.68	34.15	30.32	33.61	36.60	38.18	33.35	39.29
SRNTT-ℓ_2-Y [44]	36.90	32.97	35.52	36.04	32.00	35.79	38.40	40.68	35.01	40.92
Ours-RGB	36.07	32.82	36.35	35.88	33.88	36.97	39.64	40.99	35.12	44.79
Ours-Y	38.00	34.49	38.18	37.65	35.86	38.92	41.54	43.08	36.71	46.47

TABLE IV
PSNR AND SSIM GAINS OF EACH STEP IN OUR SOLUTION.

	BBB	*Shark*	*Knuffelgooi*	*Ballroom*	*Undo_Dancer*	*MicroWorld*
no LLE	-4.295 / -0.034	-7.718 / -0.024	-0.123 / -0.004	-0.751 / -0.001	-7.920 / -0.704	-3.566 / -0.103
no low-rank	-0.926 / -0.007	-0.427 / -0.002	-0.786 / -0.026	-1.318 / -0.036	-0.323 / -0.004	-0.315 / -0.005
no pre-frames	-0.906 / -0.004	-0.982 / -0.001	-0.647 / -0.022	-0.201 / -0.024	-0.804 / -0.048	-1.088 / -0.017
one ref patch	-2.242 / -0.014	-3.729 / -0.006	-1.382 / -0.035	-1.685 / -0.042	-3.870 / -0.035	-2.677 / -0.047
no LFRef	-2.043 / -0.014	-1.222 / -0.001	-0.428 / -0.012	-0.057 / -0.022	-2.054 / -0.024	-1.798 / -0.041

TABLE V
PSNR AND SSIM METRICS FOR MORE COMPLEX CASES (SR FACTOR IS 3).

	left	central	right	*BBB*	*Shark*	*Knuffelgooi*	*Ballroom*	*Undo_Dancer*	*MicroWorld*
case1	#1	#2	#3	37.204/0.965	44.247/0.984	34.085/0.845	35.168/0.948	38.589/0.972	33.852/0.925
case2	#1	#2		34.879/0.951	42.895/0.989	33.496/0.829	33.617/0.912	37.136/0.957	32.121/0.896
case3	#0	#2		33.962/0.944	42.810/0.983	33.156/0.823	32.924/0.903	36.698/0.955	30.454/0.851
case4	#0,#1	#2		35.903/0.958	44.184/0.984	33.609/0.831	33.964/0.917	38.489/0.966	32.767/0.909
case5	#0	#2	#3	36.264/0.961	44.232/0.984	33.730/0.832	34.233/0.919	38.502/0.967	32.829/0.909
case6	#0	#2	#4	35.844/0.958	43.893/0.984	33.559/0.830	33.923/0.915	38.269/0.966	31.621/0.881
case7		#2	#9	32.167/0.923	40.766/0.979	—	—	34.541/0.937	28.863/0.785

our solution. Tab. V presents the evaluation results of seven different cases under such complex situations. It is interesting that in case4 the PSNR and SSIM values are higher than that of case3. In other words, when introducing one more reference HR video from a neighboring view, our HR reconstruction solution could learn more useful texture information from such neighboring view. All these results show that, no matter the multiview baseline becomes wider or two reference views are only from the same side in contrast to the target view (or even one reference view is missing), our solution still works well. This especially demonstrates that our solution is easily adaptable to various complex situations.

E. Our Gradient-aware Evaluation Metric

Here we introduce a novel SR evaluation metric specified for MR multiview video, where we focus on exploiting the correlation of the gradient distribution between the target HR image and its neighboring HR views. For digital images there is a well-known heavy-tailed distribution [10] when the logarithmic function of the image's gradient is calculated, and such distribution highly reflects how much the image is reconstructed compared to the original one. For instance, the left sub-figure of Fig. 9 shows the calculated gradient density distribution in the horizontal direction, where the distribution curves dramatically change when a HR image is degraded into its LR version. Moreover, for the LR image the distribution range of its gradients also obviously changes, which means that it changes the diversity of the original texture details.

