Region-Based Multiple Description Coding for Multiview Video Plus Depth Video

Chunyu Lin, Yao Zhao, *Senior Member, IEEE*, Jimin Xiao, and Tammam Tillo, *Senior Member, IEEE*

***Abstract*—Interframe and interview predictions are widely employed in multiview video coding. This technique improves the coding efficiency, but it also increases the vulnerability of the coded bitstream. Thus, one packet loss will affect many subsequent frames in the same view and probably in other referenced views. To address this problem, a region-based multiple description coding scheme is proposed for robust 3-D video communication in this paper, in which two descriptions are formed by setting the left and right view as dominant in the first and second description, respectively. This approach exploits the fact that most regions in the reference view could be synthesized from the base view. Hence, these regions could be skipped or only coarsely encoded. In our work, the disoccluded regions, illumination-affected regions, and remaining regions are first determined and extracted. By assigning different quantization parameters for these three different regions according to the network status, an efficient multiple description scheme is formed. Experimental results demonstrate that the proposed scheme achieves considerably better performance compared with the traditional approach.**

***Index Terms*—Multiple description coding, multiview video plus depth, video coding.**

I. Introduction

3D VIDEOS are able to provide depth perception through appropriate 3D display devices, which increases the immersive experience for the audience. Depending on whether glasses are required, 3D displays can be classified as stereoscopic or auto-stereoscopic. Stereoscopic displays require two texture/color views, and each view is projected to one of the eyes of the viewer through special glasses. Since wearing such glasses in a living room is uncomfortable and inconvenient, many studies focus instead on the auto-stereoscopic format. Auto-stereoscopic format provide different views depending on viewers' position and angle. Hence, a viewer can switch views by shifting his head position. However, to achieve this motion parallax feature of the auto-stereoscopic format, more views must be provided, which increases the burden of encoding and transmission.

Multiview video coding (MVC) standard was developed to efficiently compress multiple view data through inter-frame and inter-view predictions [1]. However, this approach only reduces the transmission burden partly because many views are still required. Multiview video plus depth (MVD) format was introduced as a new 3D video format [2] that includes texture images and their associated depth maps. By employing the depth image-based rendering (DIBR) technique, arbitrary virtual views can be generated; thus only a small number of views are required to be processed and transmitted [3]. Because of this advantage, the MVD format is being widely studied in industry and academia [4], [5], [6]. Among the MVD formats, a scheme based on two views plus two depth maps is the most popular because it requires relative little data and shows good synthesis performance. The use of two views plus two depth maps allows the disocclusion problem to be much more effectively mitigated compared with the use of just one view plus one depth map. Hence, this MVD format is also our focus in this paper. In this type of MVD format, one view is selected as the base/dominant view and is encoded using traditional intra/inter prediction, and the other view is designated as the enhancement/reference view and is encoded using intra/inter and inter-view predictions. Unless otherwise specified, the terms base view and dominant view will be used interchangeably throughout this paper, as will enhancement view and reference view.

In addition to the inter-frame prediction adopted in classical 2D video coding, the codec for MVD employs inter-view prediction and view synthesis prediction. Due to the complex prediction structure, the coding efficiency of the MVD format is improved; however, this prediction structure also increases the vulnerability of the coded bitstream to packet loss.

Multiple description coding (MDC) has been proposed as an efficient solution to combat packet loss. It provides a promising framework for video applications in which retransmission is unacceptable [7]. The classical MDC diagram is shown in Fig. 1, in which one source is encoded into two representations (descriptions) that are mutually refinable and can be decoded independently. The two descriptions are then transmitted over separate channels. When the network is experiencing no loss and all the descriptions are received, the best quality is obtained, with a

Manuscript received October 11, 2016; revised May 22, 2017 and August 2, 2017; accepted September 29, 2017. This work was supported in parts by the National Natural Science Foundation of China (No.61772066, No.61210006 and 61501379) and by the Beijing Natural Science Foundation (No. KZ201610005007). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaoqing Zhu. *(Corresponding author: Chunyu Lin)*

C. Lin and Y. Zhao are with the Beijing Key Laboratory of Advanced Information Science and Network, Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: cylin@bjtu.edu.cn; yzhao@bjtu.edu.cn).

J. Xiao is with the Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: jimin.xiao@xjtlu.edu.cn).

T. Tillo is with the Libera Universit di Bolzano-Bozen (unibz), Bolzano 39100, Italy, and also with Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: Tammam.Tillo@unibz.it).

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/TMM.2017.2766043

1520-9210 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

IEEE Proof

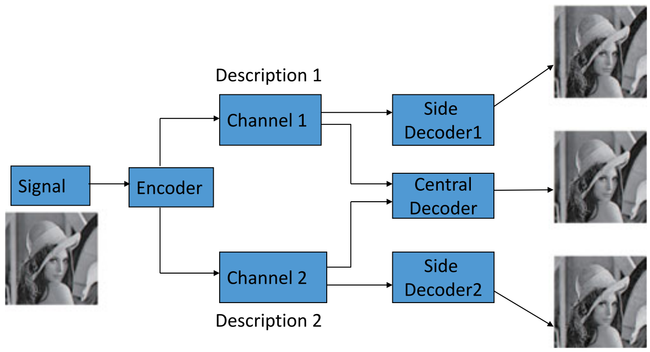

Fig. 1. Classical multiple description diagram.

so-called central distortion. If only one channel is working, the side decoder can reconstruct the source with a certain desired side distortion. To achieve this resiliency, some redundancy should be introduced in the descriptions, which is useful for mitigating packet loss but is detrimental to the central performance when no packets are lost. This redundancy should be tuned according to the network status. For example, the side distortion should be minimized at a high packet loss rate, whereas the central distortion should be minimized at a low packet loss rate. Thus, flexible tuning of the redundancy is a key task for any MDC scheme.

Many MDC works have been proposed for robust 2D video coding [8], [9]. Some studies have also been conducted on the stereoscopic video format [10]. In [11], the spatial scaling MDC scheme (SS-MDC) and the multi-state MDC scheme (MS-MDC) were proposed for stereoscopic videos. In SS-MDC, an asymmetric stereo pair is used to form descriptions, such that one view is at full resolution and the other view is down-sampled. In MS-MDC, temporal down-sampling is applied. For example, the odd frames of both the left and right views are grouped to form one description, whereas the other description contains the information for the even frames. In [12], multiview videos are subsampled in both the horizontal and vertical directions to form four sub-sequences. Then these four sub-sequences are paired to form two descriptions. In each description, one sub-sequence is directly encoded, whereas the other uses mode duplication based on the mode of the sub-sequence in the other description. This scheme is simple and efficient; however, its redundancy allocation is not flexible. In [13], an MDC video coding scheme for stereoscopic video was proposed based on a stagger frame order. All these schemes are very efficient; however, little research has yet been performed on MVD format. In fact, as more predictions are introduced, a bitstream of the MVD format becomes more vulnerable and requires greater protection. Otherwise, one packet loss in one frame will seriously affect the current view and the other reference views, as well as the virtual synthesized view. In addition, most MDC schemes for 3D videos are merely simple extensions of their 2D versions, such as spatial subsampling or temporal subsampling [11], [14]. Thus, features of MVD are not sufficiently utilized.

In this paper, we propose a region-based multiple description coding scheme (RB-MDC) that attempts to optimize the expected performance considering region importance and channel status. The proposed scheme first differentiates each region in the texture and depth videos with respect to its importance. Based on the differentiated regions, unequal protection is provided according to the importance of each region and the network status. Compared with classical schemes, gains of up to 2 dB can be achieved on both the texture videos and the synthesized views in the case of high packet loss rates.

The remainder of this paper is organized as follows. In Section II, an outline of the proposed scheme is provided, with introductions to region classification in Section II-A and redundancy allocation in Section II-B. Experimental results are presented and analyzed in Section III. Finally, conclusions are drawn in Section IV.

II. Proposed Scheme

The proposed multiple description scheme is illustrated in Fig. 2. Since the two descriptions are formed in the same way, we will take description 1 as an example to describe our algorithm. For description 1, as shown in Fig. 2, the left view is chosen to be the dominant view, whereas the right view is designated as the enhancement view. First, a virtual right view is synthesized from the left view plus depth. Based on the virtual right view, the original right view can be classified into disoccluded regions, illumination-affected regions and the remaining regions. These three types of regions have different effects on the quality of the synthesized views, as will be explained further in the next subsection. Based on this classification, lower bit rates can be assigned to unimportant regions that constitute higher percentages of the overall images. Therefore, redundancy can be flexibly allocated, and the total bit rate can be reduced. For description 2, the right view is the dominant view; otherwise, the process is similar to that for description 1.

If only one description is received, normal quality of the dominant view can be achieved along with a relatively lower quality for the enhancement view. Since the disoccluded regions and illumination-affected regions, which have a higher impact on the virtual view, have been better encoded, we can still obtain well-synthesized virtual views. When both descriptions are received, good central performance can be achieved with both the dominant left view and the dominant right view. Because the two dominant views are employed, better synthesized quality is expected.

A. Region Classification

In Fig. 2, one important step of the scheme is to classify different regions based on their contributions to the virtual left/right views. In the proposed scheme, three types of regions are classified: disoccluded regions, illumination-affected regions and the remaining regions. For the example of description 1, the disoccluded regions, or the regions that appear as a result of view switching, are the pixels in the right view that cannot be rendered from the left view. Regions of this type are the most important because the synthesized views require them but they exist only in the original right view. Notice, the holes due to large baseline are also regarded as disoccluded regions since they cannot be rendered from the base view. The illumination-affected regions

IEEE Proof

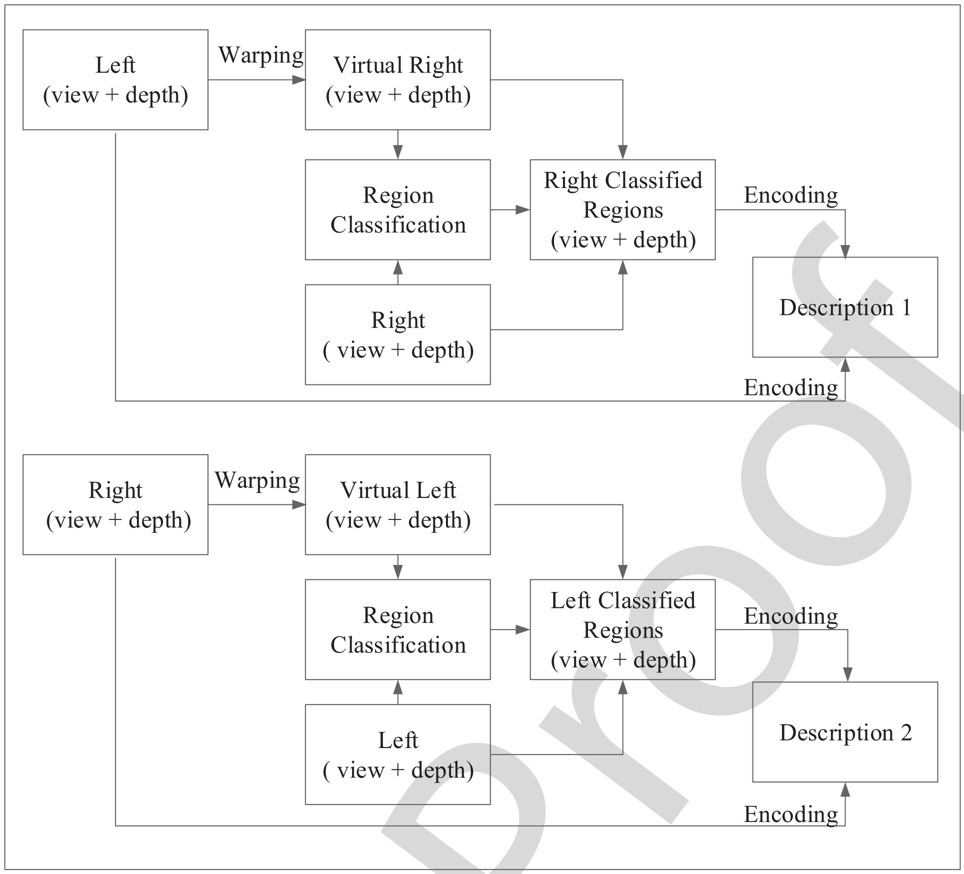

Fig. 2. Region-base multiple description scheme.

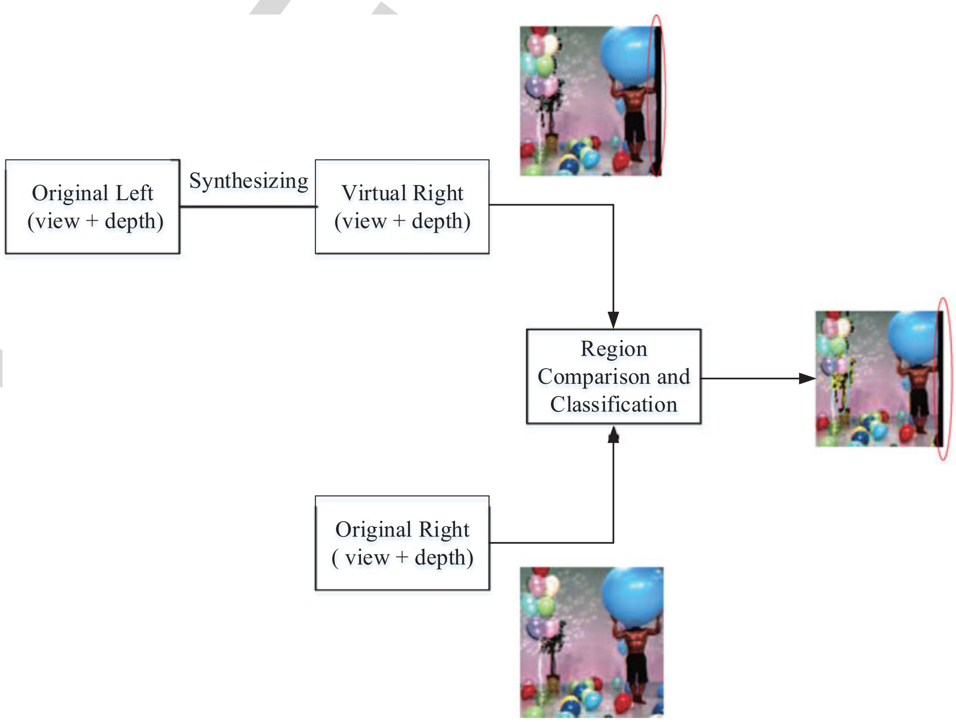

Fig. 3. Region classification process, where regions are highlighted.

Proof

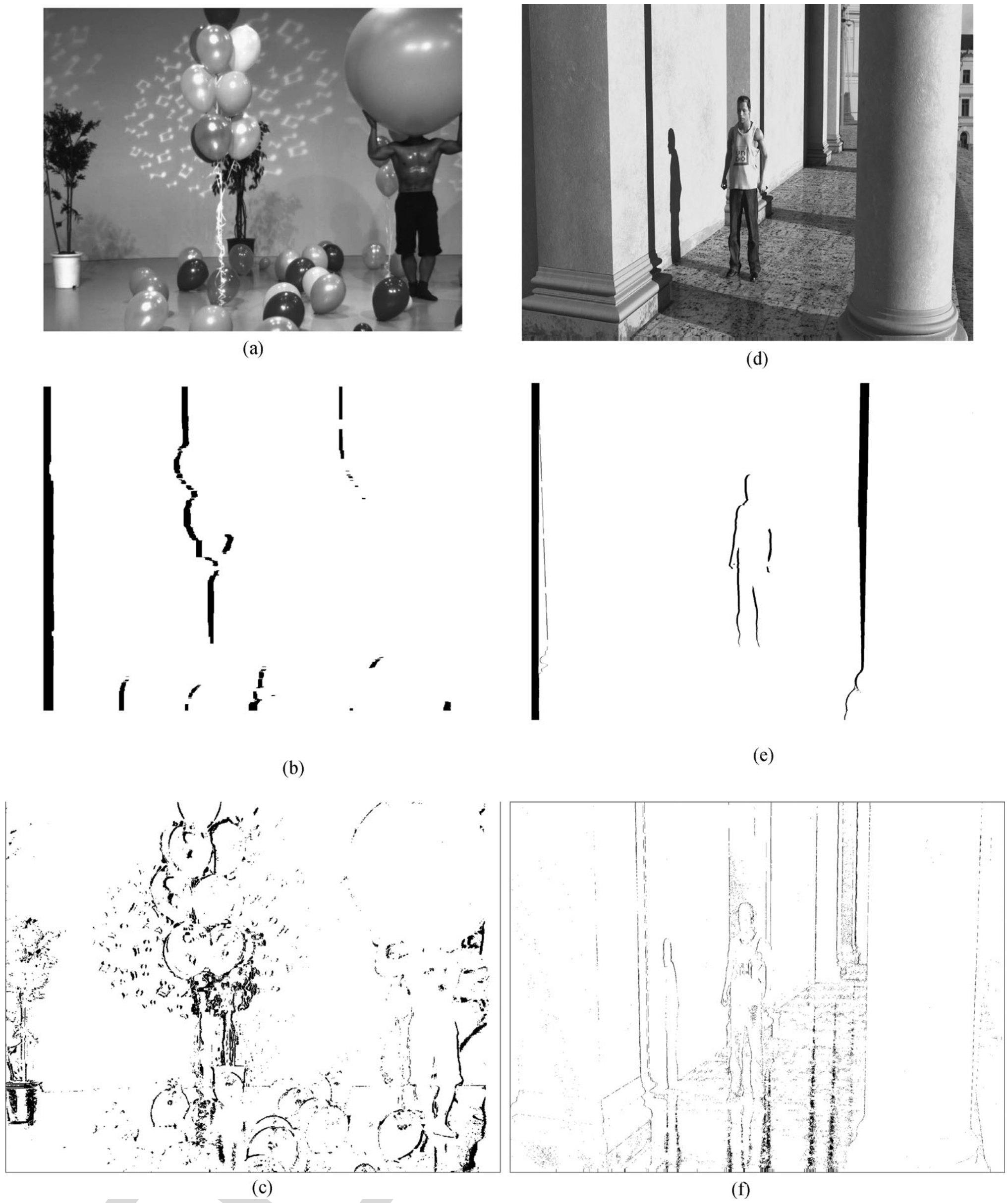

Fig. 4. Region classification example with *Balloons* and *Undodancer*. (a) Original. (b) Disoccluded. (c) Illumination-affected. (d) Original. (e) Disoccluded. (f) Illumination-affected.

are the regions in the right view that can be rendered from the left view but only with low quality. Because of the differences in the illumination conditions between the left and right views, some regions in the rendered virtual right view will differ from those in the original right view, and these regions should be encoded with sufficiently good quality to correct for these differences. Regions of the last type, called the remaining regions, can be rendered from the left view with a sufficient level of quality.

