WLDISR: Weighted Local Sparse Representation-Based Depth Image Super-Resolution for 3D Video System

Huan Zhang, Yun Zhang, *Senior Member, IEEE*, Hanli Wang, *Senior Member, IEEE*, Yo-Sung Ho, *Fellow, IEEE*, and Shengzhong Feng

***Abstract*— In this paper, we propose a Weighted Local sparse representation based Depth Image Super-Resolution (WLDISR) schemes aiming at improving the Virtual View Image (VVI) quality of 3D video system. Different from color images, depth images are mainly used to provide geometrical information in synthesizing VVI. Due to the view synthesis characteristics difference between textural structures and smooth regions of depth images, we divide the depth images into edge and smooth patches and learn two local dictionaries, respectively. Meanwhile, the weight term is derived and incorporated explicitly in the cost function to denote different importance of edge structures and smooth regions to the VVI quality. Then, local sparse representation and weighted sparse representation are jointly used in both dictionary learning and reconstruction phases in depth image super-resolution. Based on different optimizations on learning and reconstruction modules, three WLDISR schemes, WLDISR-D, WLDISR-R, and WLDISR-ALL, are proposed. Experimental results on 3D sequences demonstrate that the proposed WLDISR-D, WLDISR-R, and WLDISR-ALL schemes can achieve more than 1.9-, 2.03-, and 2.16-dB gains on average, respectively, in terms of the VVIs' quality, as compared with the state-of-the-art schemes. In addition, the visual quality of VVIs is also improved.**

***Index Terms*— 3D video, depth image, super-resolution, sparse representation, virtual view image quality.**

Manuscript received September 29, 2017; revised March 20, 2018 and July 22, 2018; accepted August 11, 2018. Date of publication August 23, 2018; date of current version October 1, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61471348 and 61871372, in part by the Key Project for Guangdong Provincial Science and Technology Development under Grant 2017B010110014, in part by the Shenzhen Science and Technology Development Project under Grants JSGG20160229202345378 and JCYJ20170811160212033, in part by the Shenzhen International Collaborative Research Project under Grant GJHZ20170314155404913, in part by the Guangdong Natural Science Foundation for Distinguished Young Scholar under Grant 2016A030306022, in part by the Guangdong Special Support Program for Youth Science and Technology Innovation Talents under Grant 2014TQ01X345, and in part by the Membership of Youth Innovation Promotion Association, Chinese Academy of Sciences, under Grant 2018392. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Oleg V. Michailovich. *(Corresponding author: Yun Zhang.)*

H. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: huan.zhang@siat.ac.cn).

Y. Zhang and S. Feng are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn; sz.feng@siat.ac.cn).

H. Wang is with the Department of Computer Science and Technology, Tongji University, Shanghai 200092, China, and also with the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China (e-mail: hanliwang@tongji.edu.cn).

Y.-S. Ho is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500-712, South Korea (e-mail: hoyo@gist.ac.kr).

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.2018.2866959

I. Introduction

NOWADAYS, 3D and Free Viewpoint Video (FVV) system is becoming more and more prevalent, since it can provide interactive and immersive visual experiences at any viewpoint and angle for users. Multiview depth images, which reflect geometrical information of a 3D world scene, are one of the key components of 3D content. To enable the arbitrary view generation and interactive functionality of the 3D and FVV system, the multi-view depth videos shall be encoded and transmitted with the multi-view color videos to the clients. High quality and High-Resolution (HR) depth images are highly demanded in rendering high quality VVIs [1]–[3]. However, depth images captured by the current depth camera based on Time-of-Flight mechanism, are usually with very limited resolution compared with corresponding color images [4], [5]. Though depth images generated from stereo matching algorithms have the same high resolution as the color images, under the transmission bit rate constraints, reduced resolution depth image coding is often used in transmission [6]–[8]. In view of the above two typical situations, depth images with Low-Resolution (LR) are often adopted in 3D video system. Therefore, depth image Super-Resolution (SR) method is highly desired in order to improve the visual quality in 3D video system.

Many image SR methods have been developed recently. Yang *et al*. [9] proposed the groundbreaking work of image SR via sparse representation called Sparse Coding Super Resolution (ScSR), which was based on the assumption that the corresponding LR and HR patches share the same coefficients represented by the coupled LR-HR dictionary. Zeyde *et al*. [10] improved the dictionary learning method in [9] and reduced the dimension of the LR features, so as to improve the quality of the reconstructed HR images. In [11] and [12], inspired by neighbor embedding and sparse coding, a very effective and relatively much faster SR method was proposed. Based on [11] and [12], Zhang *et al*. [13]

incorporated the clustering and collaborative representation methods, and proposed an effective and faster SR method. In addition, there are also some deep-learning based image SR methods [4], [14]. However, it is not an effective way to directly apply image SR methods into depth image SR, since depth images have different characteristics from color images, e.g. more sharp edges and fewer textures. Depth image SR shall be specifically designed in view of the characteristics of depth images themselves.

Example-based SR methods have gained popularity in depth image SR recently. These methods include sparse representation based [9], [15], [16], Markov Random Field (MRF)-based [17], [18], and the neighbor embedding based [19] methods. The main idea of example-based SR methods is to learn LR-HR image priors from external or interior image patches, which helps reconstruct high frequency details from LR depth images with the learned image priors. Since sharp edges contain higher frequency details than smooth regions in depth images, it is more difficult to restore edges or textures in depth image SR. In order to alleviate this problem, some depth image SR methods focus on edges to reconstruct better quality edge structures [5], [15], [20]–[22]. Mandal *et al*. [5] trained multiple sub-dictionaries via K-means clustering and added an edge-preserving regularization term to localize the discontinuities in depth images. Liu *et al*. [15] applied a combined wavelet-contourlet dictionary in the depth image SR reconstruction and proposed an efficient depth-gradient related randomized sampling scheme. Ferstl *et al*. [20] employed the LR-HR patches to learn not only the LR-HR dictionary pairs but also the edge priors. The edge priors were then used as a regularization constraint in a variational SR. An edge-guided depth image SR method was proposed in [21], where a HR edge map was first generated based on the exemplar-based method via MRF framework. The edge map was then used as a guide to help up-sample the LR depth image through a modified joint bilateral filter. To sharpen edges and reduce the jagged noises in depth images, Xie *et al*. [22] added an adaptively regularized shock filter in reconstructing the HR depth image via the coupled dictionary learning. These methods put much emphasis on the edges in depth images and have achieved improvements in reconstructing the HR depth images. However, depth images are not used for watching directly in FVV but to provide geometrical information in rendering VVIs [1], [2]. In this case, image SR methods of maximizing the Peak Signal-to-Noise Ratio (PSNR) of depth images, cannot guarantee the efficiency in promoting the quality of VVIs.

In addition, several depth image SR works have further taken the quality of VVI into consideration [23]–[25]. Hu *et al*. [23] used the original texture image of one single view and the corresponding LR depth image to synthesize the neighboring texture image. Then, the error between the original and the synthesized neighboring texture image was used as a regularization term in the depth image SR of this single view, taking advantage of multiple views to enhance the quality of VVIs. A patch-based SR method was proposed in [24] by using the synthesis error as a criterion to select the best SR result out of various SR methods. To make the LR depth image values more reliable, Lei *et al*. [25] first proposed a credibility based multi-view depth images fusion strategy which took both the VVI quality and interview correlation into consideration. A VVI quality oriented trilateral depth-image SR method was then proposed, which incorporated VVI quality as well in the weighting coefficient of the SR filter. These methods have achieved good performance in improving VVI quality. However, these methods didn't consider different view synthesis characteristics of texture and smooth regions in depth image SR.

Recently, Zhang *et al*. [26] proposed multi-view depth video coding and bit allocation optimization schemes considering view synthesis characteristics of regions with different textures. In the inspiration of Zhang *et al*. [26], we distinguish edge regions from smooth regions and consider view synthesis characteristics for these two regions in depth image SR. In this paper, we propose a Weighted Local sparse representation based Depth Image Super-Resolution (WLDISR) method. First of all, different from previous depth image SR methods, our goal is to maximize the VVI quality rather than the depth images quality in depth image SR. Towards this goal, the view synthesis distortion model is incorporated into the optimization objective function. Moreover, due to different view synthesis characteristics, edge and smooth regions are reconstructed separately, and the view synthesis distortion models with corresponding parameters are employed. The weighted terms are derived for the two regions accordingly. Local sparse representation and weighted sparse representation are then assembled in dictionary learning and reconstruction phases in depth image SR. Lastly, three WLDISR schemes are proposed based on different optimizations on learning and reconstruction modules. The rest of this paper is organized as follows. Our proposed schemes are presented in Section II. Then, detailed experimental results and analyses are elaborated and presented in Section III. In addition, the effects of some key factors are analyzed and discussions are described in Section IV. Finally, conclusions are drawn in Section V.

II. The Proposed WLDISR

A. Proposed WLDISR Framework

3D and FVV video system mainly consists of six major components: 3D and multiview video acquisition, encoding, transmission, decoding, view generation, and display [2], [26]. The mainstream data format of 3D system is Multiview Video plus Depth (MVD), i.e. multiview color and depth video. Multiview color video is generated by multiple cameras with HR texture images. Multiview depth video is composed of LR depth images captured by multiple depth cameras or less precise HR depth images generated by stereo matching based algorithms. These MVD are encoded at the server, and transmitted to the client. At the client, they are decoded, and the decoded MVD are used to synthesize intermediate VVIs through Depth Image Based Rendering (DIBR) technology. Due to the reduced resolution depth coding and limited resolution of depth camera, depth image SR is often required. Therefore, in this paper, we propose a depth image SR method aiming at improving the VVI quality in 3D video system.

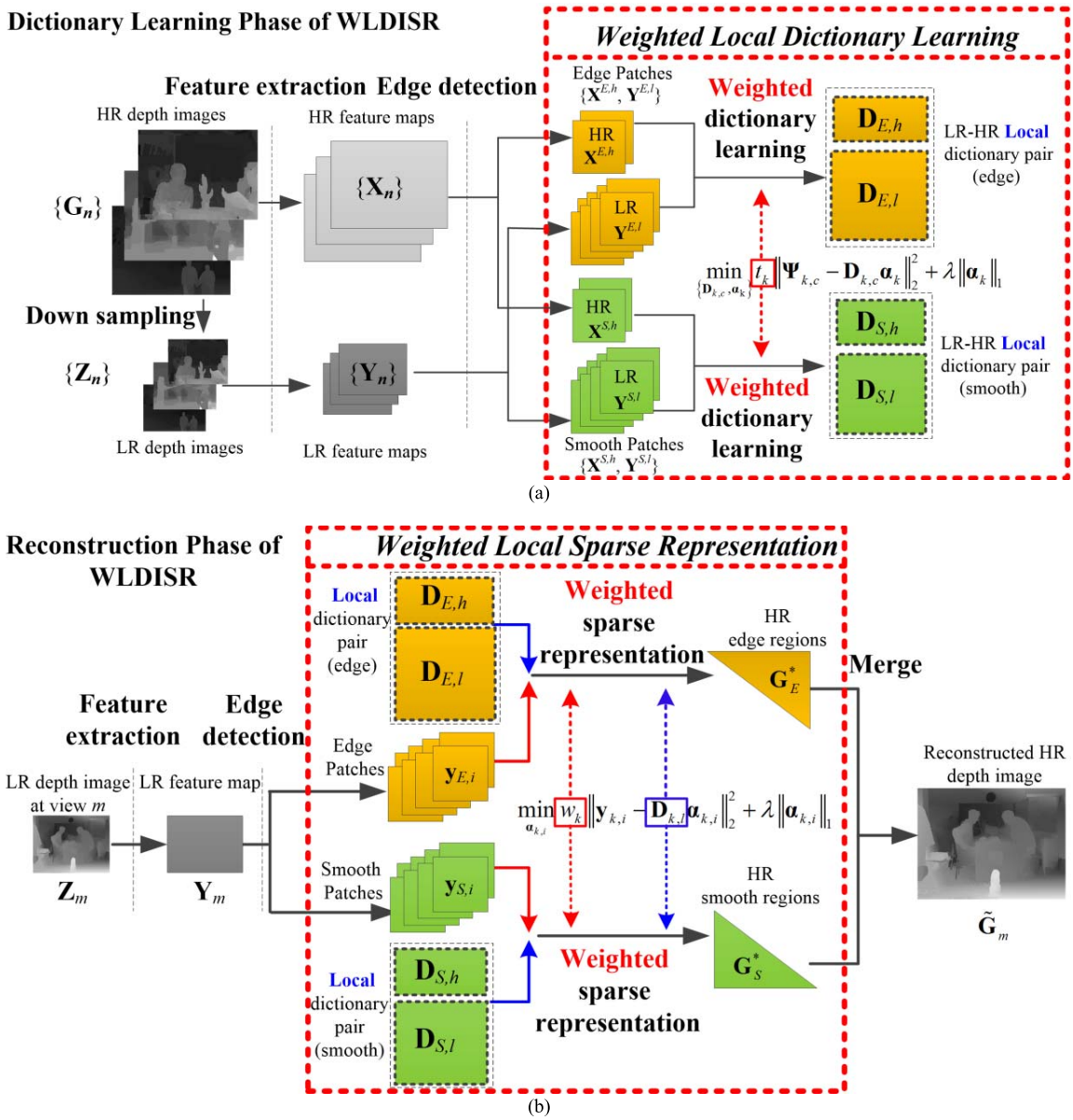


Fig. 1. The framework of WLDISR. (a) Dictionary learning phase; (b) Reconstruction phase.