According to this observation, we first compute the difference of the gradient density distribution curves between the reconstructed HR image $\hat{I}^{HR}$ and the ground truth I^{HR}. Secondly, we also compare the distribution range of such curves to quantify the preservation of the texture diversity of the original image. We thus formulate the following equation:

$$M_{gd} = \log_3((e^{-RM_x(\hat{I}^{HR}, I^{HR})} \times GR_x(\hat{I}^{HR}, I^{HR}) \times e^{-RM_y(\hat{I}^{HR}, I^{HR})} \times GR_y(\hat{I}^{HR}, I^{HR}))^{\frac{1}{3}} + 2). \quad (22)$$

Here $RM_x(\hat{I}^{HR}, I^{HR})$ and $RM_y(\hat{I}^{HR}, I^{HR})$ are the normalized Root-Mean-Squard-Error (RMSE) of the accumulated gradient error of the x and y directions between $\hat{I}^{HR}$ and I^{HR}. $GR_x(\hat{I}^{HR}, I^{HR})$ is used to calculate the diversity of gradient

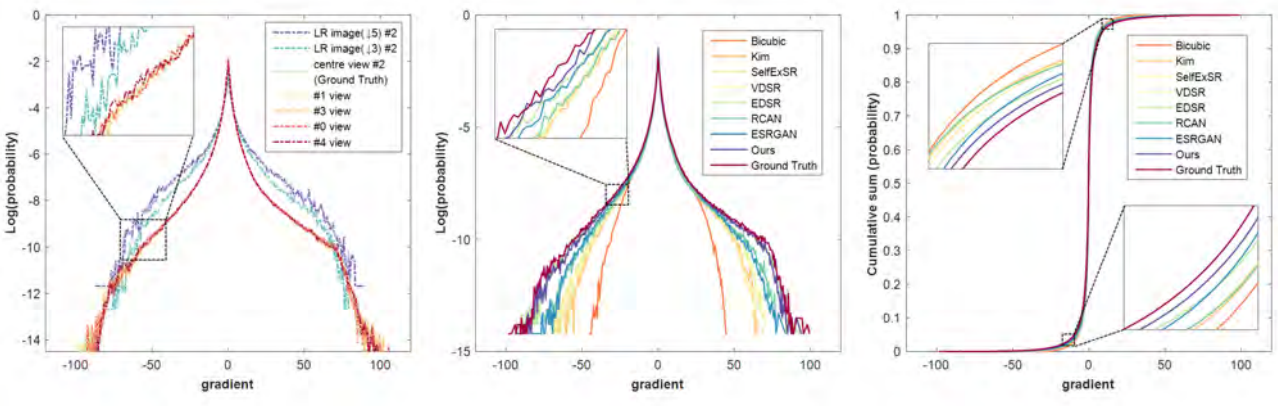


Fig. 9. Left: Gradient distribution of HR frames of different viewpoints; Middle and Right: Gradient distribution of reconstructed images using different SR methods and their cumulative sums.

distribution in the x direction, and it can be computed by

$$GR_x(\hat{I}^{HR}, I^{HR}) = GR_x^{ov}(\hat{I}^{HR}, I^{HR}) - GR_x^{nov}(\hat{I}^{HR}, I^{HR}). \quad (23)$$

Here, in Fig. 9, once we project the gradient distribution curves onto the horizontal direction, $GR_x^{ov}(\cdot,\cdot)$ and $GR_x^{nov}(\cdot,\cdot)$ are respectively the overlapped and non-overlapped percentages of $\hat{I}^{HR}$ with respect to I^{HR}. The diversity of gradient distribution $GR_y(\cdot,\cdot)$ in the y direction shares the same formulation.

The calculation of our gradient distribution metric against the ground truth is shown in the middle and right sub-figures of Fig. 9. This reveals that the reconstructed gradient distribution using our solution is very close to the ground truth, while those using the other methods are farther away. The top of Tab. VI presents more detailed results compared against other existing methods. The calculated values M_{gd} corresponding to the reconstructed gradient distribution demonstrate again that our solution outperforms other SR methods.