To classify such regions, a synthesis process is required to render the virtual right view from the left view, as shown in Fig. 3. In this synthesis process, only one texture video and one depth map can be employed; hence, many holes will be generated because of a lack of pixel information at the corresponding locations. These holes are represented as black regions in the figure. Note that except in the classification step, the synthesis process in our scheme can generally employ two texture videos

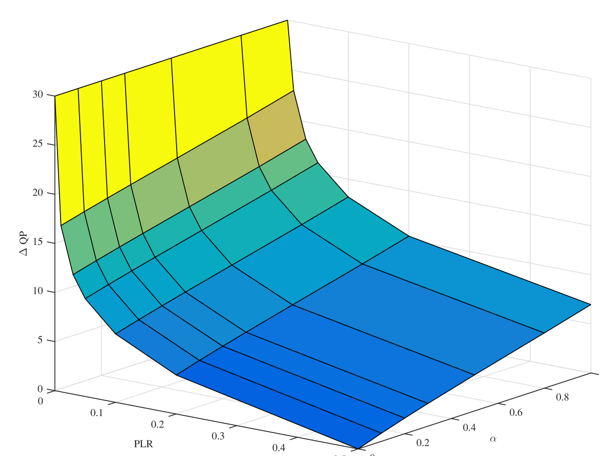

Fig. 5. ΔQP as a function of packet loss rate (PLR) and α.

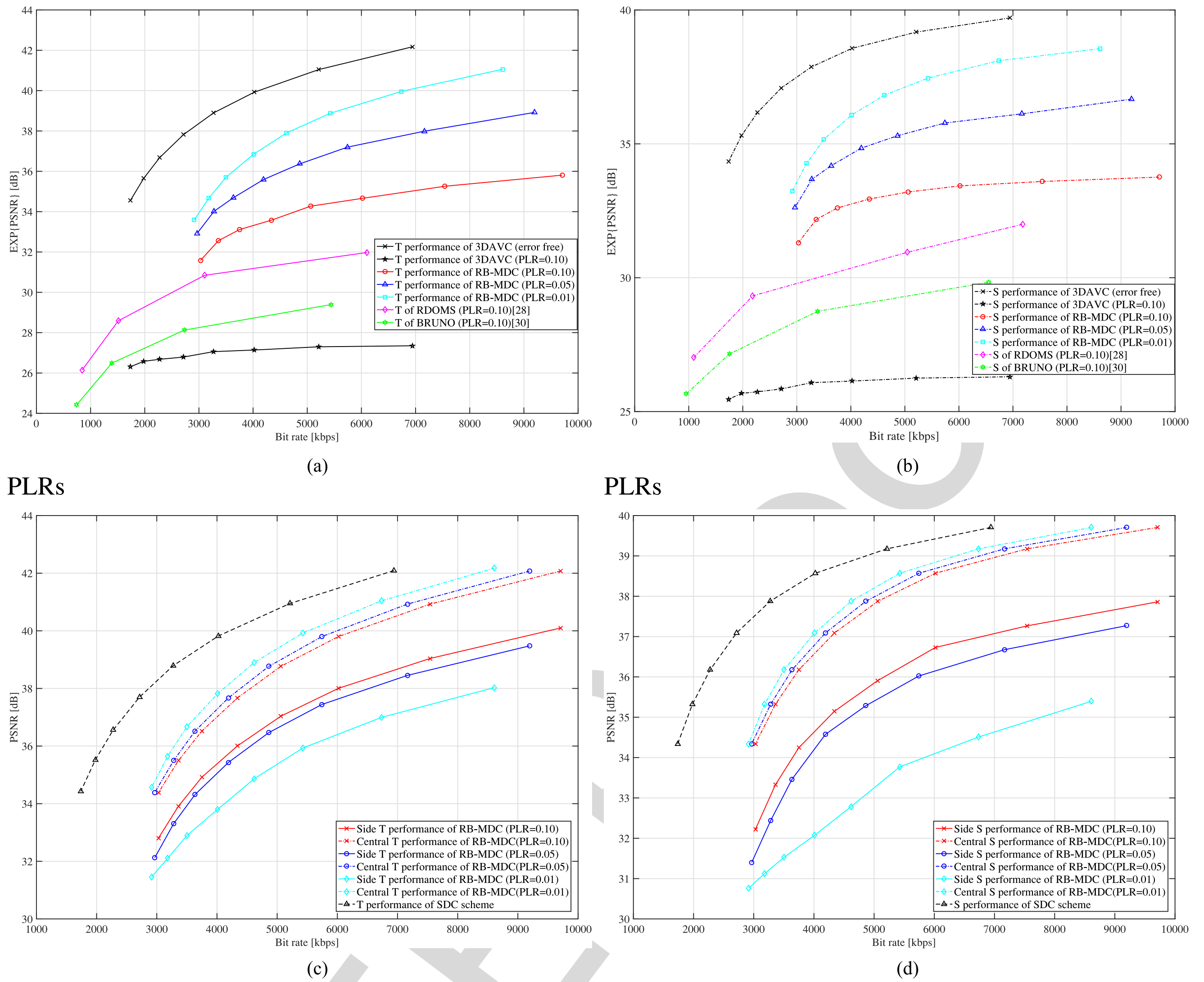

Fig. 6. The rate-PSNR performance of *Newspaper*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

plus two depth maps, which will enable the generation of considerably better virtual views at the cost of increased computation. We can simplify this synthesis process as described in [15]. Compared with the original right view, disoccluded regions can be located easily since these regions are in fact holes.

To determine the illumination-affected regions, the difference between the synthesized virtual view and the original view is first calculated. Regions with value differences larger than a certain threshold are identified as illumination-affected regions. The threshold for illumination-affected region will highly depend on the video contents and it is still an open topic yet. In our case, we set the threshold by a just noticeable difference(JND) [16]. JND is the least perceptible difference that human can notice. In [16], the JND calculation considered both the contrast and pattern complexity, which achieves very good performance. With a given sequence, its JND value is first calculated frame by frame, if a pixel difference between the warped view and the original view is larger than its corresponding JND value, it will be labeled as illumination-affected pixel. After the classification of these two types of regions, the remainder are regarded as remaining regions that can be warped from the dominant view with sufficiently good quality. Hence, the classification process is quite simple.

Examples of region classification are presented in Fig. 4, where the sequences *Balloons* [17] and *UndoDancer* [18] are divided into regions of the three different types. The second and third rows present the disoccluded and illumination-affected regions, respectively, whereas the others show the remaining regions. Here, illumination-affected regions are pixels in which the value difference between the original view and the virtual view is greater than its corresponding JND value. It can be observed that disoccluded regions and illumination-affected regions account for only a small percentage of the entire image. Hence, the allocation of a lower bit rate to the remaining regions, which constitute a large percentage, could considerably reduce the total bit rate. Note that the classification applies to both the color videos and the depth videos. For simplicity, for the depth maps, we just use the classified maps determined for the color videos.

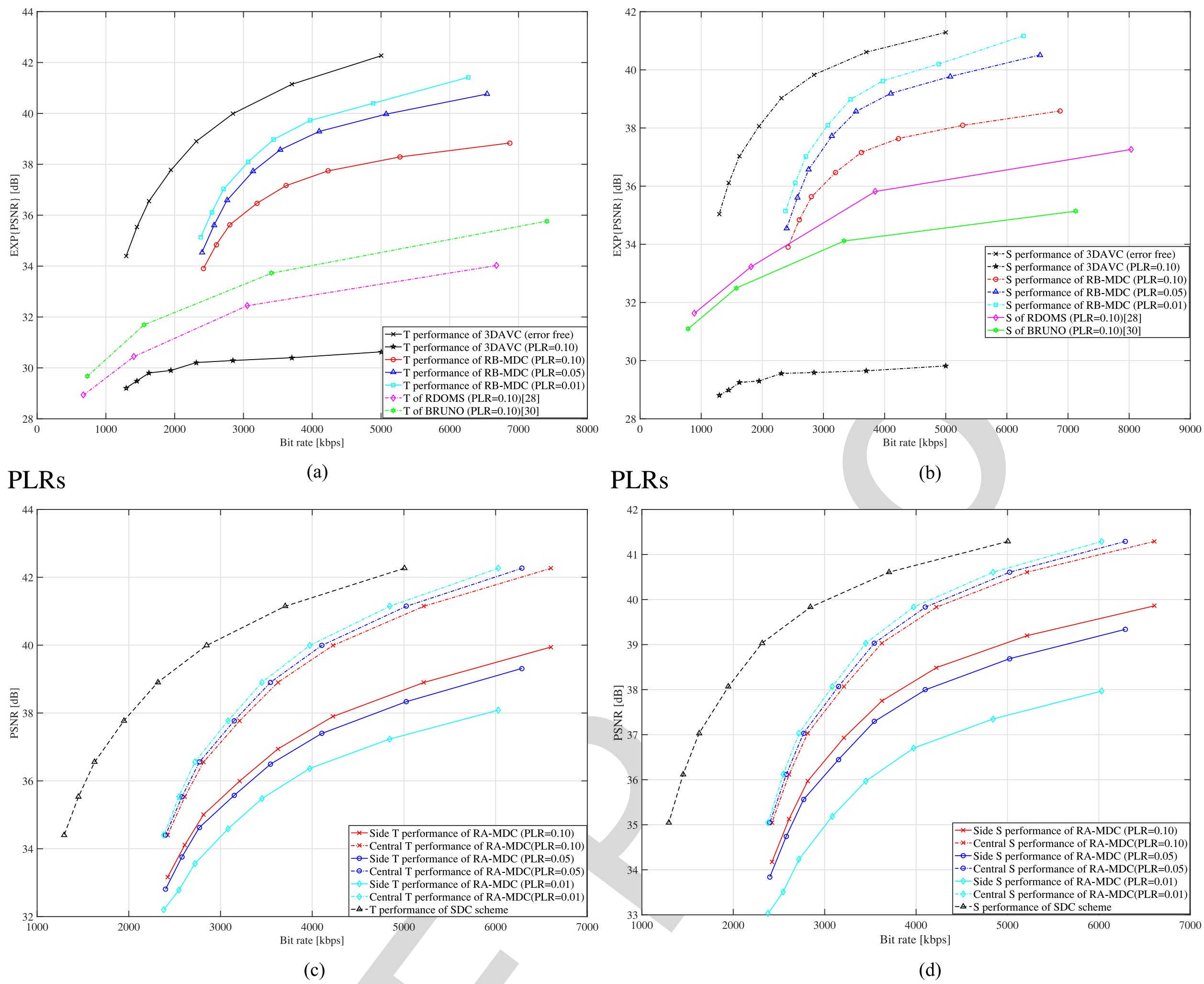

Fig. 7. The rate-PSNR performance of *Lovebird*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

B. Redundancy Allocation

Based on the classified regions, we need to design an effective redundancy allocation algorithm that considers region importance and network status to minimize the expected distortion, including both the real and virtual views, constrained by a fixed total rate. As shown in Fig. 2, additional data representing the two views are required in comparison with the single description scheme (SDC), in which only one pair of views is encoded. The bitstream of the additional views provides redundancy. When the channel quality is not good and the packet loss rate is high, more bits should be assigned to the additional views. By contrast, fewer bits are required when the channel quality is good. Hence, redundancy allocation is a key problem in any MDC scheme. In practice, the disoccluded regions and illumination-affected regions should receive higher protection compared with the remaining regions.

Since these three types of regions have different contributions to the overall performance, different levels of protection or redundancy should be allocated accordingly. Our final goal is to design a rate allocation strategy that considers the relationship among the different types of regions.

First, we need to estimate the expected distortion (left view, right view and virtual views) at the encoder end, considering the network status and the classified regions, under the relevant constraint on the total bit rate. During this process, the distortions of synthesized virtual views must be approximated. Then, we can obtain the rate-distortion function for each region and construct the relationship among the regions accordingly. Finally, we can perform bit-rate allocation based on the different quantization parameter (QP) values calculated from the rate-distortion functions. We will introduce the entire process in detail in the following.

1) Expected Distortion: The expected total distortion should include the distortions of the left and right views as well as of synthesized virtual views. It can be evaluated as

$$\begin{aligned}\bar{D} = (1-p)^2(D_L + D_R + D_V) + p(1-p)(D'_R + D_L \\ + D_{LV}) + p(1-p)(D'_L + D_R + D_{RV}) \\ + p^2(D''_L + D''_R + D''_V) \quad (1)\end{aligned}$$

where $\bar{D}$ denotes the total expected distortion and p is the packet loss rate. The subscripts L and R denote the left and right views,

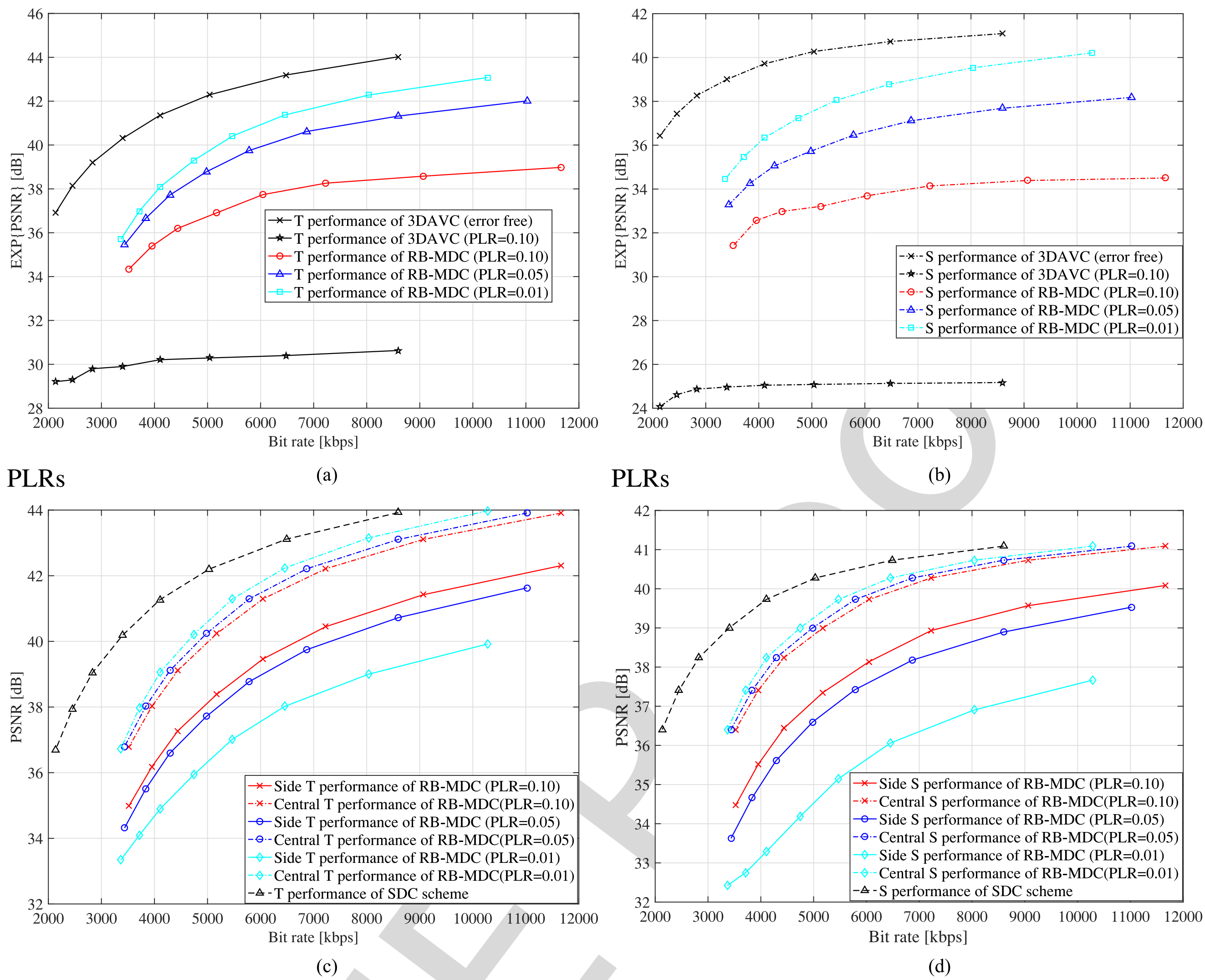

Fig. 8. The rate-PSNR performance of *Balloons*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

respectively, whereas the subscript V represents a synthesized virtual view. D_L and D_R represent the distortions of the left view and right view, respectively, when the dominant mode is used, whereas D'_L and D'_R are the corresponding distortions for the enhancement mode. D_V is the distortion of the view synthesized using the dominant left and right views, whereas D_{LV} and D_{RV} are the distortions of the views synthesized using only the dominant left view or the dominant right view, respectively. Finally, D''_L, D''_R and D''_V are the corresponding distortions with error concealment when the same frames are lost in both the left and right views. The distortions of the left and right views, such as D_L, D_R, D'_L, D'_R, D''_L and D''_R, can be calculated during encoding, whereas those of synthesized views must be estimated and approximated.