Fig. 1 shows the framework of our proposed WLDISR depth image SR method, which has two major components: dictionary learning phase, and reconstruction phase. We employ local and weighted sparse representation jointly for edge and smooth regions in both dictionary learning and reconstruction phases.

The framework of the dictionary learning phase is shown in Fig. 1(a). In the dictionary learning phase, the inputs are a set of depth images of HR-LR image pairs from several 3D sequences denoted as $\{\mathbf{G}_n, \mathbf{Z}_n\}$. The LR image $\mathbf{Z}_n$ is down-sampled from the HR image $\mathbf{G}_n$ by bicubic interpolation method, and these LR images are up-sampled to the same resolution as HR depth images by bicubic method as well. HR and LR feature maps $\{\mathbf{X}_n, \mathbf{Y}_n\}$ are extracted from HR-LR image pairs $\{\mathbf{G}_n, \mathbf{Z}_n\}$ by feature extraction [9]. The extracted HR and LR feature maps are divided into edge and smooth feature patches set $\{\mathbf{X}^h, \mathbf{Y}^l\}$ respectively after edge detection of LR images. The HR and LR feature patches are classified as edge or smooth feature patches based on the number of edge pixels in the corresponding LR patches after canny edge detection. The patch size is set as 5×5 and the patches are classified as edge patches if the number of edge pixels is larger than 1. In addition, only patches with patch variance greater than 10 are kept for training. Afterwards, the remaining edge and smooth feature patches pairs are organized as LR-HR pairs to train the coupled LR-HR edge and smooth dictionaries $\{\mathbf{D}_E, \mathbf{D}_S\}$, where $\mathbf{D}_E$ and $\mathbf{D}_S$ consist of the dictionary pairs $\{\mathbf{D}_{E,l}, \mathbf{D}_{E,h}\}$ and $\{\mathbf{D}_{S,l}, \mathbf{D}_{S,h}\}$, respectively.

Fig. 1(b) shows the reconstruction phase of WLDISR. Given an LR depth image $\mathbf{Z}_m$, LR feature map $\mathbf{Y}_m$ is then generated from $\mathbf{Z}_m$. After edge detection, $\mathbf{Y}_m$ is divided into overlapped LR edge and smooth feature patches. Then, HR edge and

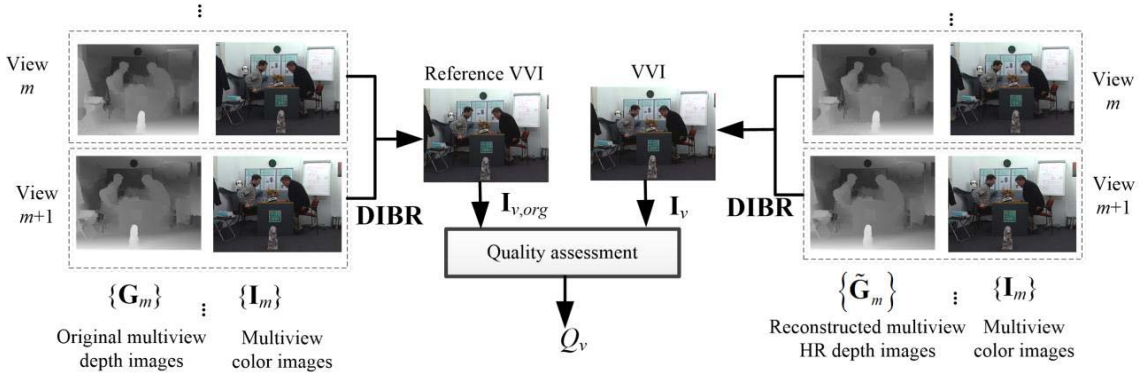


Fig. 2. The validation process of WLDISR.

smooth patches are reconstructed from LR edge and smooth patches through dictionaries $\mathbf{D}_E$, $\mathbf{D}_S$, respectively. HR edge and smooth patches are then merged to reconstruct the final HR depth image $\tilde{\mathbf{G}}_m$.

At the validation process, for any two views, e.g. view m and view $m+1$, the reconstructed HR depth images $\tilde{\mathbf{G}}_m$ and $\tilde{\mathbf{G}}_{m+1}$ generated from the reconstruction phase are combined with the corresponding HR color images, $\mathbf{I}_m$ and $\mathbf{I}_{m+1}$, to synthesize the intermediate VVI $\mathbf{I}_v$ via the DIBR module. Meantime, via the DIBR module, original HR depth images $\mathbf{G}_m$, $\mathbf{G}_{m+1}$ combining with HR color images $\mathbf{I}_m$, $\mathbf{I}_{m+1}$ are used to synthesize the VVI $\mathbf{I}_{v,org}$, which is used as a reference VVI. Finally, the quality of $\mathbf{I}_v$, denoted as Q_v, is calculated based on the comparison between the rendered $\mathbf{I}_v$ and reference $\mathbf{I}_{v,org}$. The overview of validation process is shown in Fig. 2.

To improve the VVI quality Q_v, we develop three depth image SR schemes. WLDISR-D and WLDISR-R schemes are proposed to optimize the dictionary learning and reconstruction modules individually. In addition, the WLDISR-ALL scheme is developed to optimize both dictionary learning and reconstruction modules with the weighted local sparse representation. Overall, these schemes will be presented in detail in the following subsections.

B. Dictionary Learning Phase of WLDISR

In this subsection, we first describe the dictionary learning process of the WLDISR scheme, and then determine the optimal weight for the dictionary learning in WLDISR.

Instead of using the HR/LR depth image pairs $\{\mathbf{G}_n, \mathbf{Z}_n\}$ directly, we use their corresponding feature maps $\{\mathbf{X}_n, \mathbf{Y}_n\}$ in the dictionary learning, which are divided into patch sets $\{\mathbf{X}^h, \mathbf{Y}^l\}$. $\mathbf{X}^h = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N\}$ represents the set of HR depth image feature patches, and $\mathbf{Y}^l = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_N\}$ represents the set of LR depth image feature patches. N is the total number of HR/LR depth image feature patches. Note that each HR depth image feature patch $\mathbf{x}_i$ is obtained by subtracting the mean value of each patch in HR depth image $\mathbf{G}_n$; each LR depth image feature patch $\mathbf{y}_i$ comes from the LR feature map $\mathbf{Y}_n$, which is acquired by using high pass filter directly on the interpolated LR depth image $\mathbf{Z}_n$ [9]. The dictionary learning objective function for depth images using ScSR [9] can be formulated as

$$\begin{cases} \min\limits_{\{\mathbf{D}_h, \mathbf{D}_l, \boldsymbol{\alpha}_i\}} \left(\sum\limits_i \sum\limits_{\phi} \Delta_{\phi,i} + \lambda \left\| \boldsymbol{\alpha}_i \right\|_1 \right) \\ \Delta_{h,i} = \dfrac{1}{u} \left\| \mathbf{x}_i - \mathbf{D}_h \boldsymbol{\alpha}_i \right\|_2^2, \Delta_{l,i} = \dfrac{1}{v} \left\| \mathbf{y}_i - \mathbf{D}_l \boldsymbol{\alpha}_i \right\|_2^2 \end{cases}, \quad (1)$$

where $\{\mathbf{x}_i, \mathbf{y}_i\}$ represents the i-th HR and LR training image feature patch pair, $\{\mathbf{D}_h, \mathbf{D}_l\}$ represents HR and LR dictionary pair. $\|\boldsymbol{\alpha}_i\|_1$ is the sparsity term with L_1 norm, $\boldsymbol{\alpha}_i$ is the sparse coefficient of patch pair $\{\mathbf{x}_i, \mathbf{y}_i\}$, and λ regulates the sparsity. $\phi \in \{h, l\}$. 'h' denotes HR, and 'l' denotes LR. The fidelity term $\Delta_{\phi,i}$ is the difference between the i-th original and reconstructed HR/LR depth image feature patch, i.e. $\mathbf{x}_i/\mathbf{y}_i$ and $\mathbf{D}_h \boldsymbol{\alpha}_i/\mathbf{D}_l \boldsymbol{\alpha}_i$. In fact, $\Delta_{\phi,i}$ represents the depth image feature patch distortion. u and v are the dimensions of HR and LR depth image feature patches $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively.

The objective in (1) is to minimize the reconstructed HR/LR depth image distortion subject to sparsity. However, the depth image is mainly used as the geometrical information for view rendering in 3D video system instead of being watched directly. Thus, the quality of VVIs that rendered from the depth images shall be considered in learning dictionaries for depth images, which can be formulated as

$$\begin{cases} \min\limits_{\{\mathbf{D}_h, \mathbf{D}_l, \boldsymbol{\alpha}_i\}} \left(\sum\limits_i \sum\limits_{\phi} \Lambda_{\phi,i} + \lambda \left\| \boldsymbol{\alpha}_i \right\|_1 \right), \\ \Lambda_{\phi,i} = F\left(\Delta_{\phi,i}\right), \end{cases} \quad (2)$$

where $\Lambda_{\phi,i}$ is the VVI feature difference from reconstruction in dictionary learning. $F(\cdot)$ is a function mapping the depth image feature difference $\Delta_{\phi,i}$ to the VVI feature difference $\Lambda_{\phi,i}$.