The proposed novel SR evaluation model is applicable to the practical MR multiview video systems, where there is no ground truth HR video of the target viewpoint. Note that here the PSNR and SSIM metrics are unsuitable without the exact color image of the target HR video. Nevertheless, as seen in the left sub-figure of Fig. 9, the gradient distribution curves of different HR videos from different viewpoints are very close, we thus can easily compute the desired gradient distribution curve of the target viewpoint, by averaging that of the given neighboring HR views with appropriate weighting using following formulation:

$$\overline{M}_{gd}(\cdot) = \sum_{i=1}^{n} \frac{1}{D_{c_i}} \{M_{gd}\}_{c_i} \Big/ \sum_{i=1}^{n} \frac{1}{D_{c_i}}. \quad (24)$$

Here D_{c_i} is the weighting factor of each SR metric obtained from Eq. 22. As an example, camera #2 has been chosen as the target camera view, we simply define D_{c_i} with the normalized physical distance between this view and its neighboring ones. After that, we can directly apply our evaluation model on the reconstructed video according to the calculated gradient distributions. The bottom sub-table of Tab. VI shows the evaluation result between our reconstructed image and the multiview images. Obviously, the value of $\overline{M}_{gd}$(#0, #1, #3, #4) is 0.887 using the formulation of Eq. 24, it is very similar to 0.894. This demonstrates that our metric is very efficient in those cases without ground truth.

F. Discussions

When employing our solution in extreme operational conditions, including video acquisition with a very sparse camera setup, complex lighting, unusual noise levels, or heavy lossy compression, the reconstructed result of our solution might be affected. To overcome these issues, our approach must be combined with appropriate multiview color correction, view synthesis, denoising and smart compression systems. Our current solution takes the RGB color distance to find the most similar patches in the reference images. Some recent matching methods using neural features [44] open an interesting door in this area. Adopting other potential metrics to improve both patch matching and representation accuracy could further increase the final reconstruction quality.

Our gradient-aware evaluation metric aims to investigate the gradient distribution of the reconstructed HR image. There might be some extreme cases when the gradient-aware scoring of the reconstructed HR image is high, while its texture is unexpected with respect to the LR image and the references. One potential solution is combining our metric with one more PSNR metric between the downsampled version of the reconstructed HR image and the given LR image; however, such a mixed style might affect the capacity to evaluate how much the HR gradient distribution is preserved.

Our current solution performs well on various widely-used multiview videos which are captured by typical multiview camera rigs. However, once there are no good references from the neighbouring HR views, the proposed method will produce low-quality reconstruction results. In this sense, a potential direction is to effectively exploit useful information from some external HR datasets such as CrossNet [43] and SRNTT [44].

Finally, our current solution needs to independently process the 3 color channels, and it spends about 600s for a HD-resolution image, from which about 83% of the computational costs are spent on the optimization procedure. Improving

TABLE VI
OBJECTIVE COMPARISON WITH OUR GRADIENT-AWARE EVALUATION METRIC, THE SR FACTOR IS 4 FOR THE *Knuffelgooi* SEQUENCE.