The quality of a synthesized view depends on the qualities of the left view and right views as well as on the rendering mode. If the qualities of the left view and the right view are similar, then an averaging mode in which both views are equally important is preferred. Otherwise, an extrapolating mode that uses one dominant view with a higher weight is adopted. Hence, the virtual distortion can be represented as follows:

$$\begin{cases} D_V &= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha)S(\hat{X}_R)\big) - X_V\Big)^2 \\ D_{LV} &= E\Big(\big(\alpha_L S(\hat{X}_L) + (1-\alpha_L)S(\hat{X'}_R)\big) - X_V\Big)^2 \\ D_{RV} &= E\Big(\big(\alpha_R S(\hat{X}_R) + (1-\alpha_R)S(\hat{X'}_L)\big) - X_V\Big)^2 \end{cases} \quad (2)$$

where $S()$ is the synthesis function that renders the left and right views $\hat{X}_L$ and $\hat{X}_R$ into the virtual view; X_V is the original virtual view synthesized from the original left view X_L and the original right view X_R; and α, α_L and α_R are the rendering mode parameters. For example, α can be set to 0.5 when the left view and right view are of similar quality. In practice, the rendering process is also affected by different types of regions. Suppose that one view is designated as the dominant view; then, most regions in the virtual view will be rendered from this dominant view, whereas the disoccluded regions must be rendered from the other view.

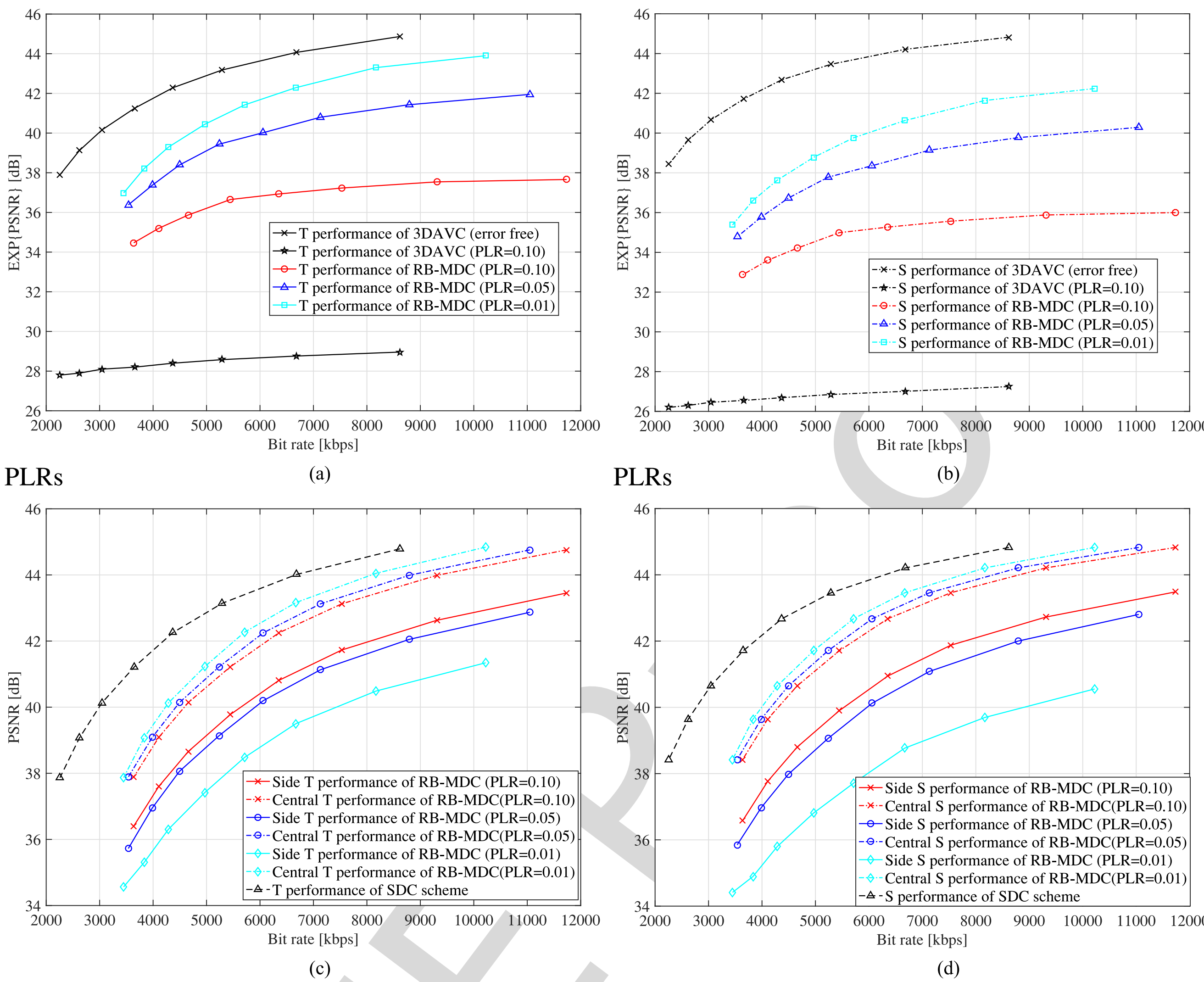

Fig. 9. The rate-PSNR performance of *Kendo*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

Since the three distortions D_V, D_{LV} and D_{RV} can be calculated in a similar manner, we will simply discuss D_V as an example.

$$
\begin{aligned}
D_V &= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha)S(\hat{X}_R)\big) - X_V\Big)^2 \\
&= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha)S(\hat{X}_R)\big) \\
&\quad - \big(\alpha S(X_L) + (1-\alpha)S(X_R)\big)\Big)^2 \\
&= E\Big(\alpha\big(S(\hat{X}_L - X_L)\big) + (1-\alpha)\big(S(\hat{X}_R - X_R)\big)\Big)^2 \\
&\approx \alpha^2 D_L + (1-\alpha)^2 D_R \\
&\quad + 2\alpha(1-\alpha)E\big(S(\hat{X}_L - X_V)S(\hat{X}_R - X_V)\big) \\
&= \alpha^2 D_L + (1-\alpha)^2 D_R
\end{aligned} \tag{3}
$$

The virtual distortion primarily depends on the views to be rendered; hence, we approximate the distortions $(S(\hat{X}_L) - X_V)^2$ and $(S(\hat{X}_R) - X_V)^2$ as $(\hat{X}_L - X_L)^2 = D_L$ and $(\hat{X}_R - X_R)^2 = D_R$, respectively. In addition, $E\big(S(\hat{X}_L - X_V)S(\hat{X}_R - X_V)\big)$ is assumed to be zero since these two errors are uncorrelated [1].

In the same way, we can obtain the other two virtual distortion formulas

$$
\begin{cases}
D_{LV} = \alpha_L^2 D_L + (1-\alpha_L)^2 D'_R \\
D_{RV} = \alpha_R^2 D_R + (1-\alpha_R)^2 D'_L
\end{cases} \tag{4}
$$

2) Rate-Distortion Functions: In our scheme depicted in Fig. 2, the left view is encoded as the base view and the right view is encoded as the enhancement view in description 1, and vice versa for description 2. We set the quality of the base view as an anchor, and our key objective is to determine the quality of the enhancement view depending on its region classification and the network status. Let the bit rates of the base views be R_L and R_R, whereas the bit rates of the enhancement views are R'_L and R'_R. The problem can be expressed as follows:

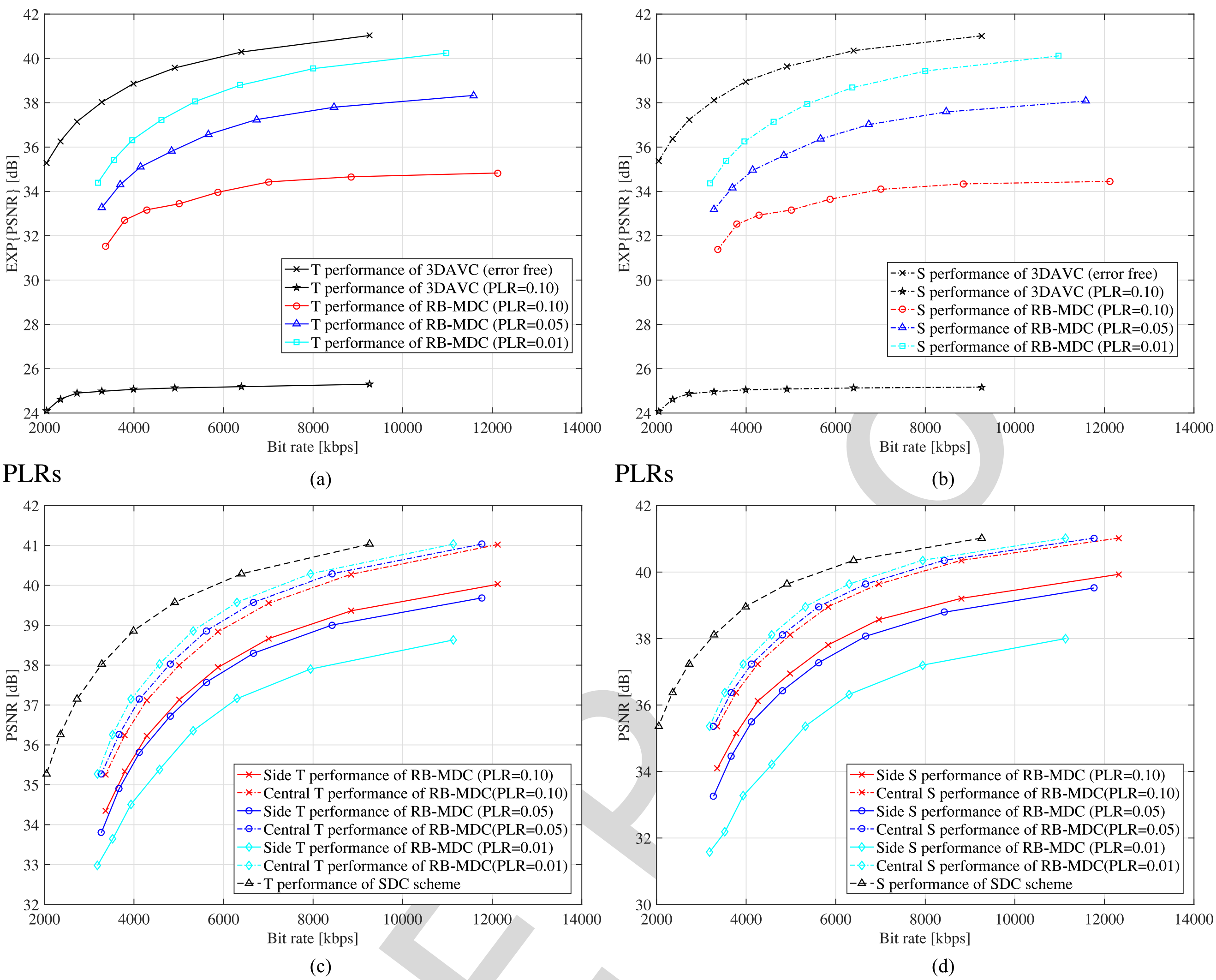

Fig. 10. The rate-PSNR performance of *Bookarrival*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

$$\begin{cases} \min \quad \bar{D} = \sum_{i=1}^{N} \bar{d}[i] \\ \text{s.t.} \ \ R_t = \sum_{i=1}^{N} (R_L[i] + R_R[i] + R'_L[i] + R'_R[i]) \end{cases} \tag{5}$$

where $\bar{d}[i]$ denotes the expected distortion of the ith macroblock(MB) among N total MBs and R_t represents the limit on the total bit rate imposed by the available bandwidth. This problem can be solved using the standard Lagrangian approach as follows

$$L = \bar{D} + \lambda \sum_{i=1}^{N} (R_L[i] + R_R[i] + R'_L[i] + R'_R[i]) \tag{6}$$

where λ is the Lagrangian multiplier. Because the two descriptions are symmetric, we will take description 1 as an example. In description 1, the left view and right view are treated as the base view and enhancement view, respectively, whose bit rates are R_L and R'_R, respectively. Using formula (1) and imposing $\nabla L = 0$, we obtain

$$\frac{\partial L}{\partial R_{L,i}} = (1-p)^2 \left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D_{V,i}}{\partial R_{L,i}} \right) + p(1-p) \left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D'_{R,i}}{\partial R_{L,i}} + \frac{\partial D_{LV,i}}{\partial R_{L,i}} \right) + \lambda = 0 \tag{7}$$

$$\frac{\partial L}{\partial R'_{R,i}} = p(1-p) \left(\frac{\partial D'_{R,i}}{\partial R'_{R,i}} + \frac{\partial D_{LV,i}}{\partial R'_{R,i}} \right) + \lambda = 0 \tag{8}$$

Here, $\frac{\partial D'_{R,i}}{\partial R_{L,i}}$ denotes the relationship between the distortion of the right view and the rate of the left view. Generally, a good left view will provide a good prediction of the right view, thereby resulting in a low distortion of the right view. To bridge $\frac{\partial D_{L,i}}{\partial R_{L,i}}$ and $\frac{\partial D'_{R,i}}{\partial R'_{R,i}}$ directly, we need to approximate $\frac{\partial D'_{R,i}}{\partial R_{L,i}}$ for the different types of regions. First, for the disoccluded regions, $\frac{\partial D'_{R,i}}{\partial R_{L,i}} = 0$ because these regions cannot be predicted from the base view. Second, regarding the illumination-affected regions, these