Fortunately, the relationship between VVIs distortion and depth images distortion has been explored in [26]–[29]. These works include the allowable depth distortion model in view synthesis [28], and view synthesis distortion models considering regional selective properties of depth images [26], and color-depth joint distortions [27], [29]. In this paper, we use the view synthesis distortion model proposed by Zhang *et al.* [26], in which the relationship between depth image distortion and VVI distortion is analyzed for edge and smooth regions. Over all the regions, edge regions, and smooth regions in depth images, the VVI distortion measured by

Mean Squared Error (MSE), MSE_{V}, can be approximated as a linear model of depth distortion measured by MSE, MSE_{D}, which can be presented as [26]

$$MSE_V = t_\varphi MSE_D + \varepsilon_\varphi, \tag{3}$$

where t_φ represents the view synthesis weight for region φ, and ε_ϕ is a constant denoting initial VVI distortion from DIBR. $\varphi \in \{ALL,\ E,\ S\}$, where 'ALL' denotes the entire image, 'E' denotes edge regions, and 'S' denotes smooth regions. t_φ is content dependent and correlates with camera parameters, rendering positions, distortions in color image, and video contents. For one sequence, t_E is usually larger than t_S since the distortions in edge regions have severer impacts on VVI quality than those in smooth regions. Note that (3) is a mapping function from depth distortion MSE_{D} to VVI distortion MSE_{V} in the image domain. Since the feature map extraction for $\{\mathbf{X}^h,\ \mathbf{Y}^l\}$ is a linear process, the relationship between the quality of VVI feature maps and depth image feature maps could be approximated as a linear function through experiments, as illustrated in Appendix. Thus, the relationship in (3) is also applicable to map the patch-wise depth image feature difference $\Delta_{\phi,i}$ to the VVI feature difference $\Lambda_{\phi,i}$, which is in feature domain and can be presented as

$$\Lambda_{\phi,i} = F\left(\Delta_{\phi,i}\right) = t_\varphi \Delta_{\phi,i} + \varepsilon_\varphi. \tag{4}$$

Apply (4) to (2), and the VVI quality oriented dictionary learning objective function for depth images can be written as

$$\min_{\{\mathbf{D}_h, \mathbf{D}_l, \boldsymbol{\alpha}_i\}} \left(\sum_i \sum_\phi t_{ALL} \Delta_{\phi,i} + \lambda \left\| \boldsymbol{\alpha}_i \right\|_1 \right), \tag{5}$$

where t_{ALL} is the weighting factor that transforms $\Delta_{\phi,i}$ to $\Lambda_{\phi,i}$ for an entire depth image. When t_{ALL} equals to 1, (5) will be degraded to (1), which is the dictionary learning for depth images by using ScSR [9].

Due to different view synthesis characteristics of different texture regions [26], VVI feature map distortions or VVI distortions of edge and smooth regions should be considered separately in depth image SR. It means the process of learning dictionaries shall be considered separately. In [30] and [31], local structures or block areas of a face were constrained to share the same dictionary atoms of a dictionary or represented by a local dictionary. The local structures or patches sharing the similar characteristics could be locally represented by local dictionaries. In order to exploit different view synthesis characteristics of edge and smooth regions, the HR and LR edge and smooth patches shall be trained to learn local edge and smooth HR-LR dictionaries separately.

We divide training feature patch sets $\{\mathbf{X}^h,\ \mathbf{Y}^l\}$ into two subsets $\{\mathbf{X}^{k,h},\ \mathbf{Y}^{k,l}\}$, where $k \in \{E,\ S\}$. 'E' denotes edge region, 'S' denotes smooth region. Let $\mathbf{X}^{k,h} = \{\mathbf{x}_{k,1},\ \mathbf{x}_{k,2},\ \mathbf{x}_{k,3}, \ldots, \mathbf{x}_{k,N(k)}\}$ represent the set of HR depth image feature patches of region k, where $N(k)$ is the number of patches for region k, and let $\mathbf{Y}^{k,l} = \{\mathbf{y}_{k,1},\ \mathbf{y}_{k,2},\ \mathbf{y}_{k,3}, \ldots, \mathbf{y}_{k,N(k)}\}$ represent the set of LR depth image feature patches of region k. To learn the dictionary for region k, VVI quality oriented local dictionary learning objective can be derived as

$$\begin{cases} \min\limits_{\{\mathbf{D}_{k,h}, \mathbf{D}_{k,l}, \boldsymbol{\alpha}_{k,i}\}} \left(\sum\limits_i \sum\limits_\phi F_k\left(\Delta_{\phi,k,i}\right) + \lambda \left\| \boldsymbol{\alpha}_{k,i} \right\|_1 \right) \\ \Delta_{h,k,i} = \dfrac{1}{u} \left\| \mathbf{x}_{k,i} - \mathbf{D}_{k,h} \boldsymbol{\alpha}_{k,i} \right\|_2^2 \\ \Delta_{l,k,i} = \dfrac{1}{v} \left\| \mathbf{y}_{k,i} - \mathbf{D}_{k,l} \boldsymbol{\alpha}_{k,i} \right\|_2^2 \end{cases}, \tag{6}$$

where $\Delta_{\phi,k,i}$ represents the distortion of the i-th patch in the region k of LR/HR depth images, $\{\mathbf{x}_{k,i},\ \mathbf{y}_{k,i}\}$ represents the i-th LR and HR training depth feature patch pair of region k. $\{\mathbf{D}_{k,l},\ \mathbf{D}_{k,h}\}$ represents LR and HR dictionary pair for region k. $F_k(\cdot)$ is a mapping function from the depth feature map distortion to the VVI feature map distortion for region k. According to (4), we find that the linear relationship is applicable for edge and smooth regions, i.e., $F_k(\cdot)$ is the $F(\cdot)$ with different t_k.

Therefore, applying (4) to (6), the VVI feature map quality oriented objective function of dictionary learning for edge and smooth regions can be written as

$$\min_{\{\mathbf{D}_{k,h}, \mathbf{D}_{k,l}, \boldsymbol{\alpha}_k\}} \left\{ t_k \left\| \boldsymbol{\Psi}_{k,c} - \mathbf{D}_{k,c} \boldsymbol{\alpha}_k \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha}_k \right\|_1 \right\}, \tag{7}$$

where $\boldsymbol{\Psi}_{k,c}$ equals to $\left[1/\sqrt{u}\mathbf{X}^{k,h}, 1/\sqrt{v}\mathbf{Y}^{k,l}\right]^T$, representing the set of LR and HR training depth image feature patch pairs for region k, and $\mathbf{D}_{k,c}$ equals to $\left[1/\sqrt{u}\mathbf{D}_{k,h}, 1/\sqrt{v}\mathbf{D}_{k,l}\right]^T$, representing LR and HR edge or smooth dictionary pairs. $\boldsymbol{\alpha}_k$ is a simplified form of the sparse coefficients $\boldsymbol{\alpha}_{k,i}$ for all the patches in region k. t_k denotes different impacts of the reconstruction loss of edge or smooth patches on the VVI feature map quality. It may have similar effects as the regularization parameter λ does. t_k plays the role like those weighted terms employed in [32]–[34], in which weights were added into data fidelity terms or dictionary atoms to strengthen different contributions of data or dictionary atoms. Generally, (7) learns two different local dictionaries with different weight t_k and learning patches, which is regarded as the weighted local dictionary learning in this paper. Moreover, feature-sign search algorithm and Lagrange dual algorithm [35] are used to solve the L_1-regularized least squares problem and L_2-constrained least squares problem in (7), respectively.

To learn an optimal dictionary for the depth image, we shall determine t_k. The relationship between the weights t_E, t_S and VVI quality Q_v were experimentally analyzed. We set a number of different weight sets $\{t_E,\ t_S\}$, which were then used to learn a number of different local dictionaries. Since the reconstruction of edge regions is independent from that of smooth regions, we optimized the dictionary learning of edge and smooth regions individually. Candidate t_k, $k \in \{E,\ S\}$, was set in the range $\{t_k | -2.00 \le \log_{10} t_k \le 1.10\}$, i.e. $\{t_k | 0.01 \le t_k \le 12.60\}$. These learned dictionaries were then used to reconstruct the HR depth images by using the ScSR. Then, the reconstructed HR depth images were used in rendering the VVI, $\mathbf{I}_v$, at the validation phase.

The second to eighth rows of Table I show the configurations for the training sequences, including Kendo, Lovebird1, Newspaper, PoznanHall2, PoznanStreet, Shark, and Undo-Dancer. They have various contents and resolutions, and also

TABLE I
PROPERTIES AND SETTINGS FOR WLDISR TRAINING, VALIDATION, AND TESTING

3D Sequences	Resolution	Depth Image Generation Method	Views	Rendered View	Training Frame	Validation Frame	Testing-Short Term Frames	Testing-Long Term Frames
Lovebird1	1024×768	stereo matching	4, 6	5	5^{th}	/	101^{th}~110^{th}	/
Kendo		stereo matching	1, 3	2	5^{th}	/	101^{th}~110^{th}	1^{th}~200^{th}
Newspaper		depth cameras	2, 4	3	5^{th}	/	101^{th}~110^{th}	/
PoznanHall2	1920×1088	stereo matching	5, 7	6	5^{th}	/	101^{th}~110^{th}	/
Shark		computer graphics	1, 5	3	5^{th}	/	101^{th}~110^{th}	/
Undodancer		computer graphics	1, 3	2	5^{th}	/	101^{th}~110^{th}	1^{th}~200^{th}
Poznanstreet		stereo matching	3, 5	4	5^{th}	/	/	/
Balloons	1024×768	stereo matching	1,3	2	/	100^{th}	101^{th}~110^{th}	/
Bookarrival		stereo matching	8, 10	9	/	100^{th}	90^{th}~99^{th}	1^{th}~100^{th}
GhostTownFly	1920×1088	computer graphics	1, 5	3	/	100^{th}	101^{th}~110^{th}	/
Café	1920×1080	depth cameras	2, 4	3	/	/	101^{th}~110^{th}	1^{th}~200^{th}
Poznan_CarPark	1920×1088	stereo matching	3, 5	4	/	/	101^{th}~110^{th}	/

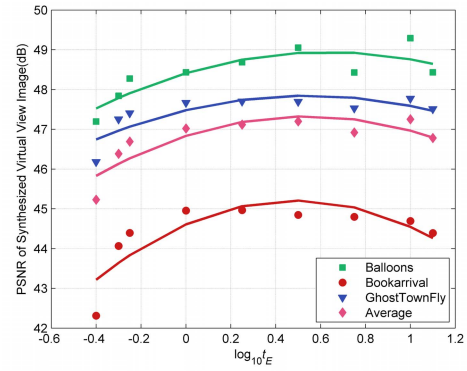
Note that symbol "/" indicates it is not used in training, validation or testing.

generated by different depth generation methods, including stereo matching, computer graphic, or captured by depth camera. Two views of each sequence listed in the fourth column and one frame of each view listed in the sixth column were used in learning the dictionaries, thus 14 depth images in total were used for training. The 14 depth images are enough for the dictionary training for two main reasons: one is that the number of the valid edge and smooth patches obtained from these 14 depth images is up to 35,000 and 6,000 respectively, which is sufficient for edge and smooth dictionary training; the other is these 14 depth images possess the diversity since they cover different spatial resolutions, depth image generation methods, image contents, and so on. The ninth to eleventh rows of Table I list the related configurations of the three sequences, i.e. GhostTownFly, Balloons, and Bookarrival, which were adopted in the reconstruction and validation phases. Two views of each sequence, and one frame, i.e. 100^{th} frame, of each view were used in validation. Note that the original ScSR method was used in reconstructing the HR depth image in order to analyze the performance of the learned dictionaries. In this depth image SR experiment, the scaling factor was 2. The intermediate VVIs were synthesized from two views of the reconstructed HR depth images and color images by using DIBR algorithm. Meanwhile, VVIs synthesized with the color images and the original depth images of each sequence were taken as reference VVIs for quality comparison.

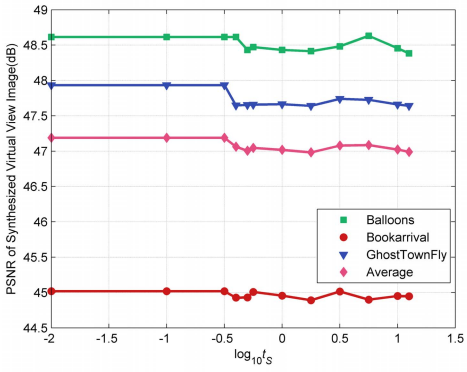
Fig. 3(a) illustrates the relationship between the weights t_E and VVI quality, where the y-axis is PSNR of VVI and the x-axis is $\log t_E$. It can be observed that the curve of each sequence and the average curve of three sequences can be approximated as a quadratic model, and the quadratic model used to fit each curve can be formulated as

$$Q_v = f(r_E) = ar_E^2 + br_E + c, \tag{8}$$

where r_E equals to $\log t_E$, a, b, and c are model parameters. Q_v is the VVI quality and $f(.)$ is a mapping function from r_E to VVI quality Q_v. The fitting R-square for the average curve is 0.73. Actually, other fitting algorithms, such as higher rank polynomial functions, can be used to achieve higher fitting



(a)



(b)

Fig. 3. Relationship between the weights $\{t_E, t_S\}$ and VVI quality at the dictionary learning phase. (a) Edge patches, (b) Smooth patches.

accuracy. However, to prevent over-fitting and obtain more reliable results, we use this quadratic model in (8). By taking the derivative of $f(r_E)$ with respect to r_E and setting its value as zero, we then get the optimal weight of r_E, 0.53, for the average curve. Since the difference between the optimal weights of each sequence is slight, we adopt this optimal r_E to learn the dictionaries for all depth images for simplicity. Accordingly, the optimal weight t_E for the dictionary learning, denoted as $t_{E,D}$, is $10^{0.53}$, i.e. 3.40.