	gradient range (x direction)			RM_x	gradient range (y direction)			RM_y	M_{gd}
	GR_x^{ov}	GR_x^{nov}	GR_x		GR_y^{ov}	GR_y^{nov}	GR_y		
Ground Truth	1	0	1	0	1	0	1	0	1
Bicubic	0.296	0	0.296	2.117	0.325	0	0.325	2.464	0.675
ScSR [5]	0.409	0	0.409	2.080	0.442	0	0.442	2.091	0.693
Kim [50]	0.439	0	0.439	1.891	0.526	0	0.526	1.937	0.706
SelfExSR [14]	0.553	0	0.553	1.998	0.586	0	0.586	1.714	0.717
SRCNN [27]	0.503	0	0.503	1.765	0.568	0	0.568	1.762	0.719
VDSR [51]	0.567	0	0.567	1.694	0.608	0	0.608	1.610	0.731
SRResNet [29]	0.513	0	0.513	1.781	0.528	0	0.528	1.474	0.725
EDSR [52]	0.549	0	0.549	1.639	0.540	0	0.540	1.533	0.731
RCAN [53]	0.509	0	0.509	1.463	0.552	0	0.552	1.528	0.735
SRGAN [29]	0.590	0	0.590	1.851	0.612	0	0.612	1.653	0.727
ESRGAN [54]	0.602	0	0.602	1.706	0.604	0	0.604	1.493	0.736
SRNTT-ℓ_2 [44]	0.676	0	0.676	1.137	0.723	0	0.723	1.035	0.790
SRNTT [44]	0.753	0	0.753	0.715	0.767	0	0.767	1.692	0.787
CrossNet [43]	0.918	0	0.918	1.459	0.857	0	0.857	1.112	0.794
Ours	0.839	0	0.839	0.578	0.946	0	0.946	0.515	0.894
#1	1	0.223	0.777	0.623	1	0.074	0.926	0.560	0.871
#3	1	0.091	0.909	0.454	0.994	0.038	0.956	0.504	0.902
#0	1	0.223	0.777	0.633	1	0.074	0.926	0.552	0.871
#4	1	0.223	0.777	0.553	1	0.070	0.930	0.566	0.876
$\overline{M}_{gd}$(#0,#1,#3,#4)	—	—	—	—	—	—	—	—	0.887
$\overline{M}_{gd}$(#1,#3)	—	—	—	—	—	—	—	—	0.882
$\overline{M}_{gd}$(#0,#1)	—	—	—	—	—	—	—	—	0.871
$\overline{M}_{gd}$(#0,#4)	—	—	—	—	—	—	—	—	0.874

the run time performance with algorithm- and hardware-level acceleration is subject of future investigations.

V. CONCLUSION

This paper has presented a novel SR optimization method for MR multiview video. By constructing a LLE-based representing model, our solution can effectively learn the detailed textures from the given reference HR images. Moreover, our low-rank constrained regularization ensures that the proposed method avoids various visual artifacts. The proposed gradient-aware evaluation for multiview reconstruction, which has also been presented, is suitable for those cases when the ground truth is not available, can reasonably consider the spatial correlation of scene textures captured from different viewpoints. The comprehensive set of experiments demonstrate that our SR approach outperforms the state-of-the-art methods for MR multiview video.

ACKNOWLEDGMENT

We would like to thank the editor and anonymous reviewers for their insightful comments in improving the paper. This work was in part supported by NSFC (61972216), Tianjin NSF (18JCYBJC41300 and 18ZXZNGX00110), and the Open Project Program of State Key Lab of VR in Beihang University (VRLAB2019B04).

REFERENCES

[1] T. Richter, J. Seiler, W. Schnurrer, and A. Kaup, "Robust super-resolution for mixed-resolution multiview image plus depth data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 814–828, 2016.

[2] Z. Jin, T. Tillo, C. Yao, J. Xiao, and Y. Zhao, "Virtual-view-assisted video super-resolution and enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 467–478, 2016.

[3] D. C. Garcia, C. Dorea, and R. L. de Queiroz, "Super resolution for multiview images using depth information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1249–1256, 2012.

[4] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, p. 96, 2007.

[5] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010.

[6] C. Ren, X. He, Q. Teng, Y. Wu, and T. Q. Nguyen, "Single image super-resolution using local geometric duality and non-local similarity," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2168–2183, 2016.

[7] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, 2003.

[8] J. Van Ouwerkerk, "Image super-resolution survey," *IMAGE VISION COMPUT.*, vol. 24, no. 10, pp. 1039–1052, 2006.

[9] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *arXiv preprint arXiv:1902.06068*, 2019.

[10] Q. Shan, Z. Li, J. Jia, and C.-K. Tang, "Fast image/video upsampling," *ACM Trans. Graph.*, vol. 27, no. 5, p. 153, 2008.

[11] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3194–3205, 2012.