IEEE Proof

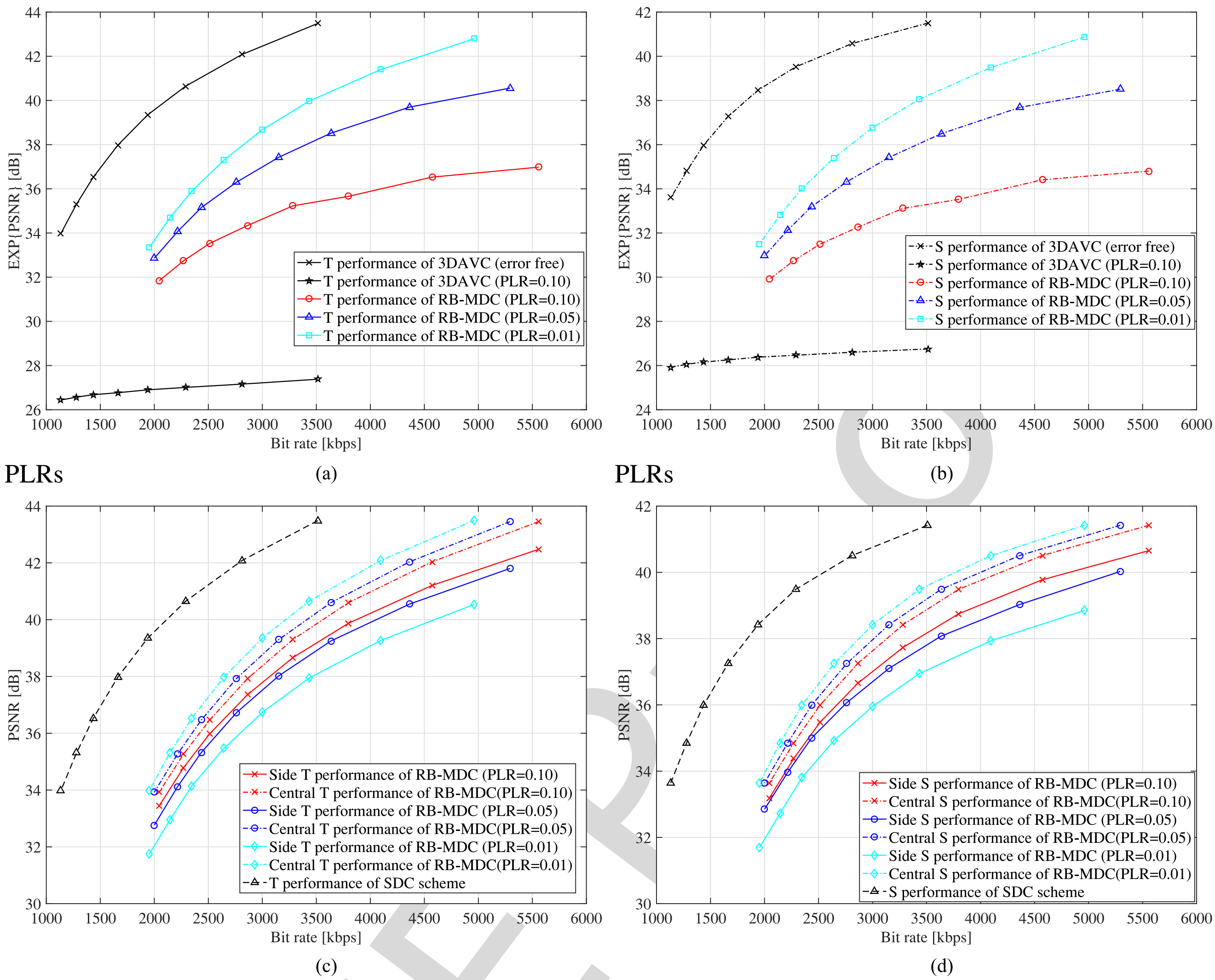

Fig. 11. The rate-PSNR performance of *Mobile*. (a) Expected texture video performance at different PLRs. (b) Expected syntheized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

regions can be predicted, but not well; consequently the following approximation is used: $\frac{\partial D'_{R,i}}{\partial R_{L,i}} = 0.5\frac{\partial D_{L,i}}{\partial R_{L,i}}$. Finally, the remaining regions can be predicted very well; hence, we approximate $\frac{\partial D'_{R,i}}{\partial R_{L,i}} = 1.0\frac{\partial D_{L,i}}{\partial R_{L,i}}$. We note that the values 0, 0.5 and 1 coincide with the rendering mode parameter α and α_L, elaborated as follows. For the disoccluded regions, α and α_L should be zero since these regions exist only in the right enhancement view. For the illumination-affected regions, α and α_L should be 0.5 since these regions in both views have the same importance. For the remaining regions, since these regions can be rendered from the left view with sufficient good quality, α and α_L are set to 1. Therefore, equation (7) can be simplified as

$$\frac{\partial L}{\partial R_{L,i}} = (1-p)^2\left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D_{V,i}}{\partial R_{L,i}}\right) + p(1-p)\left((1+\alpha^2)\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D_{LV,i}}{\partial R_{L,i}}\right) + \lambda = 0 \tag{9}$$

By substituting both (2) and (4) into (9) and (8), we obtain

$$\frac{\partial L}{\partial R_{L,i}} = (1-p)^2\left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \alpha\frac{\partial D_{L,i}}{\partial R_{L,i}}\right) + p(1-p)\left((1+\alpha^2)\frac{\partial D_{L,i}}{\partial R_{L,i}} + \alpha_L^2\frac{\partial D_{L,i}}{\partial R_{L,i}}\right) + \lambda = 0 \tag{10}$$

$$\frac{\partial L}{\partial R'_{R,i}} = p(1-p)\left(\frac{\partial D'_{R,i}}{\partial R'_{R,i}} + (1-\alpha_L)^2\frac{\partial D'_{R,i}}{\partial R'_{R,i}}\right) + \lambda = 0 \tag{11}$$

By combining (10) and (11), we can obtain the rate-distortion function describing the relationship between the base view and the enhancement view,

$$\left((1+\alpha^2)(1-p) + (1+\alpha^2+\alpha_L^2)p\right)\frac{\partial D_{L,i}}{\partial R_{L,i}} = p(2-\alpha_L^2)\frac{\partial D'_{R,i}}{\partial R'_{R,i}} \tag{12}$$

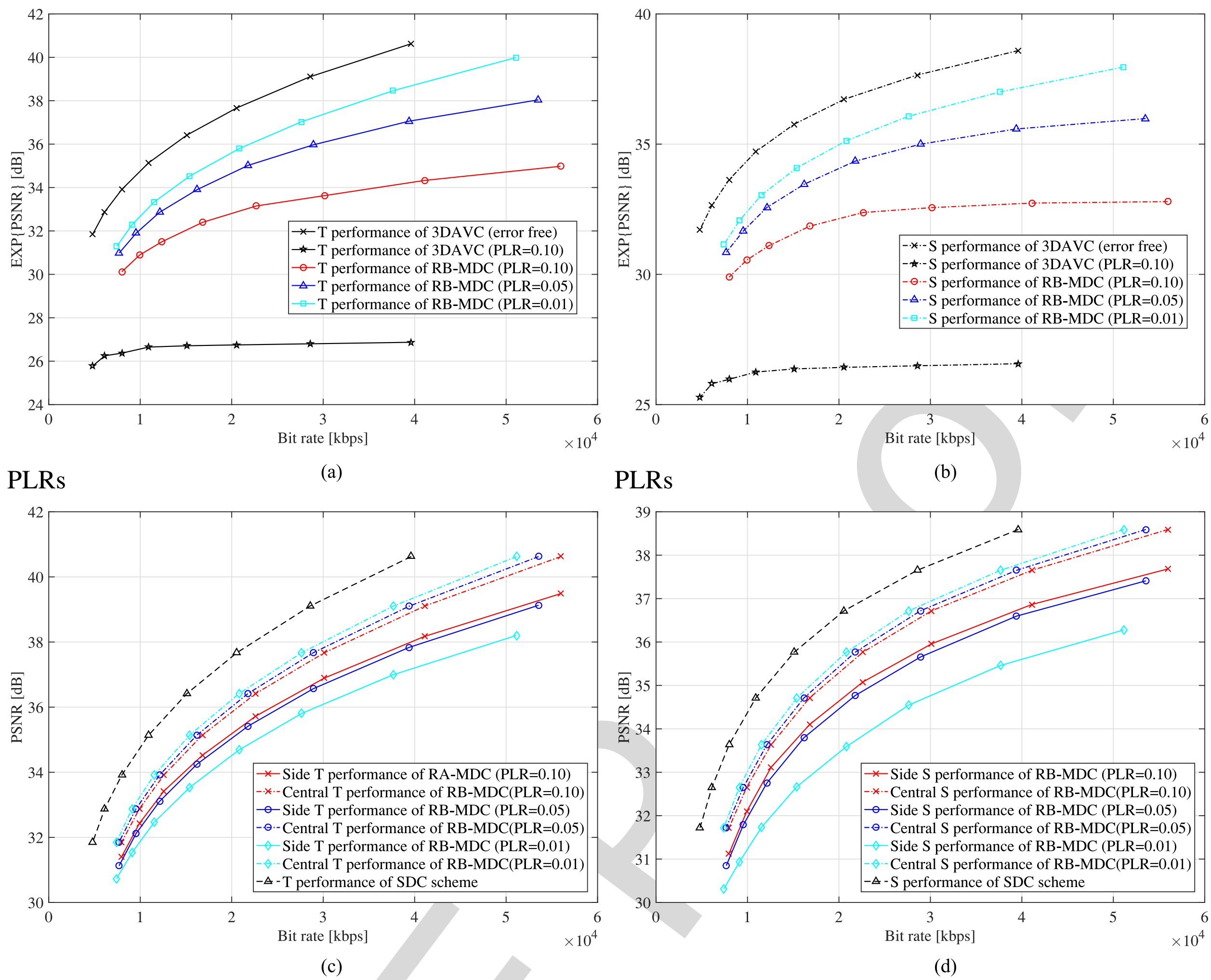

Fig. 12. The rate-PSNR performance of *Undodancer*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

3) QP Relationship: To obtain the relationship between the quantization parameters (QPs) of the two views, the standard H.264/AVC rate-distortion function is employed as follows

$$\frac{\partial D}{\partial R} = -0.85 * 2^{\frac{QP-12}{3}} \tag{13}$$

By inserting (13) into (12), we can obtain the QP values for the left base view and the right enhancement view:

$$QP_R' = QP_L + 3 * log2 \times \frac{((1+\alpha^2)(1-p) + (1+\alpha^2+\alpha_L^2)p)}{p(2-\alpha_L^2)} \tag{14}$$

where QP_L and QP_R' are the quantization parameters of the left base view and the right enhancement view, respectively. It can be observed that QP_R' depends on the packet loss rate (PLR) p and on the weight parameters α and α_L. Fig. 5 shows the relationship of $\Delta QP = QP_R' - QP_L$ with PLR, α and α_L. Here, QP_L is set to 21; therefore ΔQP is no larger than 30. We can observe that the larger the values of α and α_L are, the higher ΔQP will be. Moreover, the lower the value of p is, the higher ΔQP will be. These two trends are intuitive. The entire rate-distortion function also applies to the quantization of the depth maps.

The process of determining QP for description 2 is similar to that for description 1. With the assigned QP of the base view, we can calculate the QPs for each region under different network status using (14); thus a redundancy allocation formula is obtained. Note that QP is assigned on the macroblock(MB) level. However, some MBs are likely to contain both disoccluded pixels and general pixels.Hence, we need to calculate the ratios representing the proportions of an MB that are occupied by pixels of each different type. These ratios can be included in the QP calculation because different types of regions have different α and α_L values.

III. Experimental Results and Analysis

In this section, experiments are conducted using the following video sequences:*Newspaper* (1024 × 768) [19], *Lovebird* (1024 × 768) [20], *Balloons* (1024 × 768) [17], *Kendo* (1024 × 768) [17], *BookArrival* (1024 × 768) [21], *Mobile* (720 × 540) [18] and *UndoDancer* (1920 × 1088) [18]. The depth information is estimated versions for *Newspaper*, *Lovebird*, *Balloons*, *Kendo*

TABLE I
DISOCCLUDED RATIO OF EACH SEQUENCE

Sequence	Newspaper	Lovebird	Balloons	Kendo	Bookarrival	Mobile	Undodancer
Resolution	1024×768	1024×768	1024×768	1024×768	1024×768	720×540	1920×1088
disocclusion Ratio	0.088	0.0145	0.044	0.027	0.062	0.039	0.021

and *BookArrival*, whereas computer-generated (CG) depth is used for *Mobile* and *UndoDancer*. The description for the sequences can be found in [22]. For each sequence, two texture plus two depth videos are encoded with 3D-AVC [23] [24] to generate one description. The virtual views are synthesized using view synthesis reference software VSRS-1D-fast due to its fast and good performance [25], [26]. In detail, view 4 and view 6 of *Newspaper* were used to synthesize virtual view 5. View 6 and view 6 of *Lovebird* were used to synthesize virtual view 7. View 1 and view 3 of *Balloons* were used to synthesize virtual view 2. View 1 and view 3 of *Kendo* were used to synthesize virtual view 2. View 8 and view 10 of *BookArrival* were used to synthesize virtual view 9. View 4 and view 6 of *UndoDancer* were used to synthesize virtual view 5. View 1 and view 5 of *UndoDancer* were used to synthesize virtual view 3. The distortions of the virtual views were calculated between the virtual view images synthesized from the original texture plus depth videos and those synthesized from the decoded texture plus depth videos.

The described algorithm was implemented in the 3D-AVC reference software [27], and the important parameters are detailed in the following. The QP values for the base views were chosen from within a range of [22: 36] in step 2 to consider different rate-distortion points, whereas the QP' values for the enhancement views were determined using equation (14). The threshold for the illumination-affected regions is set as JND value frame by frame. Notice probably larger gain can be achieved if this threshold is set frame by frame according to video contents. However, high computation should be introduced to get this threshold. For description 1, the left view and right view were treated as the base view and enhancement view, respectively. The opposite view allocation was applied in description 2. Ultimately, the QP' values lay in the range [QP_P, 51]. The IPPP coding structure was used throughout the entire experiment and each row of MBs in each frame was encoded in one slice, which was then carried in one transport packet. This entire configuration was chosen to be similar to that used in the rate-distortion-optimized mode switching method [28] to facilitate a comparison of the results.

All experiments were performed in two parts: one to investigate the expected performance and one to investigate the side/central performance at different PLRs. For each part, the results for the left/right views and synthesized virtual views are presented separately. Here, the bit rate includes both descriptions (textures plus depth maps) used in our scheme. For the expected performance assessment, the Bernoulli channel model was adopted, and the performance was measured in terms of the average luminance peak signal-to-noise ratio (PSNR) obtained in 50 independent transmission trials. Side/central curves are presented to represent the performance for the case in which only one channel is working or both channels are working, where the side performance is measured as the average of the two side distortions. Three different packet loss rates of 10%, 5%, and 1% were selected for testing. Error-free results of single description coding are also presented for comparison.

Since the MVD coding structure is still new, few MDC schemes for this format have been introduced. However, several efficient error-resilient algorithms have been proposed for this format and thus can be considered for comparison here [28], [29], [30]. In [28], a rate-distortion-optimized mode switching (RDOMS) scheme was proposed that attempts to optimize the mode decision process considering the end-to-end distortion for error-resilient MVD. Bruno Macchiavello et al. have proposed a loss-resilient coding technique for free-view point videos [30]. The results of these two schemes on the *Newspaper* and *Lovebird* sequences are also reported here. To save room in the figures, our region-based multiple description scheme is abbreviated as RB-MDC, whereas T and S are used to represent a texture view and a synthesized view, respectively. To quantify the impairment caused by the introduced redundancy, we also include the results of the single description scheme (SDC), that is, the results of the classical 3D-AVC method.

In Subfigure a) and Subfigure b) of Figs. 6 and 7, the expected performances for the left/right views and synthesized views, respectively, are presented. It can be observed that our scheme is considerably superior to RDOMS [28], with gains of up to 2 dB on both the texture images and the synthesized views. However, when the bit rate is lower, the gains are relative smaller since our scheme is much more effective at normal and high bir tate cases. Note that the results of RDOMS and Bruno Macchiavello's scheme were obtained from [28]. There are many configuration parameters that can be modified during encoding, and we tried our best to make the configuration as similar as possible to that used in [28]. Furthermore, RDOMS only optimizes the mode selection, whereas our approach introduces MDC. Consequently, it is not truly fair to compare these schemes with ours since MDC has an advantage when the packet loss rate is high. However, this 2 dB gain still demonstrates the effectiveness of the proposed scheme.