In addition, the relationship between t_S and VVI quality is also analyzed, as shown in Fig. 3(b). The x-axis is the $\log t_S$ and y-axis is the quality of VVI generated with the reconstructed HR depth images. We can observe that when $\log t_S \in [-2.00, -0.50]$, Q_v of the four curves is consistent and relatively higher as compared with those larger $\log t_S$. There are small variations when $\log t_S$ is larger than -0.50. When $\log t_S \in [-2.00, -0.50]$, it is observed from experiment that the learned dictionary is made up of atoms with all zeros, which results from the small t_S. The corresponding HR depth images are actually reconstructed from the average value of LR image patches, which is similar to an averaging operation. It makes sense since the smooth region of depth images has much less texture and doesn't bother to use a dictionary to represent. We denote the optimal weight t_S, which is actually in the range of $\{t_S | 0.01 \leq t_S \leq 0.32\}$, as 'AvgLR' for learning smooth dictionary.

The dictionaries of edge and smooth regions are learned individually with the weighted local sparse representation. When only the dictionary learning module is optimized with WLDISR and the reconstruction phase uses ScSR, we denote this scheme as WLDISR-D.

C. Reconstruction Phase of WLDISR

At the reconstruction stage, the original LR image $\mathbf{Z}$ is first interpolated to the up-sampled LR image $\mathbf{Z_{up}}$ with the same resolution as the targeted resolution. Then, a LR feature map $\mathbf{Y}$ is extracted from the up-sampled LR image $\mathbf{Z_{up}}$, and $\mathbf{Z_{up}}$ and $\mathbf{Y}$ are divided into overlapped patches $\mathbf{z}_{up,i}$ and $\mathbf{y}_i$ respectively in a same partition way where $\mathbf{z}_{up,i}$ collocates with $\mathbf{y}_i$. The optimal coefficients $\{\boldsymbol{\alpha}_i^*\}$ for these overlapped LR feature patches $\{\mathbf{y}_i\}$ can be obtained by solving the following optimization problem

$$\boldsymbol{\alpha}_i^* = \arg\min_{\boldsymbol{\alpha}_i} \sum_i \|\mathbf{y}_i - \mathbf{D}_l \boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1, \tag{9}$$

where $\mathbf{D}_l$ is the LR dictionary, $\boldsymbol{\alpha}_i$ is the sparse coefficient for patch $\mathbf{y}_i$. LASSO [36] is employed to solve (9). Then, based on the optimal coefficients $\boldsymbol{\alpha}_i^*$ and the dictionary pair $\{\mathbf{D}_l, \mathbf{D}_h\}$, $\mathbf{D}_h \boldsymbol{\alpha}_i^*$, namely the reconstructed HR depth image feature patch $\mathbf{x}_i$, which corresponds to the LR feature patch $\mathbf{y}_i$, can be obtained. The associated HR patch denoted as $\mathbf{g}_i$ can be constructed as

$$\mathbf{g}_i = \mathbf{D}_h \boldsymbol{\alpha}_i^* + \mathbf{g}_i^0, \tag{10}$$

where $\mathbf{D}_h \boldsymbol{\alpha}_i^*$ can be regarded as the reconstructed texture part of $\mathbf{g}_i$. $\mathbf{g}_i^0$ is the mean patch with each pixel as g_i^0, which is calculated based on the mean value of LR image patch $\mathbf{z}_{up,i}$. All the overlapped HR patches $\mathbf{g}_i$ will be merged into an initial HR image $\mathbf{G}_0$. Then, more delicate HR solution $\mathbf{G}^*$ can be iteratively updated using back-projection method while minimizing the difference between down-sampled $\mathbf{G}$ and the original LR image $\mathbf{Z}$. This process can be expressed as

$$\mathbf{G}^* = \arg\min_{\mathbf{X}} \|\mathbf{HG} - \mathbf{Z}\|_2^2, \tag{11}$$

where $\mathbf{H}$ is a composite operator of down-sampling and blurring operations. (11) is solved by using gradient descent method.

Since the depth image is not viewed directly but used to synthesize the VVIs, the VVI quality shall also be considered in the reconstruction stage. Meanwhile, due to different view synthesis characteristics between edge and smooth regions, their dictionary pairs and reconstruction objectives shall be used and developed individually. Though there are minor mutual effects between edge and smooth regions during back-projection process, the reconstruction of the edge regions can be generally deemed as independent from that of smooth regions. Thus, given the original LR depth image at view m, $\mathbf{Z}_m$, it is divided into the edge and smooth regions, which is denoted as $\mathbf{Z}_{m,k}$, $k \in \{E, S\}$. Accordingly, the corresponding feature map of $\mathbf{Z}_{m,k}$, is $\mathbf{Y}_{m,k}$. Therefore, similar to (9)-(11), the VVI quality oriented depth image reconstruction process for region k can be presented as

$$\boldsymbol{\alpha}_{k,i}^* = \arg\min_{\boldsymbol{\alpha}_{k,i}} \left\{ \sum_i F\left(\left\| \mathbf{y}_{k,i} - \mathbf{D}_{k,l} \boldsymbol{\alpha}_{k,i} \right\|_2^2 \right) + \lambda \left\| \boldsymbol{\alpha}_{k,i} \right\|_1 \right\}, \tag{12}$$

$$\mathbf{g}_{k,i} = \mathbf{D}_{k,h} \boldsymbol{\alpha}_{k,i}^* + \mathbf{g}_{k,i}^0, \tag{13}$$

$$\mathbf{G}_k^* = \arg\min_{\mathbf{G}_k} F\left(\left\| \mathbf{H}_k \mathbf{G}_k - \mathbf{Z}_{m,k} \right\|_2^2 \right), \tag{14}$$

where $\mathbf{y}_{k,i}$ is the i-th LR depth image feature patch of region k. Here, $F(\cdot)$ is used to map the LR depth feature map distortion to the LR VVI feature map distortion for region k. $\mathbf{g}_{k,i}$ is the i-th reconstructed HR image patch of region k in a depth image and $g_{k,i}^0$ is the value of each pixel in the mean patch $\mathbf{g}_{k,i}^0$ of $\mathbf{g}_{k,i}$. $\mathbf{H}_\mathrm{k}$ is the composited down-sampling and blurring operator for region k. $\mathbf{Z}_{m,k}$ represents region k of the input LR depth image at view m, i.e. $\mathbf{Z}_m$ in Fig. 1(b). $\mathbf{G}_k^*$ is the reconstructed HR depth image of region k. Finally, after obtaining $\mathbf{G}_k^*$, $k \in \{E, S\}$, the regions $\mathbf{G}_E^*$ and $\mathbf{G}_S^*$ are non-overlapped and can be merged to form the final reconstructed HR depth image $\tilde{\mathbf{G}}_m$.

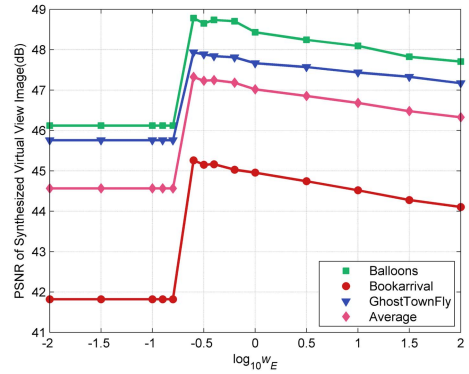
Similar to the local dictionary learning presented in Section II.B, $F(\cdot)$ also holds true at the depth image reconstruction phase. Apply (4) to (12), and we can obtain the sparse coefficient for patch i of region k by

$$\boldsymbol{\alpha}_{k,i}^* = \arg\min_{\boldsymbol{\alpha}_{k,i}} \left\{ \sum_i w_k \left\| \mathbf{y}_{k,i} - \mathbf{D}_{k,l} \boldsymbol{\alpha}_{k,i} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha}_{k,i} \right\|_1 \right\}, \tag{15}$$

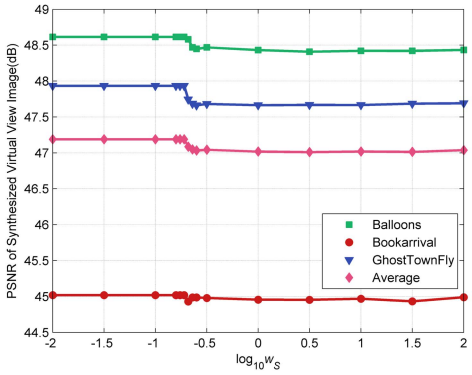
where w_k is the weighted factor indicating different impacts on VVI quality from the distortion of depth images of region k, $k \in \{E, S\}$. Similarly, apply (4) to (14), and we can obtain the reconstructed HR depth image at region k, $\mathbf{G}_k$, by

$$\begin{aligned} \mathbf{G}_k^* &= \arg\min_{\mathbf{G}_k} w_k \left\| \mathbf{H}_k \mathbf{G}_k - \mathbf{Z}_{m,k} \right\|_2^2 \\ &= \arg\min_{\mathbf{G}_k} \left\| \mathbf{H}_k \mathbf{G}_k - \mathbf{Z}_{m,k} \right\|_2^2. \end{aligned} \tag{16}$$

It can be inferred from (16) that the weight w_k has no influence on the process of back projection. However, w_k takes effect in (15) and ultimately affects the depth image SR performance at the reconstruction phase via sparse representation. Therefore, the VVI quality oriented depth image SR can be obtained by using (15) (13) and (16) successively.



(a)



(b)

Fig. 4. Relationship between the weights $\{w_E, w_S\}$ and VVI quality at the reconstruction phase. (a) Edge patches, (b) Smooth patches.

To determine the optimal weight for the reconstruction of depth images, we conducted analytical experiments between the weight w_k and VVI quality. The weight w_k was set as $\{w_k|0.01 \leq w_k \leq 100.00\}$, $k \in \{E, S\}$, for edge and smooth regions respectively. Fig. 4(a) illustrates the relationship between the weight w_E and VVI quality for edge patches, where the y-axis is PSNR of VVI and the x-axis is $\log w_E$. When $\log w_E$ is smaller than -0.75, the PSNR of the curves are consistent. It is mainly because there are no nonzero entries in sparse coefficients for LR edge patches, i.e. all zeros. When $\log w_E$ is around -0.60, all the curves reach their peaks. So, the optimal edge weight w_E, denoted as $w_{E,R}$, at the reconstruction phase is $10^{-0.6}$, i.e. 0.25.