[12] J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, 2011.

[13] Q. Yuan, L. Zhang, and H. Shen, "Regional spatially adaptive total variation super-resolution with spatial information filtering and clustering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2327–2342, 2013.

[14] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. CVPR*, 2015, pp. 5197–5206.

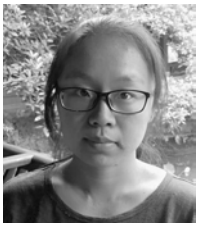
[15] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, 2004.

[16] Y.-R. Li, D.-Q. Dai, and L. Shen, "Multiframe super-resolution reconstruction using sparse directional regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 7, pp. 945–956, 2010.

[17] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1958–1975, 2009.
[18] A. W. Van Eekeren, K. Schutte, and L. J. Van Vliet, "Multiframe super-resolution reconstruction of small moving objects," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2901–2912, 2010.
[19] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. CVPR*, 2015, pp. 3791–3799.
[20] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, 2011.
[21] M. Zhang and C. Desrosiers, "High-quality image restoration using low-rank patch regularization and global structure sparsity," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 868–879, 2019.
[22] M. D. Robinson, C. A. Toth, J. Y. Lo, and S. Farsiu, "Efficient fourier-wavelet super-resolution," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2669–2681, 2010.
[23] H. Ji and C. Fermüller, "Robust wavelet-based super-resolution reconstruction: theory and algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 649–660, 2009.
[24] H. Demirel and G. Anbarjafari, "Image resolution enhancement by using discrete and stationary wavelet decomposition," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1458–1460, 2011.
[25] B. K. Gunturk, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of compressed video using transform-domain statistics," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 33–43, 2004.
[26] W. Yang, X. Zhang, Y. Tian, W. Wang, and J.-H. Xue, "Deep learning for single image super-resolution: A brief review," *arXiv preprint arXiv:1808.03344*, 2018.
[27] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.
[28] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, 2016.
[29] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017, pp. 4681–4690.
[30] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. CVPR*, 2017, pp. 3147–3155.
[31] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. CVPR*, 2018, pp. 2472–2481.
[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–14, 2018.
[33] M. Gao, S. Ma, D. Zhao, and W. Gao, "A spatial inter-view auto-regressive super-resolution scheme for multi-view image via scene matching algorithm," in *Proc. ISCAS*, 2013, pp. 2880–2883.
[34] Y. Li, X. Li, Z. Fu, and W. Zhong, "Multiview video super-resolution via information extraction and merging," in *Proc. ACM MM*, 2016, pp. 446–450.
[35] S.-P. Lu, B. Ceulemans, A. Munteanu, and P. Schelkens, "Spatio-temporally consistent color and structure optimization for multiview video color correction," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 577–590, 2015.
[36] B. Ceulemans, S.-P. Lu, G. Lafruit, P. Schelkens, and A. Munteanu, "Efficient mrf-based disocclusion inpainting in multiview video," in *Proc. ICME*, 2016, pp. 1–6.
[37] B. Ceulemans, S.-P. Lu, G. Lafruit, and A. Munteanu, "Robust multiview synthesis for wide-baseline camera arrays," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2235–2248, 2018.
[38] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 428–438, 2016.
[39] Y. Wang, Y. Liu, W. Heidrich, and Q. Dai, "The light field attachment: Turning a dslr into a light field camera using a low budget camera ring," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 10, pp. 2357–2364, 2017.
[40] H. Zheng, M. Guo, H. Wang, Y. Liu, and L. Fang, "Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system," in *Proc. ICCV Workshops*, 2017, pp. 2481–2486.
[41] V. Boominathan, K. Mitra, and A. Veeraraghavan, "Improving resolution and depth-of-field of light field cameras using a hybrid imaging system," in *Proc. ICCP*, 2014, pp. 1–10.
[42] J. Jiang, X. Ma, Chen, T. Lu, Z. Wang, and J. Ma, "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Trans. Multimedia*, pp. 15–26, 2017.
[43] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proc. ECCV*, 2018.
[44] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. CVPR*, 2019.
[45] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[46] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. CVPR*, 2014, pp. 2862–2869.
[47] P. Wang and R. Menon, "Computational spectroscopy via singular-value decomposition and regularization," *Optics express*, vol. 22, no. 18, pp. 21 541–21 550, 2014.
[48] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
[49] P. Kovacs, A. Fekete, K. Lackner, V. Adhikarla, and A. Zare, "[ftv ahg] big buck bunny light-field test sequences," *ISO/IEC JTC1/SC29/WG11, Doc. MPEG M35721*, 2015.
[50] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, 2010.
[51] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. CVPR*, 2016, pp. 1646–1654.
[52] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. CVPR Workshops*, 2017, pp. 136–144.
[53] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, 2018, pp. 286–301.
[54] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV*, 2018, pp. 63–79.
[55] R. Song, H. Ko, and C. Kuo, "MCL-3D: A database for stereoscopic image quality assessment using 2D-image-plus-depth source," *J. Inf. Sci. Eng.*, 2015.
[56] A. Vetro, M. McGuire, W. Matusik, A. Behrens, J. Lee, and H. Pfister, "Multiview video test sequences from merl," *ISO/IEC JTC1/SG29/WG11, Doc. MPEG2005/M12077*, 2005.