In Figs. 6 and 7, the curves for the single description scheme in the error-free case are also included. We note that the gap between the error-free case and the proposed method is small when the bit rate is high because our bit allocation strategy can achieve better performance at higher bit rates. When the bit rate is lower, the three curves at the different PLRs tend to be very close. This is mainly because of the higher QP for a lower bit rate. On the one hand, a large QP will cause many macroblocks to be processed in skip mode, meaning that our bit allocation

IEEE Proof

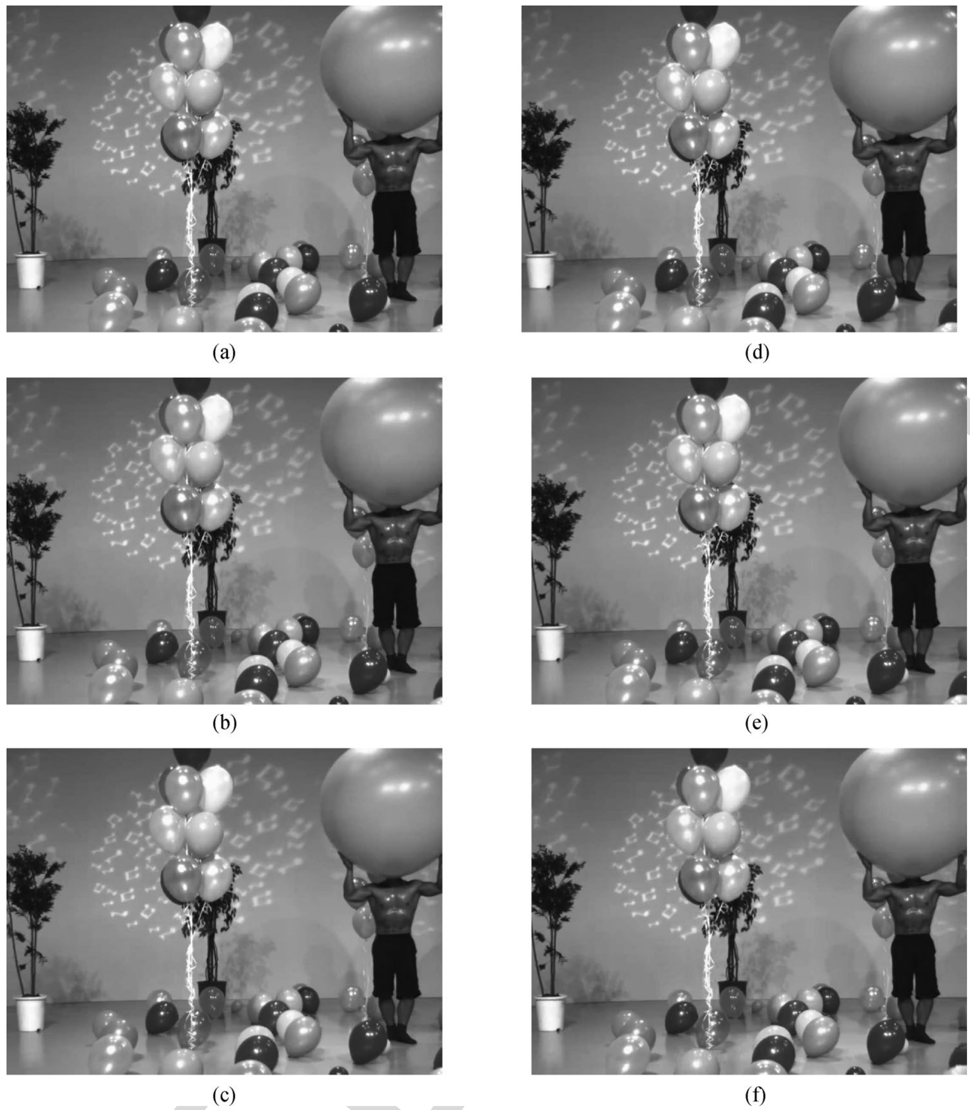

Fig. 13. Subjective visual results for *Balloons*

based on ΔQP will not work. On the other hand, with a large QP, we have less freedom to tune ΔQP because $QP + \Delta QP$ cannot be larger than 51 according to the H.264/AVC standard. For comparison, the results for the SDC scheme with $PLR = 0.10$ and $PLR = 0$ are also included. It can be observed that the proposed scheme is far superior to the SDC scheme in the presence of packet loss, in terms of both texture video performance and synthesized view performance. Moreover, packet loss affects the synthesized view performance more than the texture video performance since the synthesis depends on both the texture and depth images.

Subfigure (c) and Subfigure (d) of Figs. 6 and 7 present the side/central performances for the texture views and the synthesized views, respectively. Here, the performance of left and right views are averaged to provide that of texture view. We can observe that different trade-offs between side and central performance can be achieved under different channel statuses. In addition, the packet loss rate affects the introduced redundancy; a higher PLR corresponds to a higher redundancy. For example, the best side performance is achieved for a high PLR (0.10), whereas the best central performance is observed at a low PLR (0.01). We can determine the additional bit-rate cost for our central description that is required to achieve the same PSNR as that in the error-free SDC case, which is equivalent to the introduced redundancy. The different side description curves represent the performances achieved with different redundancy allocations when only one channel is working. We find that all performances are acceptable, even when one channel is completely nonfunctional. Note that the gain originates from our effective bit-rate allocation strategy for both the color videos and the depth maps.

Figs. 8–12 present the rate-PSNR performances on the *Balloons*, *Kendo*, *BookArrival*, *Mobile* and *UndoDancer* video sequences. These results confirm that the proposed technique exhibits good behavior regardless of the video content and resolution. Note that we treat holes as disoccluded regions. Hence, for depth maps that contain excessive noise, many holes or disocclusion regions will be generated and the efficiency of the proposed scheme will be affected. In the extreme case in which there are no holes or disoccluded regions, our scheme can achieve the maximum bit-rate savings and is the most effective. In the contrast, if the baseline it too large, many large holes will be generated. Our scheme will cost too many bits to deal with this situation. However, general baseline are not too larger, otherwise we cannot get a good 3D feeling. The disocclusion ratio for each sequence is listed in Table I. For example, *Balloons* contains relatively few disoccluded and illumination-affected regions; thus, its total redundancy is relatively low, and its expected performance is the best among all three sequences with the same resolution. *Newspaper* contains relatively more disoccluded and illumination-affected regions, and consequently, its expected performance is relatively worse. As for *Mobile* and *UndoDancer* with CG depth map, it is not fair to compare these sequences with the other four sequences since they have different resolution and bit-rate ranges. In fact, depth map of *UndoDancer* and *UndoDancer* have few disoccluded and illumination-affected regions, without any noise in depth maps; hence, for a good fixed central performance, its side performance and central performance are relative closer compared with the results of other sequences, due to its low introduced redundancy.

In addition to objective results, some subjective results are provided in Fig. 13. Here, the 10th frame of *Balloon* in view 1, together with the 10th frame in its corresponding synthesized view, are selected to demonstrate the performance. Our RB-MDC are configured at packet loss rate 5%. In order to evaluate the performance, the results of single description (SDC) case are included, in which the total bit rates of our MDC scheme and that of SDC are tuned to be similar as 5000 kbps. Since there are redundancy inserted in RB-MDC scheme, the results of ours is at disadvantage compared with that of SDC at error free case. In fact, there are some distortion around the balloons and trees, however, we cannot notice big visual difference between ours and that of SDC. Particularly, the side visual results that supposes one description is broken down are also very good, which demonstrate the efficiency of our scheme.

IV. Conclusion

In this paper, a region-based multiple description coding scheme for multiview video plus depth is proposed. First, regions are classified into disoccluded, illumination-affected and remaining regions according to their contributions to the virtual view to be synthesized. Second, an optimized expected rate-distortion function is designed based on both the texture video distortions and synthesized view distortions. By assigning different quantization parameters to the three types of regions depending on the channel status, we can minimize the expected distortion. Compared with traditional error-resilient 3D-AVC schemes, the proposed scheme can achieve gains of up to 2 dB in the case of packet loss. In addition, different prioritizations between side and central performance can be applied under different channel conditions, which is a desirable feature of any MDC scheme. An analysis of the experimental results shows that the proposed MDC scheme is a promising approach for the transmission of MVD-format 3D videos over error-prone channels.

It should be noted that our scheme achieves much better performance when the bit rate is higher. This is because our rate allocation strategy is more accurate at higher bit rates by virtue of the larger possible range of ΔQP. In addition, the performance of our scheme is also affected by the quality of the depth maps. If noise is present in the depth maps, such as noise due to depth estimation, irregular holes will be generated and the coding efficiency will consequently deteriorate. Hence, depth maps acquired via time-of-flight sensors must be subjected to noise reduction processing, which may be investigated in our further work.

References

[1] A. Vetro, A. Tourapis, K. Muller, and T. Chen, "3D-TV content storage and transmission," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 384–394, Jun. 2011.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, Sep. 2007, pp. 201–204.

[3] J. Y. Lee *et al.*, "Depth-based texture coding in AVC-Compatible 3D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1347–1361, Aug. 2015.

[4] F. Shao, G. Jiang, W. Lin, M. Yu, and Q. Dai, "Joint bit allocation and rate control for coding multi-view video plus depth based 3D video," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1843–1854, Dec. 2013.

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

[6] C. Zhu, S. Li, J. Zheng, Y. Gao, and L. Yu, "Texture-aware depth prediction in 3D video coding," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 482–486, Jun. 2016.

[7] V. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74 –93, Sep. 2001.

[8] Y. Wang, A. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.

[9] C. Lin, T. Tillo, Y. Zhao, and B. Jeon, "Multiple description coding for H.264/AVC with redundancy allocation at macro block level," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 589–600, May 2011.

[10] H. Karim, A. Sali, S. Worrall, A. Sadka, and A. Kondoz, "Multiple description video coding for stereoscopic 3D," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2048–2056, Nov. 2009.

[11] A. Norkin *et al.*, "Schemes for multiple description coding of stereoscopic video," in *Proc. Int. Conf. Multimedia Content Representation, Classification Security*, 2006, pp. 730–737.

[12] X. Wang and C. Cai, "Mode duplication based multiview multiple description video coding," in *Proc. Data Compression Conf.*, Mar. 2013, pp. 527–527.

[13] C. Lin, Y. Zhao, T. Tillo, and J. Xiao, "Multiple description coding for stereoscopic videos with stagger frame order," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 1016–1025, Jun. 2015.

[14] J. Guo, H. Bai, C. Lin, M. Zhang, and Y. Zhao, "Intra-/inter-view correlation based multiple description coding for multiview transmission," in *Proc. Data Compression Conf.*, Apr. 2015, pp. 446–446.

[15] J. Jin, A. Wang, Y. Zhao, C. Lin, and B. Zeng, "Region-aware 3-D warping for DIBR," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 953–966, Jun. 2016.

[16] J. Wu *et al.*, "Enhanced just noticeable difference model for images with pattern complexity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2682–2693, Jun. 2017.

[17] Nagoya University, "3DV test sequences." [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/ mpegftv/mpeg3dv/CfP/ Q1

[18] Nokia, "3DV test sequences." [Online]. Available: ftp:[illegible]g3dv.research.nokia.com

[19] Gwangju Institute [illegible]ience and Technology, "3DV test sequences." [Online]. Available: ftp://203.253.128.142

IEEE Proof

[20] Electronics and Telecommunications Research Institute, “3DV test sequences.” [Online]. Available: ftp://203.253.128.142

[21] Fraunhofer, “3DV test sequences.” [Online]. Available: http://sp.cs.tut.fi/mobile3dtv/stereo-video/

[22] *Call for Proposals on 3D Video Coding Technology*, International Organization for Standardization, ISO/IEC JTC1/SC29/WG11 mpeg2011/N12036, 2011.

[23] B. T. Oh and K. J. Oh, “View synthesis distortion estimation for AVC- and HEVC-compatible 3-D video coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 1006–1015, Jun. 2014.

[24] D. Rusanovskyy, F.-C. Chen, L. Zhang, and T. Suzuki, “3DAVC test model 8,” JCT-3V Document JCT3V-F1003m, Nov. 2013.

[25] Y. Chen, G. Tech, K. Wegner, and S. Yea, “Test model 8 of 3D-HEVC and MV-HEVC,” JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC1/SC 29/WG 11, Doc. JCT3V-H1003, 2014.

[26] D.-Y. Kim, W.-S. Jang, and Y.-S. Ho, “Analysis of View Synthesis Methods (VSRS 1D fast and VSRS3.5),” Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 2012.

[27] “3D-AVC reference software ATM 13.1.” [Online]. Available: http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/3DV-ATMv1 3.1

[28] P. Gao and W. Xiang, “Rate-distortion optimized mode switching for error-resilient multi-view video plus depth based 3-D video coding,” *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1797–1808, Nov. 2014.

[29] A. Vosoughi, V. Testoni, P. Cosman, and L. Milstein, “Joint source-channel coding of 3D video using multiview coding,” in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 2050–2054.

[30] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, and W. T. Tan, “Loss-resilient coding of texture and depth for free-viewpoint video conferencing,” *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 711–725, Apr. 2014.

Authors’ photographs and biographies not available at the time of publication.

IEEE Proof

QUERIES

Q1. Author: Please provide the years in Refs. [17]–[21] and [27].

IEEE Proof

Region-Based Multiple Description Coding for Multiview Video Plus Depth Video

Chunyu Lin, Yao Zhao, *Senior Member, IEEE*, Jimin Xiao, and Tammam Tillo, *Senior Member, IEEE*

***Abstract*—Interframe and interview predictions are widely employed in multiview video coding. This technique improves the coding efficiency, but it also increases the vulnerability of the coded bitstream. Thus, one packet loss will affect many subsequent frames in the same view and probably in other referenced views. To address this problem, a region-based multiple description coding scheme is proposed for robust 3-D video communication in this paper, in which two descriptions are formed by setting the left and right view as dominant in the first and second description, respectively. This approach exploits the fact that most regions in the reference view could be synthesized from the base view. Hence, these regions could be skipped or only coarsely encoded. In our work, the disoccluded regions, illumination-affected regions, and remaining regions are first determined and extracted. By assigning different quantization parameters for these three different regions according to the network status, an efficient multiple description scheme is formed. Experimental results demonstrate that the proposed scheme achieves considerably better performance compared with the traditional approach.**

***Index Terms*—Multiple description coding, multiview video plus depth, video coding.**

I. Introduction

3D VIDEOS are able to provide depth perception through appropriate 3D display devices, which increases the immersive experience for the audience. Depending on whether glasses are required, 3D displays can be classified as stereoscopic or auto-stereoscopic. Stereoscopic displays require two texture/color views, and each view is projected to one of the eyes of the viewer through special glasses. Since wearing such glasses in a living room is uncomfortable and inconvenient, many studies focus instead on the auto-stereoscopic format. Auto-stereoscopic format provide different views depending on viewers' position and angle. Hence, a viewer can switch views by shifting his head position. However, to achieve this motion parallax feature of the auto-stereoscopic format, more views must be provided, which increases the burden of encoding and transmission.

Multiview video coding (MVC) standard was developed to efficiently compress multiple view data through inter-frame and inter-view predictions [1]. However, this approach only reduces the transmission burden partly because many views are still required. Multiview video plus depth (MVD) format was introduced as a new 3D video format [2] that includes texture images and their associated depth maps. By employing the depth image-based rendering (DIBR) technique, arbitrary virtual views can be generated; thus only a small number of views are required to be processed and transmitted [3]. Because of this advantage, the MVD format is being widely studied in industry and academia [4], [5], [6]. Among the MVD formats, a scheme based on two views plus two depth maps is the most popular because it requires relative little data and shows good synthesis performance. The use of two views plus two depth maps allows the disocclusion problem to be much more effectively mitigated compared with the use of just one view plus one depth map. Hence, this MVD format is also our focus in this paper. In this type of MVD format, one view is selected as the base/dominant view and is encoded using traditional intra/inter prediction, and the other view is designated as the enhancement/reference view and is encoded using intra/inter and inter-view predictions. Unless otherwise specified, the terms base view and dominant view will be used interchangeably throughout this paper, as will enhancement view and reference view.

In addition to the inter-frame prediction adopted in classical 2D video coding, the codec for MVD employs inter-view prediction and view synthesis prediction. Due to the complex prediction structure, the coding efficiency of the MVD format is improved; however, this prediction structure also increases the vulnerability of the coded bitstream to packet loss.

Multiple description coding (MDC) has been proposed as an efficient solution to combat packet loss. It provides a promising framework for video applications in which retransmission is unacceptable [7]. The classical MDC diagram is shown in Fig. 1, in which one source is encoded into two representations (descriptions) that are mutually refinable and can be decoded independently. The two descriptions are then transmitted over separate channels. When the network is experiencing no loss and all the descriptions are received, the best quality is obtained, with a

Manuscript received October 11, 2016; revised May 22, 2017 and August 2, 2017; accepted September 29, 2017. This work was supported in parts by the National Natural Science Foundation of China (No.61772066, No.61210006 and 61501379) and by the Beijing Natural Science Foundation (No. KZ201610005007). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaoqing Zhu. *(Corresponding author: Chunyu Lin)*

C. Lin and Y. Zhao are with the Beijing Key Laboratory of Advanced Information Science and Network, Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: cylin@bjtu.edu.cn; yzhao@bjtu.edu.cn).

J. Xiao is with the Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: jimin.xiao@xjtlu.edu.cn).

T. Tillo is with the Libera Universit di Bolzano-Bozen (unibz), Bolzano 39100, Italy, and also with Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: Tammam.Tillo@unibz.it).

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/TMM.2017.2766043

1520-9210 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

IEEE Proof

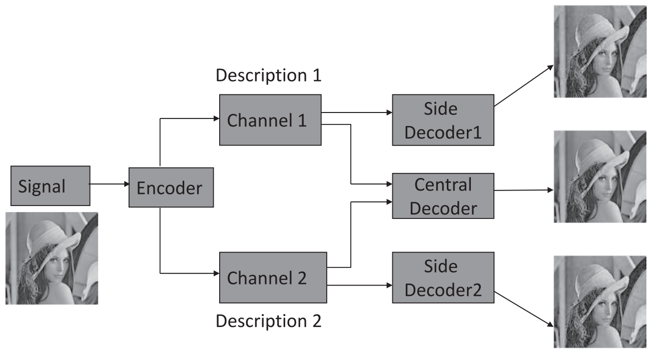

Fig. 1. Classical multiple description diagram.

so-called central distortion. If only one channel is working, the side decoder can reconstruct the source with a certain desired side distortion. To achieve this resiliency, some redundancy should be introduced in the descriptions, which is useful for mitigating packet loss but is detrimental to the central performance when no packets are lost. This redundancy should be tuned according to the network status. For example, the side distortion should be minimized at a high packet loss rate, whereas the central distortion should be minimized at a low packet loss rate. Thus, flexible tuning of the redundancy is a key task for any MDC scheme.