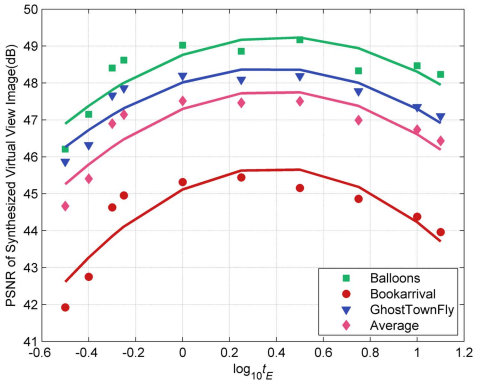
Besides, the relationship between w_S and VVI quality is also analyzed and shown in Fig. 4(b). It's demonstrated that there is no big variation for different w_S as it changes from 0.01 to 100.00. It is because that the smooth regions are simple and has few textures. when $\log w_S \in [-2.00, -0.70]$, Q_v of four curves achieves the highest PSNR and keeps unchanged as well, which is because there is no nonzero entries in sparse coefficients for LR smooth patches. In other words, the HR smooth patches are reconstructed by only using the mean patches of the LR smooth patches, i.e. $\mathbf{g}^0_{S,i}$ in (13). Thus, instead of using the sparse representation based reconstruction, we reconstruct smooth patches of the HR depth image by using the mean value of the LR patch. We denote the weight w_S as 'AvgLR' as well.

In fact, the reconstruction phase and the dictionary learning phase can be separately optimized in depth image SR. When the reconstruction phase only is optimized via WLDISR and dictionary learning phase is identically the same as that of ScSR, we denote this scheme as WLDISR-R.

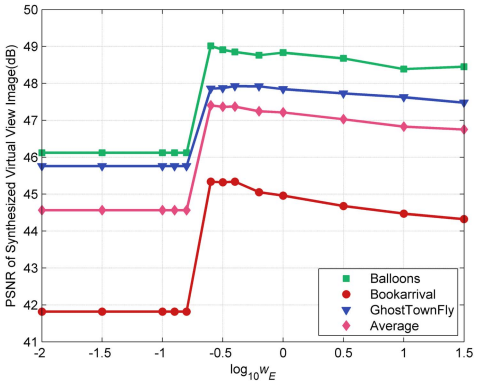
D. WLDISR Based Joint Optimization on Dictionary Learning and Reconstruction (WLDISR-ALL)

Since the dictionary learning and reconstruction phases can be jointly optimized within the proposed WLDISR, we proposed WLDISR-ALL scheme, which combines WLDISR-D and WLDISR-R. However, due to the high dependency between the dictionary learning module and reconstruction module, i.e., the optimal weights at the reconstruction phase are mutually correlated with the optimal weights at learning phase, optimal weights of the overall process shall be determined jointly. As the smooth region is simple and with much fewer textures, t_S at learning phase and w_S at the reconstruction phase are kept unchanged, i.e. both 'AvgLR', namely to construct HR smooth patches from average value of LR smooth image patches. In addition, the SR performance of smooth patches is independent from that of edge patches, so we only need to analyze the optimal parameter set for the edge patches for the WLDISR-ALL scheme. To analyze the weights dependency between the reconstruction and dictionary learning phases, we swept the weight t_E among $\{t_E|0.01 \leq t_E \leq 12.60\}$ while the weight w_E was set as the optimal value $w_{E,R}$ in WLDISR-R. After we obtained the new optimal weight for t_E, denoted as $t_{E,ALL}$, the weight of reconstruction w_E varied among $\{w_E|0.01 \leq w_E \leq 100.00\}$ while t_E was set as the optimal $t_{E,ALL}$.

Fig. 5(a) illustrates the relationship between the weights t_E and VVI quality for edge patches, where the x-axis is the $\log t_E$ and y-axis is the PSNR of synthesized VVIs. The dots are real collected data and the curves are fitting results by using the quadratic function in (8). The fitting R-square of the average curve is 0.79. We can observe that the relation is convex, which is similar to the fitting results in Fig. 3(a). In addition, we take its derivative to $\log t_E$ and set it as zero, and get the optimal weight t_E as 2.47 for edge dictionary learning. In addition, the relationship between w_E and VVI quality is also analyzed while t_E is set as 2.47, as shown in Fig. 5(b). It can be observed that the optimal weight w_E at the reconstruction phase in this joint optimization is $10^{-0.6}$ too, i.e. 0.25, which is identically the same as the optimal w_E in WLDISR-R. Based on the above experimental analyses, we can obtain that the ultimate optimal parameter set $\{t_E, w_E\}$ of edge patches in WLDISR-ALL scheme is $\{2.47, 0.25\}$. Meanwhile, the optimal parameters of smooth patches $\{t_S, w_S\}$ adopt the 'AvgLR' scheme. In summary, the complete optimal parameter set for WLDISR-ALL, $\{t_E, t_S, w_E, w_S\}$, is {2.47, AvgLR, 0.25, AvgLR}. In finding the optimal t_E and w_E, we firstly fix w_E and sweep t_E, then fix t_E and sweep w_E to find their optimal points. In fact, the same optimal t_E and w_E can be obtained as we exchange the searching order of w_E and t_E.



(a)



(b)

Fig. 5. Relationship between the weights $\{t_E, w_E\}$ and VVI quality in WLDISR-ALL. (a) t_E vs VVI quality in dictionary learning, (b) w_E vs VVI quality in reconstruction.

TABLE II
THE WEIGHTS IN SCSR AND WLDISR SCHEMES

Methods	t_E	t_S	w_E	w_S
ScSR [9]	1	1	1	1
WLDISR-D	3.40	AvgLR	1	1
WLDISR-R	1	1	0.25	AvgLR
WLDISR-ALL	2.47	AvgLR	0.25	AvgLR

Note that "AvgLR" denotes using the average value instead of dictionary in dictionary learning or reconstruction.

The optimal weight values of three schemes, i.e. WLDISR-D, WLDISR-R, and WLDISR-ALL, are shown in Table II. In edge dictionary learning, the optimal t_E of both WLDISR-D and WLDISR-ALL schemes are bigger than 1. As for WLDISR-D and WLDISR-ALL, it means edge patches of depth images in training dataset should be represented by more dictionary atoms subject to a fixed-size dictionary as compared with the original ScSR. Meanwhile, in SR reconstruction, the optimal w_E of WLDISR-R and WLDISR-ALL is smaller than 1. It implies that edge patches of depth image in the test dataset should be represented by fewer dictionary atoms subject to the fixed-size dictionary compared with ScSR in SR reconstruction. The reason may be that the depth image patches have fewer patterns and textures than color images. In addition, for smooth patches, the best way to reconstruct the HR smooth patches is 'AvgLR', in which just one pattern of dictionary atom, i.e. mean patches of smooth patches, is used in learning or reconstruction phases. Since the optimal t_S of WLDISR-D and WLDISR-ALL, 'AvgLR' $\in \{t_S | 0.01 \leq t_S \leq 0.32\}$, is smaller than the weight t_E, i.e. 3.40/2.47, and the optimal w_S of WLDISR-R and WLDISR-ALL, 'AvgLR' $\in \{w_S | 0.01 \leq w_S \leq 0.20\}$, is smaller than the weight w_E, i.e. 0.25, it indicates edge regions are more important than smooth regions, which is consistent with the finding that the distortion of the edge regions has larger impacts on VVI quality.

III. EXPERIMENTAL RESULTS AND ANALYSES

In this section, extensive experiments were performed to testify the performance of the proposed algorithm. In 3D and FVV system, two or more views of color and depth videos are encoded and transmitted. In these experiments, we mainly consider two views situation, where two views of color images are HR and the two views of depth images are LR, and one view of intermediate VVIs is rendered by View Synthesis Reference Software, VSRS 3.0 [37]. First, the LR depth images are restored to the HR depth images by our proposed algorithms and the benchmark schemes. Then, intermediate VVIs are rendered from the HR color view images and reconstructed HR depth images. We compare the SR performance among our proposed schemes, the bicubic interpolation method, the benchmark ScSR [9], and two other state-of-the-art methods [10], [21]. In addition to the PSNR of VVIs that rendered from the reconstructed depth images, their visual quality is also compared. Moreover, we compare the time complexities of our proposed methods with other schemes.

Four comparison SR methods, including bicubic interpolation method (denoted as Bicubic), the ScSR [9], 'Zeyde' scheme [10] (denoted as Zeyde), Edge-guided depth image SR scheme [21] (denoted as Edge-guided) were implemented and compared with our proposed WLDISR-D, WLDISR-R, and WLDISR-ALL schemes. The scaling factor was 2. The down-sampling and up-sampling method was Bicubic. For all the methods except Zeyde, they all used the same training set. The training images are 14 depth images in total, which have been described in Section II.B and Table I. For Zeyde, the default training database and default configurations [10] were used. For Edge-guided scheme, we used the default configurations in [21]. For ScSR, WLDISR-D, WLDISR-R, and WLDISR-ALL methods, the setting for the dictionary training is as follows: the number of the training patches was 1,000,000; the patch size was 5×5; the number of atoms of the training dictionary was 512; the variance threshold for patches was 10. We extracted 1,000,000 patches randomly from training images, since more patches in training had better and more stable performance in depth image SR as compared with default 100,000 patches setting in ScSR [9]. The way of random extraction is the same as that used in ScSR. To randomly extract the patches, there are two steps: compute the number of patches needed in every image proportional to the image size; randomly pick the patches from all the possible patches of each image.

TABLE III
VVI QUALITY COMPARISONS ON SHORT-TERM SEQUENCES (UNIT: dB)

3D Sequences	ScSR [9]	Zeyde [10]	Edgeguided [21]	WLDISR-D	WLDISR-R	WLDISR-ALL
Balloons*	49.92	48.70	47.49	49.99	50.40	50.54
Bookarrival*	45.79	43.67	43.22	45.99	46.20	46.12
Café	40.83	38.40	38.95	41.27	41.33	41.32
GhostTownFly*	47.68	46.41	45.07	47.87	48.07	48.27
Lovebird1	45.28	44.52	43.94	45.50	45.67	45.33
Kendo	54.32	51.75	50.36	54.76	54.70	54.86
Newspaper	45.29	43.44	42.05	45.52	45.55	45.49
Poznan_CarPark	36.28	36.65	36.64	36.25	36.18	36.17
PoznanHall2	51.05	49.44	48.47	51.31	51.26	51.70
Shark	43.46	41.77	41.99	43.70	43.78	43.90
Undodancer	49.05	46.12	44.27	49.69	50.10	50.99
Average gain	**46.27**	**44.63**	**43.86**	**46.53**	**46.66**	**46.79**
Average gain exclude *	**45.70**	**44.01**	**43.33**	**46.01**	**46.08**	**46.22**

TABLE IV
VVI QUALITY COMPARISONS ON LONG-TERM SEQUENCES (UNIT: dB)

3D Sequences	Bicubic	ScSR [9]	Zeyde [10]	Edgeguided [21]	WLDISR-D	WLDISR-R	WLDISR-ALL
Undodancer	40.59	44.83	42.60	41.73	46.25	46.64	47.82
Kendo	50.75	53.89	51.97	50.22	54.33	54.26	54.46
Café	38.65	40.88	38.79	39.04	41.27	41.35	41.38
Bookarrival	43.16	46.12	44.53	43.75	46.47	46.55	46.62
Average	**43.29**	**46.43**	**44.47**	**43.69**	**47.08**	**47.20**	**47.57**

A. Objective VVI Quality Comparisons

To evaluate the performance of our proposed three schemes, comparison experiments were performed among seven depth image SR schemes, which include the Bicubic, ScSR, Edge-guided, Zeyde, our proposed three schemes WLDISR-D, WLDISR-R, and WLDISR-ALL. All the sequences except Poznanstreet listed in Table I were employed as the test sequences, including Balloons, Bookarrival, Poznan_carpark, Café, GhostTownFly, Lovebird1, Newspaper, PoznanHall2, Kendo, PoznanStreet, and Undodancer. The resolution, views to be upsampled, the rendered view, and testing frames of each test sequences are illustrated in Table I. Specifically, short-term videos with 10 consecutive frames and long-term videos with 100 or 200 consecutive frames were tested.