Shao-Ping Lu is currently an associate professor at Nankai University, China. Prior to that he was a postdoc and senior researcher at Vrije Universiteit Brussels (VUB) in Belgium. He received the Ph.D. degree in Computer Science from Tsinghua University, China. His research interests lie primarily in the intersection of visual computing, with particular focus on computational photography, 3D video representation, visual scene analysis and machine learning.

Sen-Mao Li is currently a Master student of computer science at Nankai University, Tianjin, China. His research interests include multiview signal super-resolution and machine learning.



Rong Wang is currently a Master student of computer science at Nankai University, Tianjin, China, where she received her Bachelor degree in electronic information science and technology in 2019. Her research interests include 3D video representation and machine learning.

Gauthier Lafruit is professor at l'Université Libre de Bruxelles (ULB), Belgium, in the Laboratory of Image Synthesis and Analysis (LISA). He received his Ph.D. degree in Electrical Engineering from the Vrije Universiteit Brussel, Belgium, in 1995. His current research includes Virtual Reality from camera captured content, Light Fields, Computational Imaging and GPU acceleration. Gauthier Lafruit served as Associate Editor for IEEE TCSVT. He is currently co-chair of the MPEG-I Visual working group.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012, then did 2 years research fellow in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision and image processing. He has published $50+$ refereed papers, with $15,000+$ Google Scholar citations. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, *etc*. He is an IEEE senior member and on the editor board of IEEE TIP.

Adrian Munteanu (M'07) is professor at the Electronics and Informatics (ETRO) department of the Vrije Universiteit Brussel (VUB), Belgium. He received the MSc degree in Electronics and Telecommunications from Politehnica University of Bucharest, Romania, in 1994, the MSc degree in Biomedical Engineering from University of Patras, Greece, in 1996, and the Doctorate degree in Applied Sciences (Summa Cum Laudae) from Vrije Universiteit Brussel, Belgium, in 2003. In the period 2004-2010 he was post-doctoral fellow with the Fund for Scientific Research Flanders (FWO), Belgium, and since 2007, he is professor at VUB. His research interests include image, video and 3D graphics compression, 3D video, deep-learning, distributed visual processing, error-resilient coding, and multimedia transmission over networks. Adrian Munteanu is the author of more than 350 journal and conference publications, book chapters, and contributions to standards and holds 7 patents in image and video coding. He is the recipient of the 2004 BARCO-FWO prize for his PhD work, the (co-)recipient of the Most Cited Paper Award from Elsevier for 2007, and of 10 other scientific prizes and awards at international conferences. Adrian Munteanu served as Associate Editor for IEEE Transactions on Multimedia and currently serves as Associate Editor for IEEE Transactions on Image Processing.