Many MDC works have been proposed for robust 2D video coding [8], [9]. Some studies have also been conducted on the stereoscopic video format [10]. In [11], the spatial scaling MDC scheme (SS-MDC) and the multi-state MDC scheme (MS-MDC) were proposed for stereoscopic videos. In SS-MDC, an asymmetric stereo pair is used to form descriptions, such that one view is at full resolution and the other view is down-sampled. In MS-MDC, temporal down-sampling is applied. For example, the odd frames of both the left and right views are grouped to form one description, whereas the other description contains the information for the even frames. In [12], multiview videos are subsampled in both the horizontal and vertical directions to form four sub-sequences. Then these four sub-sequences are paired to form two descriptions. In each description, one sub-sequence is directly encoded, whereas the other uses mode duplication based on the mode of the sub-sequence in the other description. This scheme is simple and efficient; however, its redundancy allocation is not flexible. In [13], an MDC video coding scheme for stereoscopic video was proposed based on a stagger frame order. All these schemes are very efficient; however, little research has yet been performed on MVD format. In fact, as more predictions are introduced, a bitstream of the MVD format becomes more vulnerable and requires greater protection. Otherwise, one packet loss in one frame will seriously affect the current view and the other reference views, as well as the virtual synthesized view. In addition, most MDC schemes for 3D videos are merely simple extensions of their 2D versions, such as spatial subsampling or temporal subsampling [11], [14]. Thus, features of MVD are not sufficiently utilized.

In this paper, we propose a region-based multiple description coding scheme (RB-MDC) that attempts to optimize the expected performance considering region importance and channel status. The proposed scheme first differentiates each region in the texture and depth videos with respect to its importance. Based on the differentiated regions, unequal protection is provided according to the importance of each region and the network status. Compared with classical schemes, gains of up to 2 dB can be achieved on both the texture videos and the synthesized views in the case of high packet loss rates.

The remainder of this paper is organized as follows. In Section II, an outline of the proposed scheme is provided, with introductions to region classification in Section II-A and redundancy allocation in Section II-B. Experimental results are presented and analyzed in Section III. Finally, conclusions are drawn in Section IV.

II. Proposed Scheme

The proposed multiple description scheme is illustrated in Fig. 2. Since the two descriptions are formed in the same way, we will take description 1 as an example to describe our algorithm. For description 1, as shown in Fig. 2, the left view is chosen to be the dominant view, whereas the right view is designated as the enhancement view. First, a virtual right view is synthesized from the left view plus depth. Based on the virtual right view, the original right view can be classified into disoccluded regions, illumination-affected regions and the remaining regions. These three types of regions have different effects on the quality of the synthesized views, as will be explained further in the next subsection. Based on this classification, lower bit rates can be assigned to unimportant regions that constitute higher percentages of the overall images. Therefore, redundancy can be flexibly allocated, and the total bit rate can be reduced. For description 2, the right view is the dominant view; otherwise, the process is similar to that for description 1.

If only one description is received, normal quality of the dominant view can be achieved along with a relatively lower quality for the enhancement view. Since the disoccluded regions and illumination-affected regions, which have a higher impact on the virtual view, have been better encoded, we can still obtain well-synthesized virtual views. When both descriptions are received, good central performance can be achieved with both the dominant left view and the dominant right view. Because the two dominant views are employed, better synthesized quality is expected.

A. Region Classification

In Fig. 2, one important step of the scheme is to classify different regions based on their contributions to the virtual left/right views. In the proposed scheme, three types of regions are classified: disoccluded regions, illumination-affected regions and the remaining regions. For the example of description 1, the disoccluded regions, or the regions that appear as a result of view switching, are the pixels in the right view that cannot be rendered from the left view. Regions of this type are the most important because the synthesized views require them but they exist only in the original right view. Notice, the holes due to large baseline are also regarded as disoccluded regions since they cannot be rendered from the base view. The illumination-affected regions

IEEE Proof

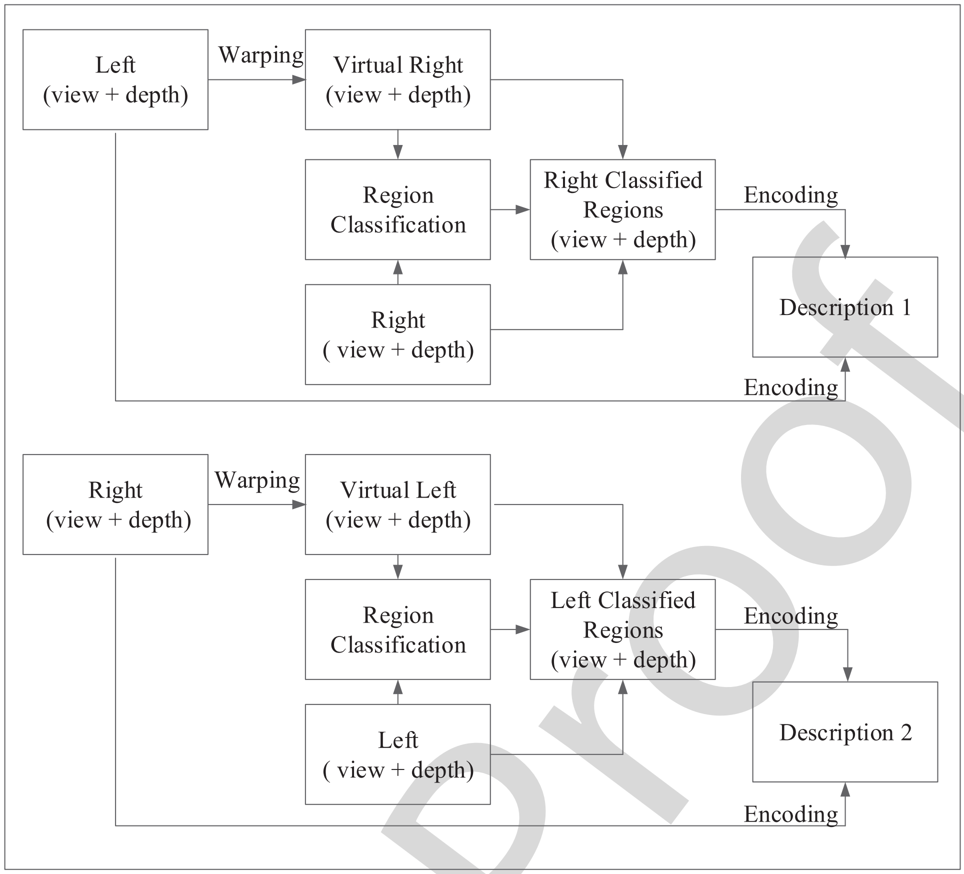

Fig. 2. Region-base multiple description scheme.

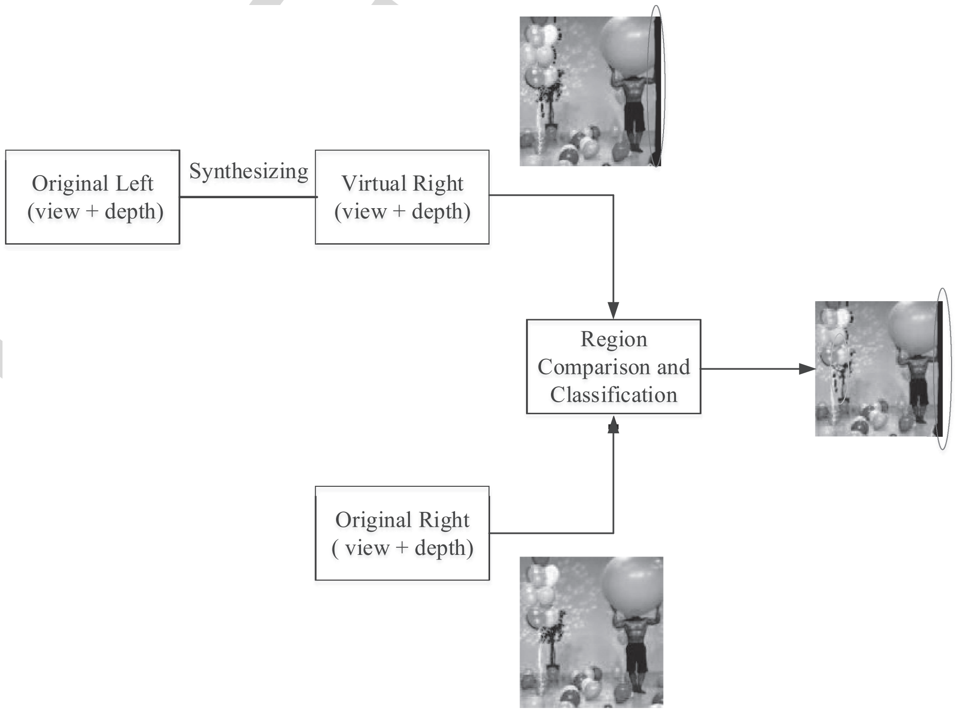

Fig. 3. Region classification process, where regions are highlighted.

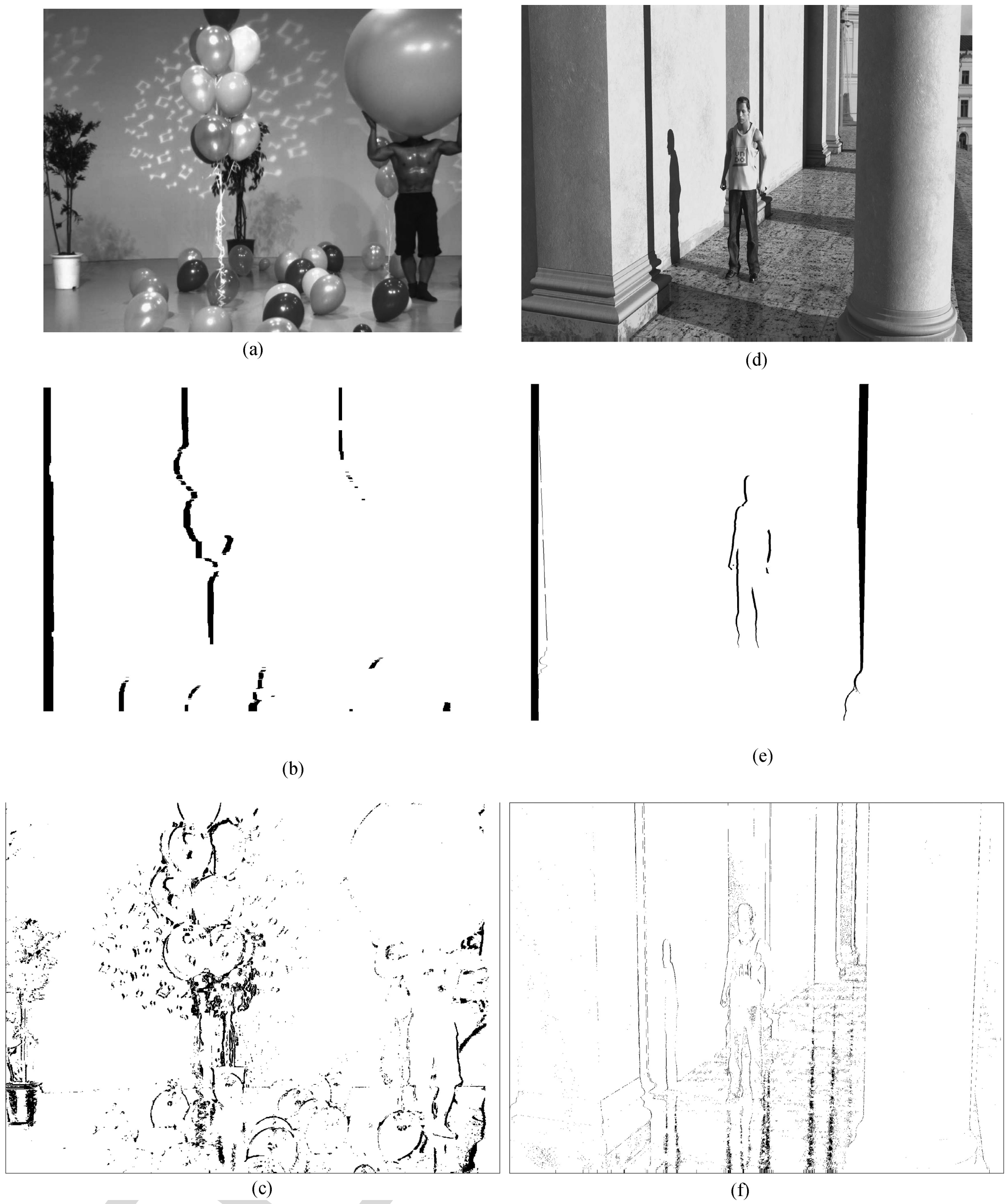

Fig. 4. Region classification example with *Balloons* and *Undodancer*. (a) Original. (b) Disoccluded. (c) Illumination-affected. (d) Original. (e) Disoccluded. (f) Illumination-affected.

are the regions in the right view that can be rendered from the left view but only with low quality. Because of the differences in the illumination conditions between the left and right views, some regions in the rendered virtual right view will differ from those in the original right view, and these regions should be encoded with sufficiently good quality to correct for these differences. Regions of the last type, called the remaining regions, can be rendered from the left view with a sufficient level of quality.

To classify such regions, a synthesis process is required to render the virtual right view from the left view, as shown in Fig. 3. In this synthesis process, only one texture video and one depth map can be employed; hence, many holes will be generated because of a lack of pixel information at the corresponding locations. These holes are represented as black regions in the figure. Note that except in the classification step, the synthesis process in our scheme can generally employ two texture videos

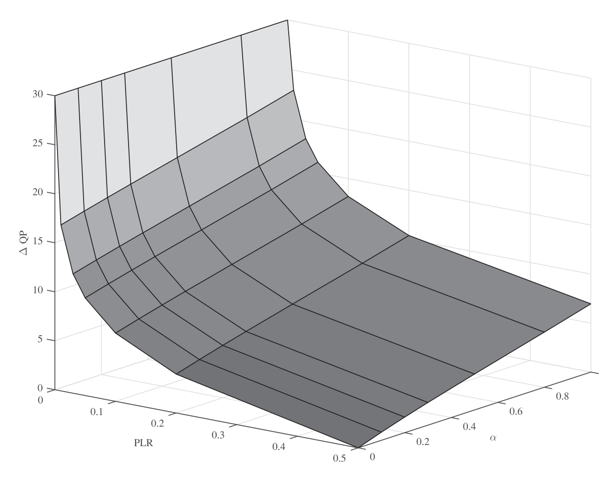

Fig. 5. ΔQP as a function of packet loss rate (PLR) and α.