Table III presents the average PSNR of the VVIs for the eleven short-term sequences. Bookarrival, GhostTown-Fly, and Balloons, which have been used as validation sequences in determining the weight parameters in Section II, are labeled with symbol '*'. We observe that the average quality of the VVIs rendered from the reconstructed depth image using ScSR, Edge-guided, and Zeyde methods are 46.27 dB, 44.63 dB, and 43.86 dB, respectively, for all the test sequences. The average PSNR of the VVIs from our three proposed schemes are 46.53 dB, 46.66 dB, and 46.79 dB, respectively. We can also observe that the SR performance ranks as WLDISR-ALL, WLDISR-R, WLDISR-D, ScSR, Zeyde, and Edge-guided in terms of VVI quality. The proposed WLDISR schemes are the top three. Moreover, for the test sequences, WLDISR-ALL excels method ScSR, Zeyde and Edge-guided by 0.52 dB, 2.16 dB, and 2.93 dB, respectively, which are significant improvements. While excluding the sequences labeled with symbol '*', similar improvements for our proposed schemes against ScSR, Zeyde, and Edge-guided can also be inferred. These improvements on short-term videos have proved that our schemes are more effective in improving the VVI quality as compared with the state-of-the-art depth image SR methods.

In addition, comparison experiments on long-term sequences were also performed to further testify the performance of the proposed algorithms. These four long-term sequences were selected since they are from four different providers, with various resolutions, camera settings, and video contents. Depth videos of Bookarrival and Kendo are from stereo matching, depth videos of Café are generated based on depth camera imaging, and Undodancer is animation video generated by computer graphics. Table IV presents the average PSNR of VVIs generated from the seven depth image SR methods, which were tested on the four long-term consecutive frames. It can be found that Bicubic, ScSR, Zeyde, and Edge-guided achieve 43.29 dB, 46.43 dB, 44.47 dB, and 43.69 dB respectively on average. Our proposed three WLDISR schemes can achieve 47.08 dB, 47.20 dB, and 47.57 dB on average. WLDISR-ALL excels method ScSR, Zeyde, and Edge-guided by 1.14 dB, 3.10 dB, and 3.88 dB on average over the four sequences. Our three schemes achieve the largest gains on average for sequence Undodancer, and the smallest gains on average for sequence Café. These substantial improvements on long-term consecutive frames further indicate that our proposed schemes are effective in improving VVI quality for 3D videos.

More specific frame-by-frame VVI quality comparisons on sequence Café and Undodancer are demonstrated in Fig. 6.

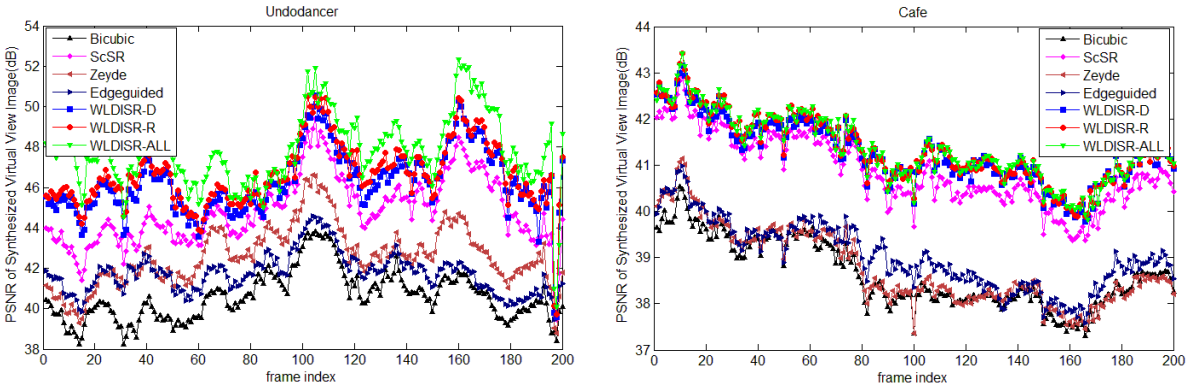


Fig. 6. VVI quality comparisons on sequences with 200 consecutive frames. (a) Undodancer; (b) Café.

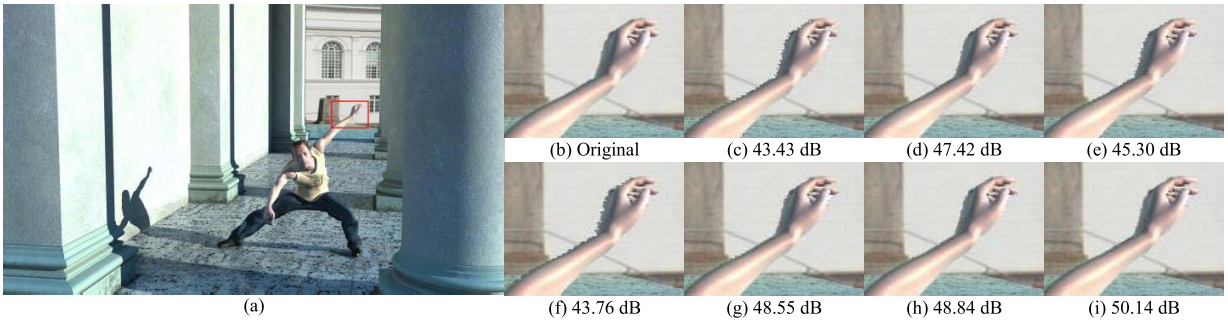


Fig. 7. Visual comparisons among VVIs from different SR schemes (Undodancer-111th frame). (a) Original virtual view; (b) enlarged image of the original virtual view; (c) to (i) enlarged images synthesized from the reconstructed HR depth images with Bicubic, ScSR, Zeyde, Edge-guided, WLDISR-D, WLDISR-R, WLDISR-ALL methods.

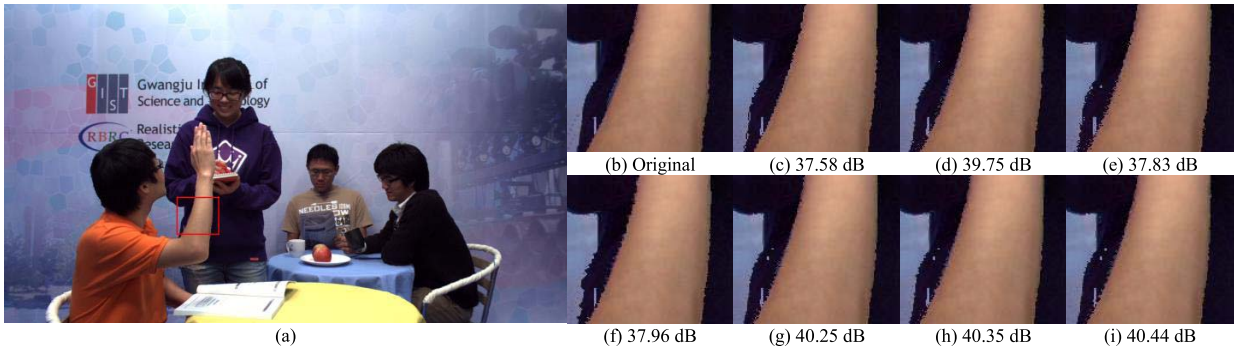


Fig. 8. Visual comparisons among VVIs from different SR schemes (Café-158th frame). (a) Original virtual view; (b) enlarged image of the original virtual view; (c) to (i) enlarged images synthesized from the reconstructed HR depth images with Bicubic, ScSR, Zeyde, Edge-guided, WLDISR-D, WLDISR-R, WLDISR-ALL methods.

The y-axis denotes the PSNR of the VVI and the x-axis denotes frame index of a sequence. It's observed that our proposed three schemes are generally better than ScSR, Edge-guided method, and Zeyde's scheme in VVI quality. The three anchor schemes are mainly designed to improve the visual quality of the depth image. Particularly, Edge-guided is specifically devised to improve the edges of depth image, such as more sharp and less jagged edges in human visual aspect. In comparison, our proposed schemes aim to improve VVI quality by considering the view synthesis impacts in objective functions. Therefore, the proposed three WLDISR schemes are of the top three VVI quality for the test sequences, while the three anchor schemes have inferior performance in terms of VVI quality. Moreover, the WLDISR-ALL is the best one among all the tested schemes for most of the frames.

B. Visual Quality Comparisons Among Rendered VVIs

In addition to comparison studies on the VVI quality among different depth image SR methods, subjective visual quality of the VVIs is also compared. Figs. 7 to 9 demonstrate the visual comparisons among different methods for Undodancer, Café, and Bookarrival sequences. The three sequences are of various types of textures. In addition, their

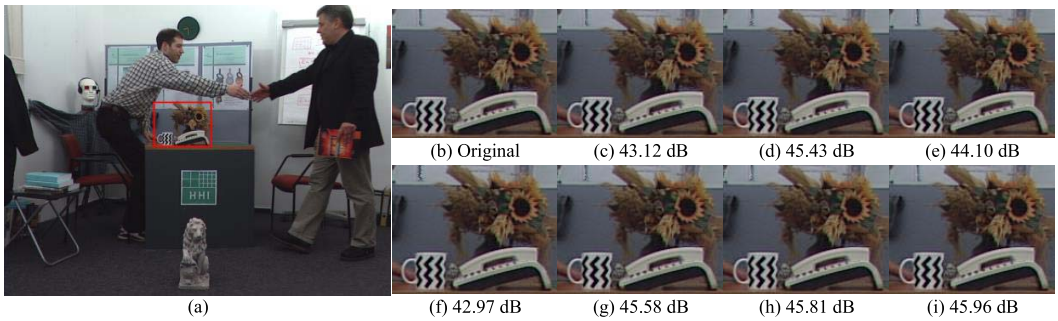


Fig. 9. Visual comparisons among VVIs from different SR schemes (Bookarrival-34th frame). (a) Original virtual view; (b) enlarged image of the original virtual view; (c) to (i) enlarged images synthesized from the reconstructed HR depth images with Bicubic, ScSR, Zeyde, Edge-guided, WLDISR-D, WLDISR-R, WLDISR-ALL methods.

depth images are generated from computer graphic, camera capturing, and stereo matching, respectively. For all figures from Figs. 7 to 9, (a) is VVI synthesized by color images and original depth images, where the red rectangle is zoomed for comparison; (b) is the zoomed region of the original VVI, (c) to (i) are enlarged images synthesized from the depth images up-sampled by Bicubic, ScSR, Zeyde, Edge-guided, WLDISR-D, WLDISR-R, and WLDISR-ALL schemes, respectively. Note that the PSNR values in subfigures from (c) to (i) are not the values of the enlarged images, but the quality values of entire VVIs. From Fig. 7, we can perceive annoying artifacts along the rim of hands, fingers and arms in the VVIs generated by Bicubic, Zeyde, and Edge-guided methods. By contrast, our proposed schemes have much clearer edges and fewer artifacts along the edges, and the visual quality of WLDISR-ALL scheme mostly resembles the reference VVI. Similar results can also be found for Café in Fig. 8 and Bookarrival in Fig. 9. The visual results further validate that our proposed schemes, considering different synthesis characteristics of different structures in depth images, are effective in improving the VVI quality and especially the quality of edge regions of the VVIs.

C. Comparisons on Computational Complexity

To evaluate the computation performance of our proposed methods, we implemented the proposed algorithms and benchmarks on Matlab R2014a. All the depth image SR experiments were run on a computer with an Intel I7 eight-core 4GHZ CPU, 32GB memory, and Windows 7 operating system. The testing methods except Bicubic, test sequences, and the frames to be up-sampled are the same as that in the long-term videos experiments in subsection III.A. Since the dictionaries are learned offline for the testing methods, we only need to compare the reconstruction time of the test sequences among different methods. The average computation time of super-resolving each frame is demonstrated in Fig. 10. It takes ScSR 358 seconds on average to up-sample a frame of a 3D sequence. ScSR runs slowest while Zeyde runs fastest. The computation time of the proposed schemes WLDISR-D, WLDISR-R, and WLDISR-ALL is 31.94%, 26.73%, and 27.57% respectively of the time ScSR costs. Compared with ScSR, the proposed schemes can reduce above 68.06% time complexities due to the time savings of smooth regions from

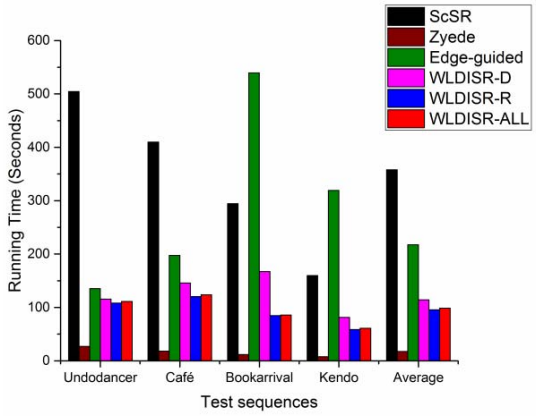


Fig. 10. Comparison of running time among different depth image SR methods.

using 'AvgLR'. The reason that the three WLDISR schemes are of different running time is mainly because the change of weight t_E and w_E leads to different learned dictionaries and number of non-zero entries in sparse coefficients, respectively. Overall, the proposed schemes have much lower computational complexity than the benchmark schemes.