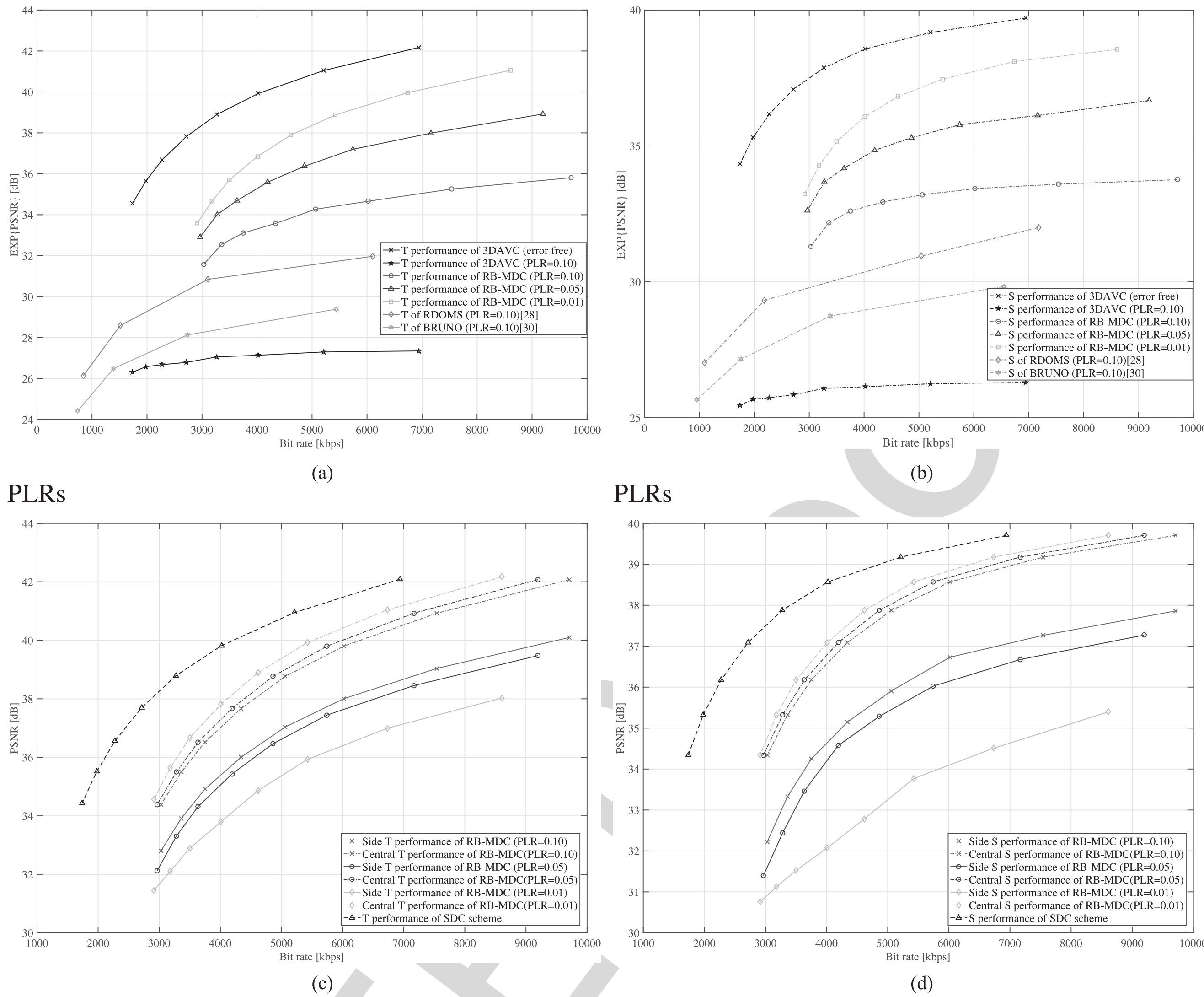

Fig. 6. The rate-PSNR performance of *Newspaper*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

plus two depth maps, which will enable the generation of considerably better virtual views at the cost of increased computation. We can simplify this synthesis process as described in [15]. Compared with the original right view, disoccluded regions can be located easily since these regions are in fact holes.

To determine the illumination-affected regions, the difference between the synthesized virtual view and the original view is first calculated. Regions with value differences larger than a certain threshold are identified as illumination-affected regions. The threshold for illumination-affected region will highly depend on the video contents and it is still an open topic yet. In our case, we set the threshold by a just noticeable difference(JND) [16]. JND is the least perceptible difference that human can notice. In [16], the JND calculation considered both the contrast and pattern complexity, which achieves very good performance. With a given sequence, its JND value is first calculated frame by frame, if a pixel difference between the warped view and the original view is larger than its corresponding JND value, it will be labeled as illumination-affected pixel. After the classification of these two types of regions, the remainder are regarded as remaining regions that can be warped from the dominant view with sufficiently good quality. Hence, the classification process is quite simple.

Examples of region classification are presented in Fig. 4, where the sequences *Balloons* [17] and *UndoDancer* [18] are divided into regions of the three different types. The second and third rows present the disoccluded and illumination-affected regions, respectively, whereas the others show the remaining regions. Here, illumination-affected regions are pixels in which the value difference between the original view and the virtual view is greater than its corresponding JND value. It can be observed that disoccluded regions and illumination-affected regions account for only a small percentage of the entire image. Hence, the allocation of a lower bit rate to the remaining regions, which constitute a large percentage, could considerably reduce the total bit rate. Note that the classification applies to both the color videos and the depth videos. For simplicity, for the depth maps, we just use the classified maps determined for the color videos.

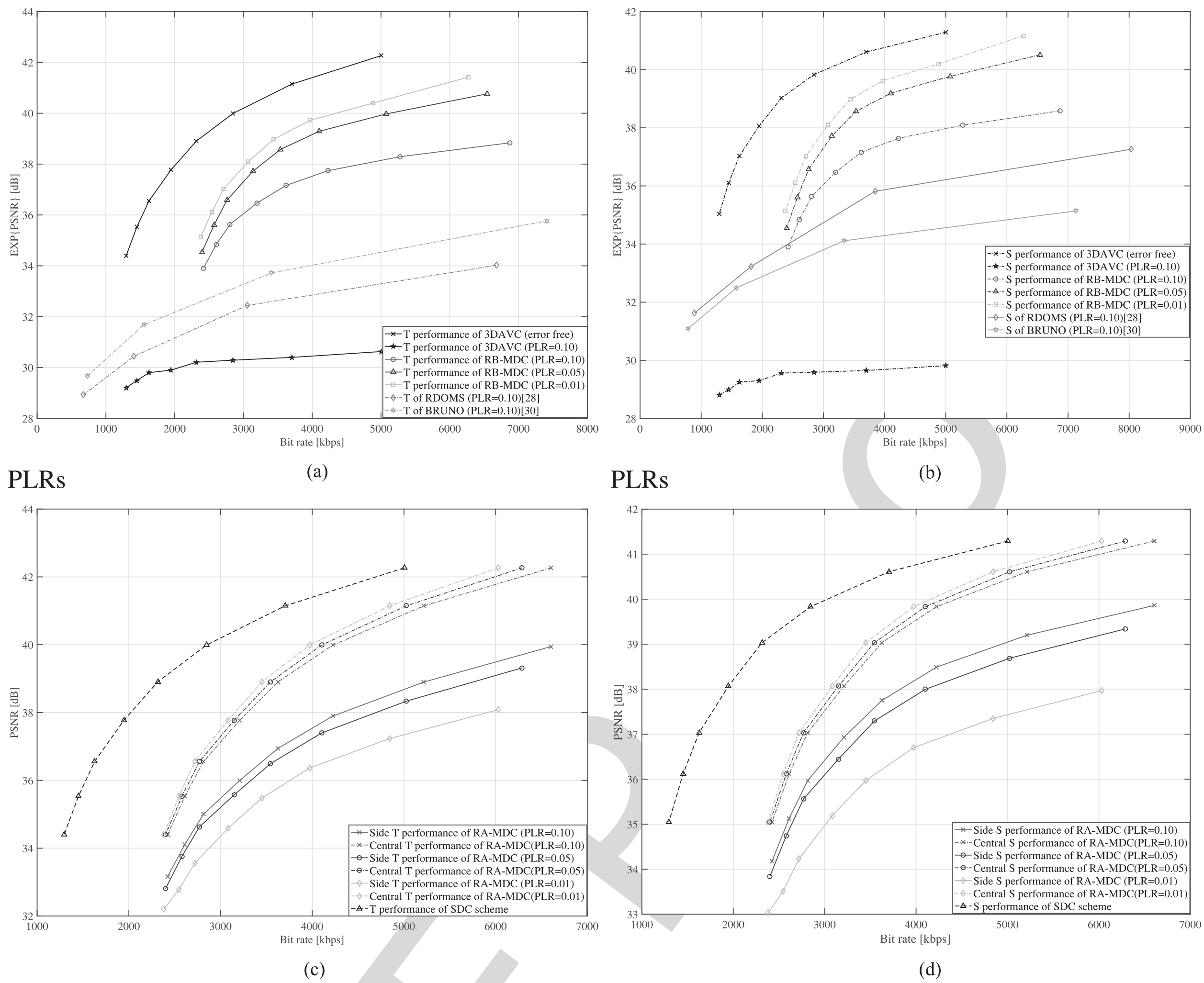

Fig. 7. The rate-PSNR performance of *Lovebird*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

B. Redundancy Allocation

Based on the classified regions, we need to design an effective redundancy allocation algorithm that considers region importance and network status to minimize the expected distortion, including both the real and virtual views, constrained by a fixed total rate. As shown in Fig. 2, additional data representing the two views are required in comparison with the single description scheme (SDC), in which only one pair of views is encoded. The bitstream of the additional views provides redundancy. When the channel quality is not good and the packet loss rate is high, more bits should be assigned to the additional views. By contrast, fewer bits are required when the channel quality is good. Hence, redundancy allocation is a key problem in any MDC scheme. In practice, the disoccluded regions and illumination-affected regions should receive higher protection compared with the remaining regions.

Since these three types of regions have different contributions to the overall performance, different levels of protection or redundancy should be allocated accordingly. Our final goal is to design a rate allocation strategy that considers the relationship among the different types of regions.

First, we need to estimate the expected distortion (left view, right view and virtual views) at the encoder end, considering the network status and the classified regions, under the relevant constraint on the total bit rate. During this process, the distortions of synthesized virtual views must be approximated. Then, we can obtain the rate-distortion function for each region and construct the relationship among the regions accordingly. Finally, we can perform bit-rate allocation based on the different quantization parameter (QP) values calculated from the rate-distortion functions. We will introduce the entire process in detail in the following.

1) Expected Distortion: The expected total distortion should include the distortions of the left and right views as well as of synthesized virtual views. It can be evaluated as

$$\begin{aligned}\bar{D} = (1-p)^2(D_L + D_R + D_V) + p(1-p)(D'_R + D_L \\ + D_{LV}) + p(1-p)(D'_L + D_R + D_{RV}) \\ + p^2(D''_L + D''_R + D''_V) \qquad (1)\end{aligned}$$

where $\bar{D}$ denotes the total expected distortion and p is the packet loss rate. The subscripts L and R denote the left and right views,

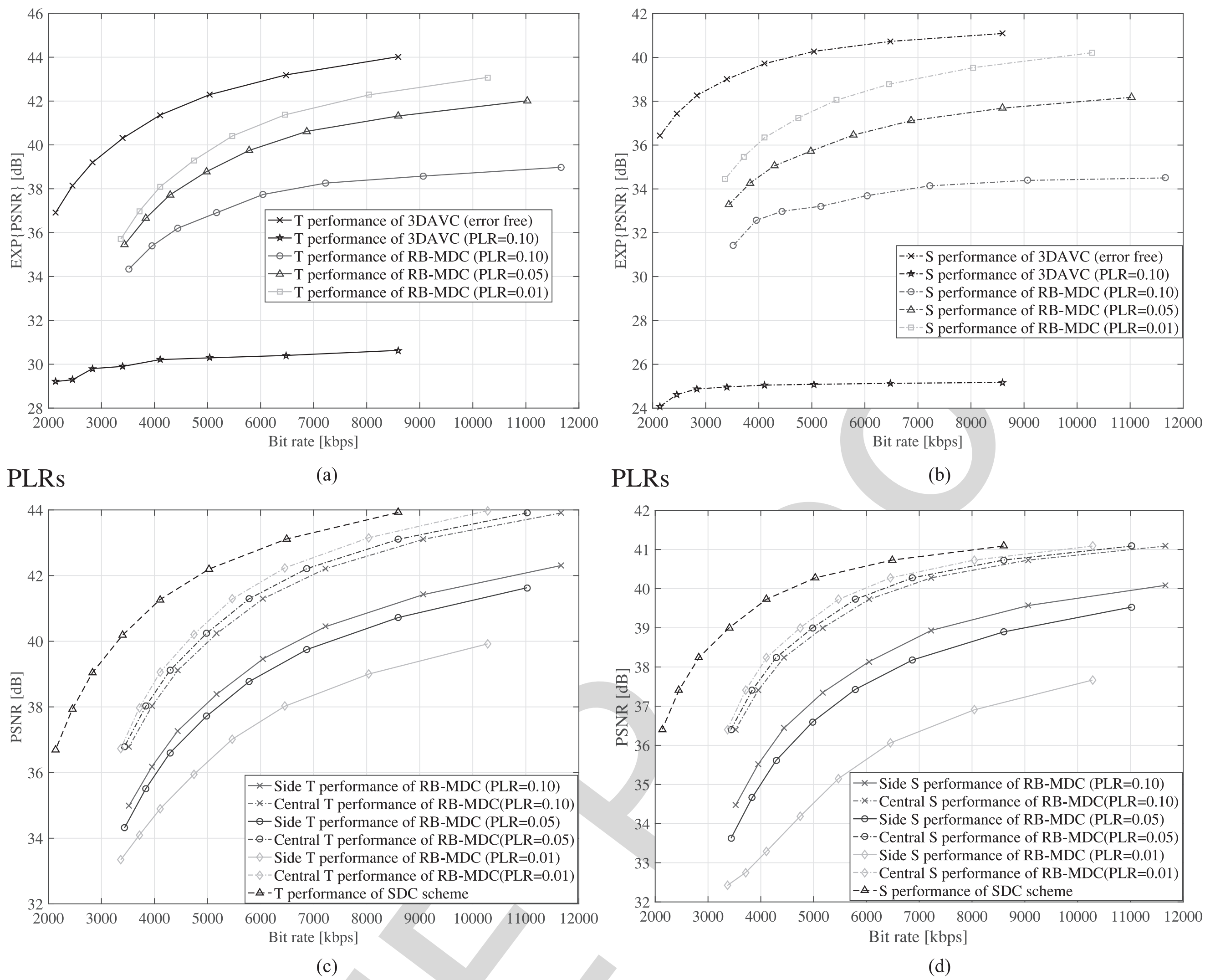

Fig. 8. The rate-PSNR performance of *Balloons*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

respectively, whereas the subscript V represents a synthesized virtual view. D_L and D_R represent the distortions of the left view and right view, respectively, when the dominant mode is used, whereas D'_L and D'_R are the corresponding distortions for the enhancement mode. D_V is the distortion of the view synthesized using the dominant left and right views, whereas D_{LV} and D_{RV} are the distortions of the views synthesized using only the dominant left view or the dominant right view, respectively. Finally, D''_L, D''_R and D''_V are the corresponding distortions with error concealment when the same frames are lost in both the left and right views. The distortions of the left and right views, such as D_L, D_R, D'_L, D'_R, D''_L and D''_R, can be calculated during encoding, whereas those of synthesized views must be estimated and approximated.

The quality of a synthesized view depends on the qualities of the left view and right views as well as on the rendering mode. If the qualities of the left view and the right view are similar, then an averaging mode in which both views are equally important is preferred. Otherwise, an extrapolating mode that uses one dominant view with a higher weight is adopted. Hence, the virtual distortion can be represented as follows:

$$\begin{cases} D_V &= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha)S(\hat{X}_R)\big) - X_V\Big)^2 \\ D_{LV} &= E\Big(\big(\alpha_L S(\hat{X}_L) + (1-\alpha_L)S(\hat{X'}_R)\big) - X_V\Big)^2 \\ D_{RV} &= E\Big(\big(\alpha_R S(\hat{X}_R) + (1-\alpha_R)S(\hat{X'}_L)\big) - X_V\Big)^2 \end{cases} \quad (2)$$

where $S()$ is the synthesis function that renders the left and right views $\hat{X}_L$ and $\hat{X}_R$ into the virtual view; X_V is the original virtual view synthesized from the original left view X_L and the original right view X_R; and α, α_L and α_R are the rendering mode parameters. For example, α can be set to 0.5 when the left view and right view are of similar quality. In practice, the rendering process is also affected by different types of regions. Suppose that one view is designated as the dominant view; then, most regions in the virtual view will be rendered from this dominant view, whereas the disoccluded regions must be rendered from the other view.