IV. Key Factors Analyses on WLDISR and Discussions

In this section, three key factors' influences on the performance of the proposed schemes are analyzed. In addition, the role of the threshold for classifying edge and smooth regions is discussed.

A. Key Factors' Effects on the VVI Quality

In this subsection, we analyze three key factors in dictionary training, i.e. dictionary size, variance threshold for patches in training, and patch size, which have important impacts on the VVI quality. Since the optimal way to acquire smooth patches of HR depth image is the averaging operation, i.e. 'AvgLR', there is actually no dictionary for the smooth patches. Therefore, we mainly analyze the impacts of three key factors for edge patches. Four sequences, including Café, Undodancer, Bookarrival, and Kendo, were tested. Ten consecutive frames of each sequence with two views were tested and the average

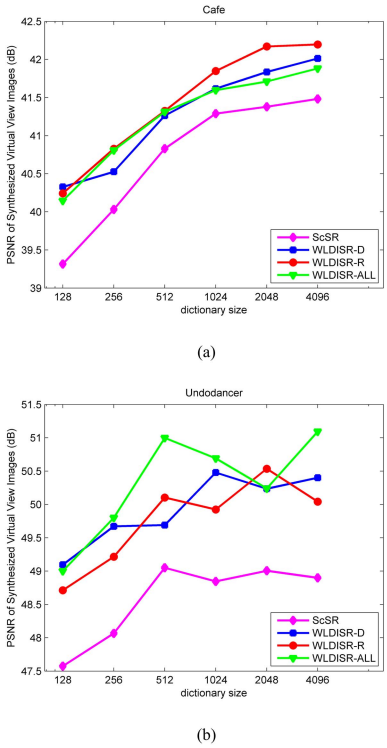


Fig. 11. Rendered VVI quality against dictionary sizes. (a) Café (b) Undodancer.

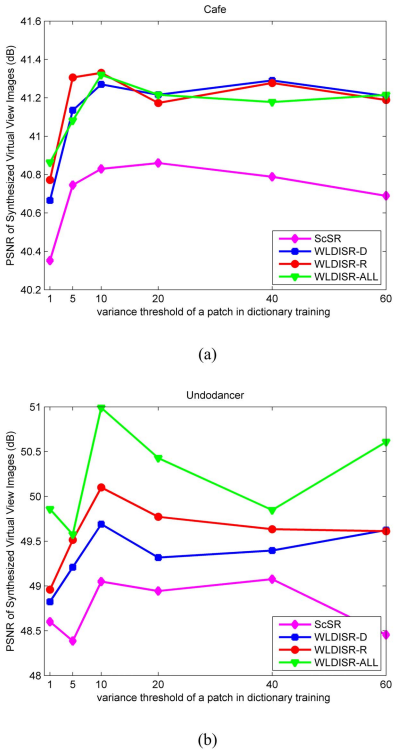


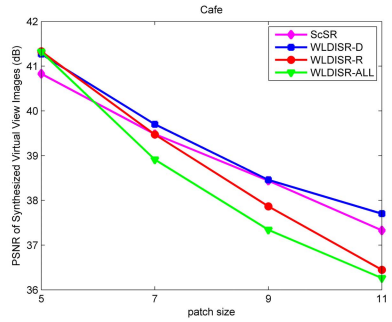
Fig. 12. Rendered VVI quality against patch variance thresholds. (a) Café (b) Undodancer.

PSNR value of their corresponding VVIs was calculated and used. Four methods, ScSR, WLDISR-D, WLDISR-R, and WLDISR-ALL, were tested and analyzed. The configurations of these methods were as same as those of subsection III.A. Due to the long length of the manuscript, only results of Café and Undodancer are illustrated.

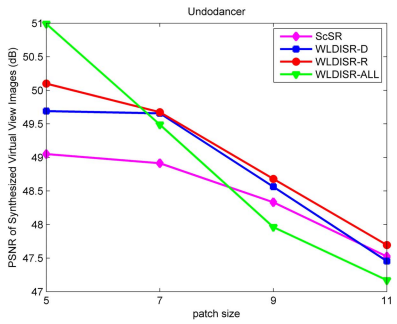
1) Effects of Dictionary Size: To analyze the relationship between the dictionary size and VVI quality with respect to the proposed algorithms, different dictionary sizes in the dictionary learning were set from 128 to 4096. Fig. 11 illustrates the relationship between the dictionary size and VVI quality for Café and Undodancer. The y-axis denotes the VVI quality measured with PSNR. The x-axis denotes the dictionary size and semi-log coordinate is employed in Fig. 11 for better observations. It can be observed that the average quality of VVIs increases with the dictionary size in general for different schemes for sequences Café and Undodancer. There are some fluctuations when the dictionary size increases for Undodancer. In addition, the WLDISR-ALL is not always the best one. It is mainly because the optimal weights for the dictionary learning and reconstruction are generated from the size with 512 and cannot guarantee the best for other sizes. Besides, the optimal weight is an average value of all sequences, which can improve the performance of mostsequences rather than all sequences. Basically, large dictionary size may improve the quality of the reconstructed depth images and VVIs; however, it will also lead to higher computational complexity. So, it's recommended to select a dictionary size from 512 to 1024 to make a trade-off between the SR performance and computing complexity.

2) Effects of Patch Variance Threshold: Patch variance threshold determines the number and diversity of patches in the dictionary learning. To analyze the relationship between the patch variance threshold and the VVI quality, we tested different patch variance thresholds, denoted as t_{pv}, with the range of $\{t_{pv} | 1 \leq t_{pv} \leq 60\}$. t_{pv} is used to remove the patches with variance below t_{pv} to control the structure information included in the training patches, which affects the number of the training patches. The larger t_{pv} is, the smaller number of patches is collected to train the dictionary. Fig. 12 illustrates the relationship between t_{pv} and the VVI quality with respect to different tested methods, where the x-axis is t_{pv} and y-axis is the VVI quality. It is observed that when t_{pv} is 1, which means more patches are included in training set, the VVI quality of the four schemes are almost the lowest for Café and Undodancer. This phenomenon suggests that the performance of the dictionary learning depends more on the distribution of patch structures than the number of patches in training. According to Fig. 12, it is suggested that t_{pv} set as 10 for WLDISR-D, WLDISR-R, and WLDISR-ALL can achieve a relatively higher performance against other settings.

3) Effects of the Patch Size: The relationship between the patch size in dictionary learning and the VVI quality was also analyzed. We tested different patch sizes, $n \times n$, where $n \in \{5, 7, 9, 11\}$. Fig. 13 illustrates the relationship between the patch size and the VVI quality for the four test schemes, where the x-axis is the patch size n and y-axis is the VVI quality. From Fig. 13, we observe that the quality of VVI decreases as the patch size increases in general for the four schemes on the tested sequences. The reasons are two-folds. One is the



(a)



(b)

Fig. 13. Rendered VVI quality against patch sizes. (a) Café (b) Undodancer.

representation fidelity will decrease as the patch size increases. The other is the overlap regions decrease as the patch size increases. In this paper, patch size is set as 5. In addition, we can observe that the WLDISR-ALL is not the best one for the patch size $n \in \{7, 9, 11\}$. It is because the optimal weighted factors $\{t_E, w_E\}$ in dictionary learning and reconstruction are determined when patch size is 5×5. These weights may not be the optimal when patch size changes.

B. Further Discussions

In the WLDISR, the depth images are divided into edge and smooth regions, and then processed individually in the dictionary learning and reconstruction. The threshold of the edge detection module would determine the division of edge and smooth regions. If the threshold is higher, more patches would be classified as smooth patches. It will consequently affect the dictionary learning and reconstruction phases in WLDISR. However, it is noteworthy that the optimal weights are determined with given edge and smooth regions classification. Once the smooth and texture regions are changed by using another edge detection threshold, the optimal weights will change accordingly. Overall, the major concept of the proposed WLDISR schemes is to do depth image SR differently for edge and smooth regions, and it has been proved to be effective. Moreover, three or more kinds of regions may also be considered under the proposed WLDISR framework.

V. Conclusions

We proposed a Weighted Local sparse representation based Depth Image Super-Resolution (WLDISR) scheme aiming at improving the virtual view image quality of 3D video system. Local sparse representation and weighted sparse representation are jointly applied in both dictionary learning and reconstruction phases for edge and smooth regions in depth image super-resolution. Three schemes WLDISR-D, WLDISR-R, and WLDISR-ALL are proposed and derived based on individual optimization on dictionary learning and reconstruction modules and the joint optimization on the two modules. Experimental results and visual comparisons validate that our proposed three schemes outperform other state-of-the-art methods. In addition, we have discussed about three key factors in affecting the performance of our proposed schemes. Overall, our work has achieved favorable quality improvement and can provide a new perspective for depth image super-resolution in 3D sequences. In future, we may further investigate depth image enhancement in temporal domain.

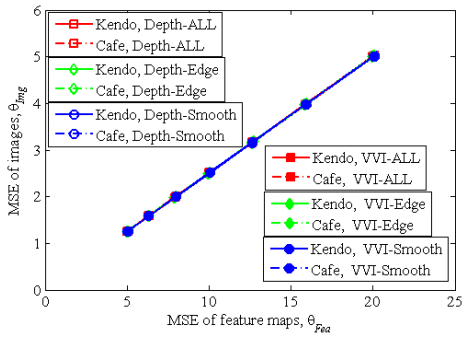


Fig. 14. Relationship between LR VVI/depth image feature map distortion and LR VVI/depth image distortion.

Appendix

In the following text, we denote the HR and LR depth image/feature patch distortion measured by MSE as $\delta_{\phi,i}/\Delta_{\phi,i}$, and HR and LR VVI distortion/feature patch distortion measured by MSE as $\sigma_{\phi,i}/\Lambda_{\phi,i}$, where $\phi \in \{h, l\}$, 'h' denotes HR, and 'l' denotes LR.

For HR depth image/VVI feature patches, HR depth image/VVI feature patches are acquired by subtracting the mean pixel value of HR depth image/VVI patches. $\mathbf{x}_i$ and $\mathbf{m}_i$ represents the original and reconstructed HR depth image feature patch, respectively, where $\mathbf{m}_i$ is reconstructed via sparse representation by HR dictionary $\mathbf{D}^h$ and sparse coefficient $\boldsymbol{\alpha}_i$. $\mathbf{g}_i$, $\mathbf{h}_i$ are the original and reconstructed HR depth image patches, respectively. $\mathbf{g}_i^0$ is the mean patch of HR depth image patch $\mathbf{g}_i$ with value of each pixel as g_i^0, and $\mathbf{g}_i^0$ can be approximately regarded as the mean patch of the reconstructed HR depth image patches. We can derive

$$\begin{aligned}\Delta_{h,i} &= \|\mathbf{x}_i - \mathbf{m}_i\|_2^2 = \left\|\mathbf{x}_i - \mathbf{D}^h\boldsymbol{\alpha}_i\right\|_2^2 \\ &\approx \left\|\mathbf{g}_i - \mathbf{g}_i^0 - (\mathbf{h}_i - \mathbf{g}_i^0)\right\|_2^2 \\ &= \|\mathbf{g}_i - \mathbf{h}_i\|_2^2 = \delta_{h,i}. \end{aligned} \tag{17}$$

Thus, the distortion of HR depth image feature patch $\Delta_{h,i}$ equals to the distortion of HR depth image patch $\delta_{h,i}$. $\mathbf{v}_i$, $\mathbf{n}_i$ are the corresponding original and reconstructed HR VVI patches

of depth image patches $\mathbf{g}_i$, $\mathbf{h}_i$, respectively. $\mathbf{v}_i^0$ is the mean patch of both the original and reconstructed HR VVI patches $\mathbf{v}_i$, $\mathbf{n}_i$. $\mathbf{p}_i$ and $\mathbf{q}_i$ represents the original and reconstructed HR VVI patch, respectively. Similarly, the HR VVI feature distortion patch can be derived as

$$\begin{aligned}\Lambda_{h,i} &= \|\mathbf{p}_i - \mathbf{q}_i\|_2^2 = \left\|\mathbf{v}_i - \mathbf{v}_i^0 - (\mathbf{n}_i - \mathbf{v}_i^0)\right\|_2^2 \\ &= \|\mathbf{v}_i - \mathbf{n}_i\|_2^2 = \sigma_{h,i}. \end{aligned} \tag{18}$$

The distortion of HR VVI feature patch $\Lambda_{h,i}$ also equals to the distortion of HR VVI patch $\sigma_{h,i}$.

Substitute (17) and (18) into (3), then we obtain

$$\Lambda_{h,i} = t_\varphi \Delta_{h,i} + \varepsilon_h. \tag{19}$$

It implies that (3) is applicable to the distortions of HR depth image/VVI feature patches for φ region, where $\varphi \in \{E, S, ALL\}$ represent edge, smooth regions and the entire image.

For LR depth image/VVI feature patches, experiments were conducted to establish the mathematical relationship between the LR depth image/VVI feature patch distortion $\Delta_{l,i}/\Lambda_{l,i}$ and the LR depth image/VVI distortion $\delta_{l,i}/\sigma_{l,i}$, respectively. The reconstructed noise of the depth image in SR and VVI can be modeled as white noise model. Series of white noise with intensity $[1, 2, 3, 4, 5, 6, 7]$ were injected into the LR depth images/VVI. Series of LR depth images/VVIs feature maps were then generated from these noise-contaminated depth images/VVIs.

Fig. 14 demonstrates the relationship between $\delta_{l,i}/\sigma_{l,i}$ and $\Delta_{l,i}/\Lambda_{l,i}$. The x-axis is θ_{Fea}, which denotes the MSE of LR feature map of depth/VVI images at φ region. The y-axis is θ_{Img}, which denotes the MSE of LR depth/VVI images at φ region. Two sequences, Kendo and Café, are demonstrated. The solid line and dotted line indicate the sequence Kendo and Café, respectively. The rectangle, diamond, and circle symbols denote the entire images, the edge, and smooth region, respectively. The solid and hollow markers denote the depth images and VVIs, respectively. From Fig. 14, it is observed that the MSE of images is strongly linear with the MSE of corresponding feature images both for depth images and VVIs. Moreover, the linear fitted lines from different sequences, regions, and types of images are overlapped. The linear model can be approximated as

$$\theta_{Img} \approx s\theta_{Fea}, \tag{20}$$

where s is the coefficient and is the same for different images, regardless of image types or contents, and for φ region, i.e. the entire image, edge region, and smooth region. The coefficient s seems only related to the high pass filter. For the high pass filter used in this paper, coefficient s approximates to 0.25. The constant term is omitted since it approaches to zero. The fitting R-square is 0.99. Thus, substitute (20) into (3) in the manuscript and we can obtain

$$s\Lambda_{l,i} = t_\varphi s\Delta_{l,i} + \varepsilon_l. \tag{21}$$

Thus, (21) can be transformed into

$$\Lambda_{l,i} = t_\varphi \Delta_{l,i} + \varepsilon_l/s. \tag{22}$$

Since ε_l approaches to zero, we can approximate ε_l/s to ε_l as they are both a small value approaching zero. Integrate (19) and (22), and we obtain

$$\Lambda_{\phi,i} = F\left(\Delta_{\phi,i}\right) = t_\varphi \Delta_{\phi,i} + \varepsilon_\varphi, \quad \phi \in \{h, l\}. \tag{23}$$

Therefore, (4) is proved and valid.

REFERENCES

[1] G. Tech, Y. Chen, K. Müller, J. R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3D extensions of high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 35–49, Jan. 2016.

[2] J. Y. Lee and H. W. Park, "Efficient synthesis-based depth map coding in AVC-compatible 3D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1107–1116, Jun. 2016.

[3] Z. Liu, G. Cheung, J. Chakareski, and Y. Ji, "Multiple description coding and recovery of free viewpoint video for wireless multi-path streaming," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 1, pp. 151–164, Feb. 2015.

[4] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, May 2017.

[5] S. Mandal, A. Bhavsar, and A. K. Sao, "Depth map restoration from undersampled data," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 119–134, Jan. 2017.

[6] M. Joachimiak, M. M. Hannuksela, and M. Gabbouj, "View synthesis quality mapping for depth-based super resolution on mixed resolution 3D video," in *Proc. 3DTV-Conf. True Vis.-Capture, Transmiss. Display 3D Video*, Jul. 2014, pp. 1–4.

[7] M. Joachimiak, M. M. Hannuksela, P. Aflaki, and M. Gabbouj, "Upsampled-view distortion optimization for mixed resolution 3D video coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1056–1060.

[8] E. Ekmekcioglu, M. Mrak, S. T. Worrall, and A. M. Kondoz, "Edge adaptive upsampling of depth map videos for enhanced free-viewpoint video quality," *Electron. Lett.*, vol. 45, no. 7, pp. 353–354, Mar. 2009.

[9] 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, Nov. 2010.

[10] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. 7th Int. Conf. Curves Surf.*, Jun. 2010, pp. 711–730.

[11] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[12] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. IEEE Asian Conf. Comput. Vis.*, Nov. 2014, pp. 111–126.

[13] Y. Zhang, Y. Zhang, J. Zhang, and Q. Dai, "CCR: Clustering and collaborative representation for fast single image super-resolution," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 405–417, Mar. 2016.

[14] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 848–852, Jun. 2017.

[15] L.-K. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1983–1996, Jun. 2015.

[16] S. Mandal, A. Bhavsar, and A. K. Sao, "Noise adaptive super-resolution from single image via non-local mean and sparse representation," *Signal Process.*, vol. 132, no. 5, pp. 134–149, Mar. 2017.

[17] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 71–84.

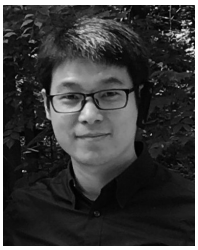
[18] Y. Zuo, Q. Wu, J. Zhang, and P. An, "Explicit modeling on depth-color inconsistency for color-guided depth up-sampling," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[19] S. Yang, J. Liu, Y. Fang, and Z. Guo, "Joint-feature guided depth map super-resolution with face priors," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 399–411, Jan. 2018.

[20] D. Ferstl, M. Rüther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 513–521.

[21] 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, Jan. 2016.
[22] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun, "Joint super resolution and denoising from a single depth image," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1525–1537, Sep. 2015.
[23] W. Hu, G. Cheung, X. Li, and O. Au, "Depth map super-resolution using synthesized view matching for depth-image-based rendering," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2012, pp. 605–610.
[24] H. Lv, Y. Zhang, K. Li, X. Wang, H. Xuan, and Q. Dai, "Synthesis-guided depth super resolution," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 125–128.
[25] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, "Depth map super-resolution considering view synthesis quality," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1732–1745, Apr. 2017.
[26] Y. Zhang, S. Kwong, L. Xu, S. Hu, G. Jiang, and C.-C. J. Kuo, "Regional bit allocation and rate distortion optimization for multiview depth video coding with view synthesis distortion model," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3497–3512, Sep. 2013.
[27] L. Fang, Y. Xiang, N.-M. Cheung, and F. Wu, "Estimation of virtual view synthesis distortion toward virtual view position," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1961–1976, May 2016.
[28] Y. Zhang, S. Kwong, S. Hu, and C.-C. J. Kuo, "Efficient multiview depth coding optimization based on allowable depth distortion in view synthesis," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4879–4892, Nov. 2014.
[29] H. Yuan, S. Kwong, X. Wang, Y. Zhang, and F. Li, "A virtual view PSNR estimation method for 3-D videos," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 134–140, Mar. 2016.
[30] C. H. Duan, C. K. Chiang, and S. H. Lai, "Face verification with local sparse representation," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 177–180, Feb. 2013.
[31] L. Wang, H. Wu, and C. Pan, "Manifold regularized local sparse representation for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 651–659, Apr. 2015.
[32] L. Liu, L. Chen, C. L. P. Chen, Y. Y. Tang, and C. M. Pun, "Weighted joint sparse representation for removing mixed noise in image," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 600–611, Mar. 2017.
[33] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5216–5223.
[34] A. Soltani-Farani and H. R. Rabiee, "When pixels team up: Spatially weighted sparse coding for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 107–111, Jan. 2015.
[35] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 801–808.
[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
[37] M. T. M. Tanimoto, T. Fujii, and K. Suzuki, *View Synthesis Algorithm in View Synthesis Reference Software 3.0 (VSRS3.0)*, document M16090, MPEG (ISO/IEC JTC1/SC29/WG11), Lausanne, Switzerland, Apr. 2009.

Huan Zhang received the B.S. degree from the Civil Aviation University of China, Tianjin, China, in 2010, and the M.S. degree from Tsinghua University, Beijing, China, in 2013. She is currently pursuing the Ph.D. degree with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China. Her research interests include image restoration and image or video quality assessment.



Yun Zhang (M'12–SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Visiting Scholar Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong. From 2010 to 2017, he was an Assistant Professor and an Associate Professor with the Shenzhen Institutes of Advanced Technology, CAS, where he has been a Full Professor since 2017. His research interests are video compression, 3D video processing, and visual perception.

Hanli Wang (M'08–SM'12) received the B.E. and M.E. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2007. From 2007 to 2008, he was a Research Fellow with the Department of Computer Science, City University of Hong Kong. From 2007 to 2008, he was also a Visiting Scholar with Stanford University, Palo Alto, CA, USA. From 2008 to 2009, he was a Research Engineer with Precoad Inc., Menlo Park, CA, USA. From 2009 to 2010, he was an Alexander von Humboldt Research Fellow with the University of Hagen, Hagen, Germany. Since 2010, he has been a Full Professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His current research interests include digital video coding, computer vision, and machine learning.

Yo-Sung Ho (SM'06–F'16) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, South Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990. In 1983, he joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. From 1990 to 1993, he was with the North America Philips Laboratories, Briarcliff Manor, NY, USA, where he was involved in the development of the advanced digital high-definition television system. In 1993, he rejoined ETRI as a Technical Staff and was involved in the development of the Korean DBS digital television and high-definition television systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), South Korea. Since 2003, he has been the Director of the Realistic Broadcasting Research Center, GIST, where he is currently a Professor with the School of Electrical Engineering and Computer Science. His research interests include digital image and video coding, image analysis and image restoration, 3D image modeling and representation, advanced source coding techniques, augmented reality and virtual reality, 3D television, and realistic broadcasting technologies. He served as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS VIDEO TECHNOLOGY.

Shengzhong Feng received the B.Sc. degree in computer science and engineering from the University of Science and Technology of China in 1991 and the Ph.D. degree in computer science and engineering from the Beijing Institute of Technology in 1997. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, where he is also an Assistant Director. His main research interests include big data, cloud computing, and networked computing systems with a specific emphasis on system architecture design and resource management for system's performance, reliability, availability, power efficiency, and security.