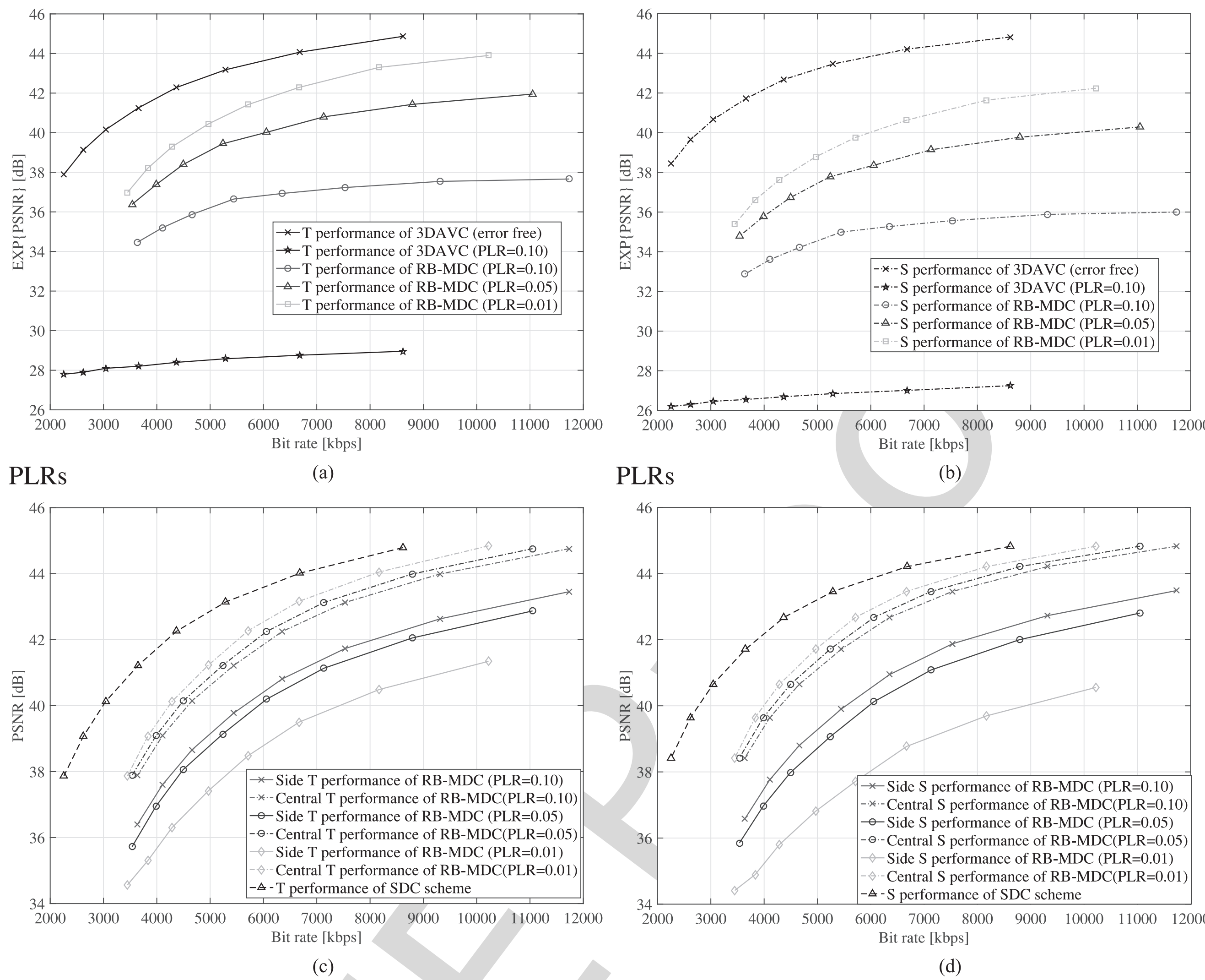

Fig. 9. The rate-PSNR performance of *Kendo*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

Since the three distortions D_V, D_{LV} and D_{RV} can be calculated in a similar manner, we will simply discuss D_V as an example.

$$
\begin{aligned}
D_V &= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha) S(\hat{X}_R)\big) - X_V\Big)^2 \\
&= E\Big(\big(\alpha S(\hat{X}_L) + (1-\alpha) S(\hat{X}_R)\big) \\
&\quad - \big(\alpha S(X_L) + (1-\alpha) S(X_R)\big)\Big)^2 \\
&= E\Big(\alpha\big(S(\hat{X}_L - X_L)\big) + (1-\alpha)\big(S(\hat{X}_R - X_R)\big)\Big)^2 \\
&\approx \alpha^2 D_L + (1-\alpha)^2 D_R \\
&\quad + 2\alpha(1-\alpha) E\big(S(\hat{X}_L - X_V) S(\hat{X}_R - X_V)\big) \\
&= \alpha^2 D_L + (1-\alpha)^2 D_R
\end{aligned} \tag{3}
$$

The virtual distortion primarily depends on the views to be rendered; hence, we approximate the distortions $(S(\hat{X}_L) - X_V)^2$ and $(S(\hat{X}_R) - X_V)^2$ as $(\hat{X}_L - X_L)^2 = D_L$ and $(\hat{X}_R - X_R)^2 = D_R$, respectively. In addition, $E\big(S(\hat{X}_L - X_V) S(\hat{X}_R - X_V)\big)$ is assumed to be zero since these two errors are uncorrelated [1].

In the same way, we can obtain the other two virtual distortion formulas

$$
\begin{cases}
D_{LV} = \alpha_L^2 D_L + (1-\alpha_L)^2 D'_R \\
D_{RV} = \alpha_R^2 D_R + (1-\alpha_R)^2 D'_L
\end{cases} \tag{4}
$$

2) Rate-Distortion Functions: In our scheme depicted in Fig. 2, the left view is encoded as the base view and the right view is encoded as the enhancement view in description 1, and vice versa for description 2. We set the quality of the base view as an anchor, and our key objective is to determine the quality of the enhancement view depending on its region classification and the network status. Let the bit rates of the base views be R_L and R_R, whereas the bit rates of the enhancement views are R'_L and R'_R. The problem can be expressed as follows:

IEEE Proof

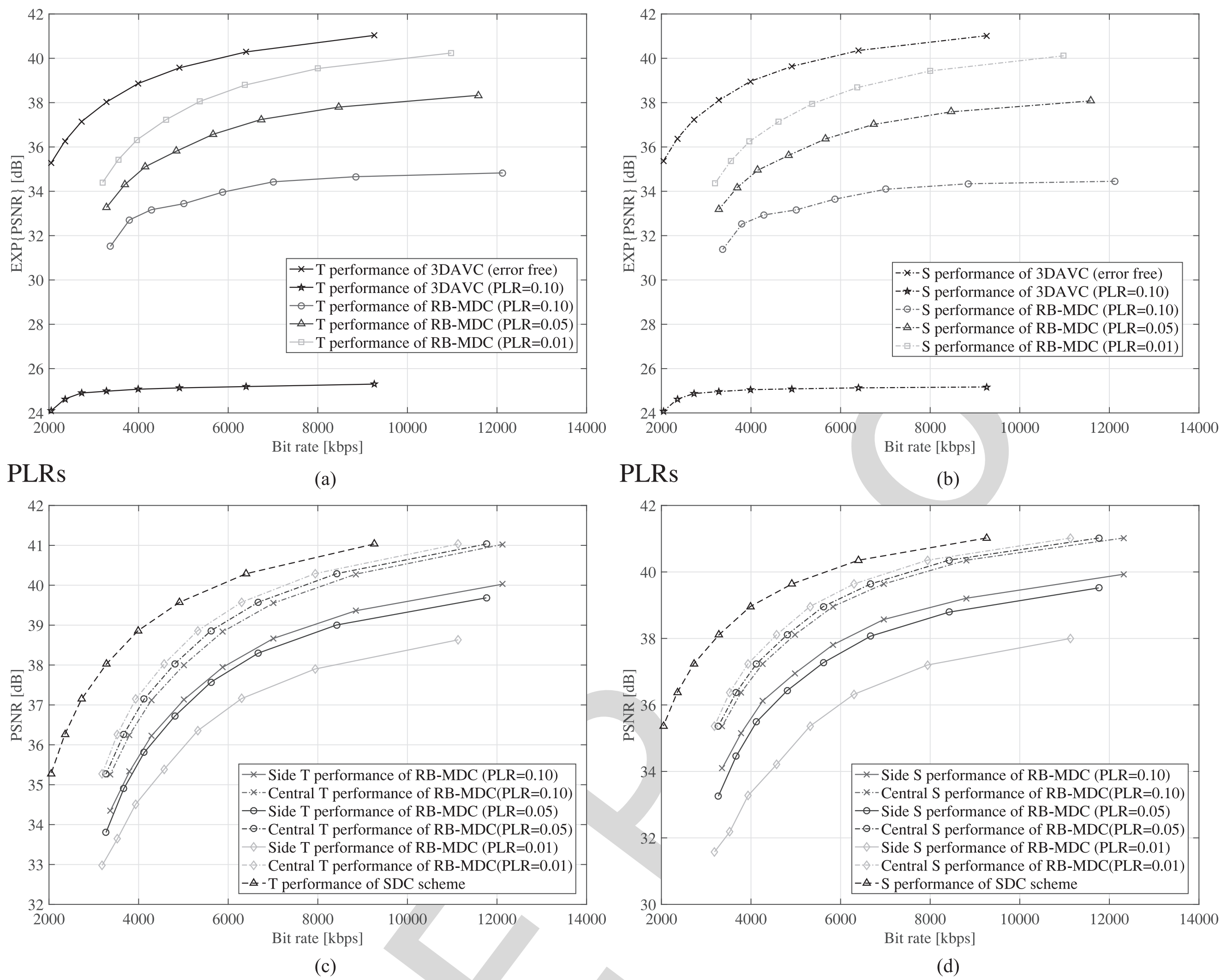

Fig. 10. The rate-PSNR performance of *Bookarrival*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

$$\begin{cases}\min \quad \bar{D}=\sum_{i=1}^{N} \bar{d}[i]\\ \text{s.t.}\ \ R_t=\sum_{i=1}^{N}(R_L[i]+R_R[i]+R'_L[i]+R'_R[i])\end{cases} \tag{5}$$

where $\bar{d}[i]$ denotes the expected distortion of the ith macroblock(MB) among N total MBs and R_t represents the limit on the total bit rate imposed by the available bandwidth. This problem can be solved using the standard Lagrangian approach as follows

$$L=\bar{D}+\lambda\sum_{i=1}^{N}(R_L[i]+R_R[i]+R'_L[i]+R'_R[i]) \tag{6}$$

where λ is the Lagrangian multiplier. Because the two descriptions are symmetric, we will take description 1 as an example. In description 1, the left view and right view are treated as the base view and enhancement view, respectively, whose bit rates are R_L and R'_R, respectively. Using formula (1) and imposing $\nabla L=0$, we obtain

$$\frac{\partial L}{\partial R_{L,i}}=(1-p)^2\left(\frac{\partial D_{L,i}}{\partial R_{L,i}}+\frac{\partial D_{V,i}}{\partial R_{L,i}}\right)+p(1-p)\left(\frac{\partial D_{L,i}}{\partial R_{L,i}}+\frac{\partial D'_{R,i}}{\partial R_{L,i}}+\frac{\partial D_{LV,i}}{\partial R_{L,i}}\right)+\lambda=0 \tag{7}$$

$$\frac{\partial L}{\partial R'_{R,i}}=p(1-p)\left(\frac{\partial D'_{R,i}}{\partial R'_{R,i}}+\frac{\partial D_{LV,i}}{\partial R'_{R,i}}\right)+\lambda=0 \tag{8}$$

Here, $\frac{\partial D'_{R,i}}{\partial R_{L,i}}$ denotes the relationship between the distortion of the right view and the rate of the left view. Generally, a good left view will provide a good prediction of the right view, thereby resulting in a low distortion of the right view. To bridge $\frac{\partial D_{L,i}}{\partial R_{L,i}}$ and $\frac{\partial D'_{R,i}}{\partial R'_{R,i}}$ directly, we need to approximate $\frac{\partial D'_{R,i}}{\partial R_{L,i}}$ for the different types of regions. First, for the disoccluded regions, $\frac{\partial D'_{R,i}}{\partial R_{L,i}}=0$ because these regions cannot be predicted from the base view. Second, regarding the illumination-affected regions, these

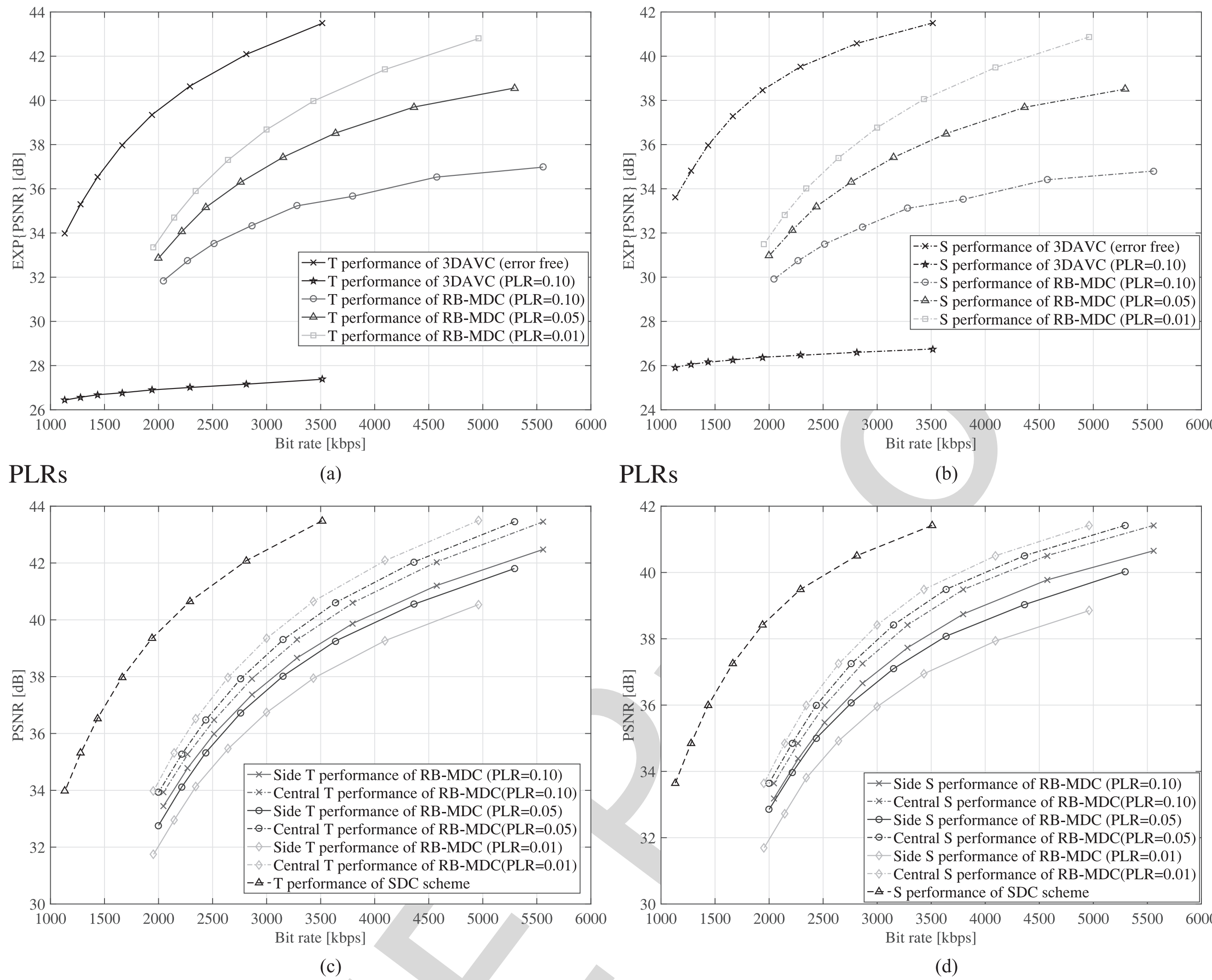

Fig. 11. The rate-PSNR performance of *Mobile*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

regions can be predicted, but not well; consequently the following approximation is used: $\frac{\partial D'_{R,i}}{\partial R_{L,i}} = 0.5\frac{\partial D_{L,i}}{\partial R_{L,i}}$. Finally, the remaining regions can be predicted very well; hence, we approximate $\frac{\partial D'_{R,i}}{\partial R_{L,i}} = 1.0\frac{\partial D_{L,i}}{\partial R_{L,i}}$. We note that the values 0, 0.5 and 1 coincide with the rendering mode parameter α and α_L, elaborated as follows. For the disoccluded regions, α and α_L should be zero since these regions exist only in the right enhancement view. For the illumination-affected regions, α and α_L should be 0.5 since these regions in both views have the same importance. For the remaining regions, since these regions can be rendered from the left view with sufficient good quality, α and α_L are set to 1. Therefore, equation (7) can be simplified as

$$\frac{\partial L}{\partial R_{L,i}} = (1-p)^2\left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D_{V,i}}{\partial R_{L,i}}\right) + p(1-p)\left((1+\alpha^2)\frac{\partial D_{L,i}}{\partial R_{L,i}} + \frac{\partial D_{LV,i}}{\partial R_{L,i}}\right) + \lambda = 0 \tag{9}$$

By substituting both (2) and (4) into (9) and (8), we obtain

$$\frac{\partial L}{\partial R_{L,i}} = (1-p)^2\left(\frac{\partial D_{L,i}}{\partial R_{L,i}} + \alpha\frac{\partial D_{L,i}}{\partial R_{L,i}}\right) + p(1-p)\left((1+\alpha^2)\frac{\partial D_{L,i}}{\partial R_{L,i}} + \alpha_L^2\frac{\partial D_{L,i}}{\partial R_{L,i}}\right) + \lambda = 0 \tag{10}$$

$$\frac{\partial L}{\partial R'_{R,i}} = p(1-p)(\frac{\partial D'_{R,i}}{\partial R'_{R,i}} + (1-\alpha_L)^2\frac{\partial D'_{R,i}}{\partial R'_{R,i}}) + \lambda = 0 \tag{11}$$

By combining (10) and (11), we can obtain the rate-distortion function describing the relationship between the base view and the enhancement view,

$$\left((1+\alpha^2)(1-p) + (1+\alpha^2+\alpha_L^2)p\right)\frac{\partial D_{L,i}}{\partial R_{L,i}} = p(2-\alpha_L^2)\frac{\partial D'_{R,i}}{\partial R'_{R,i}} \tag{12}$$

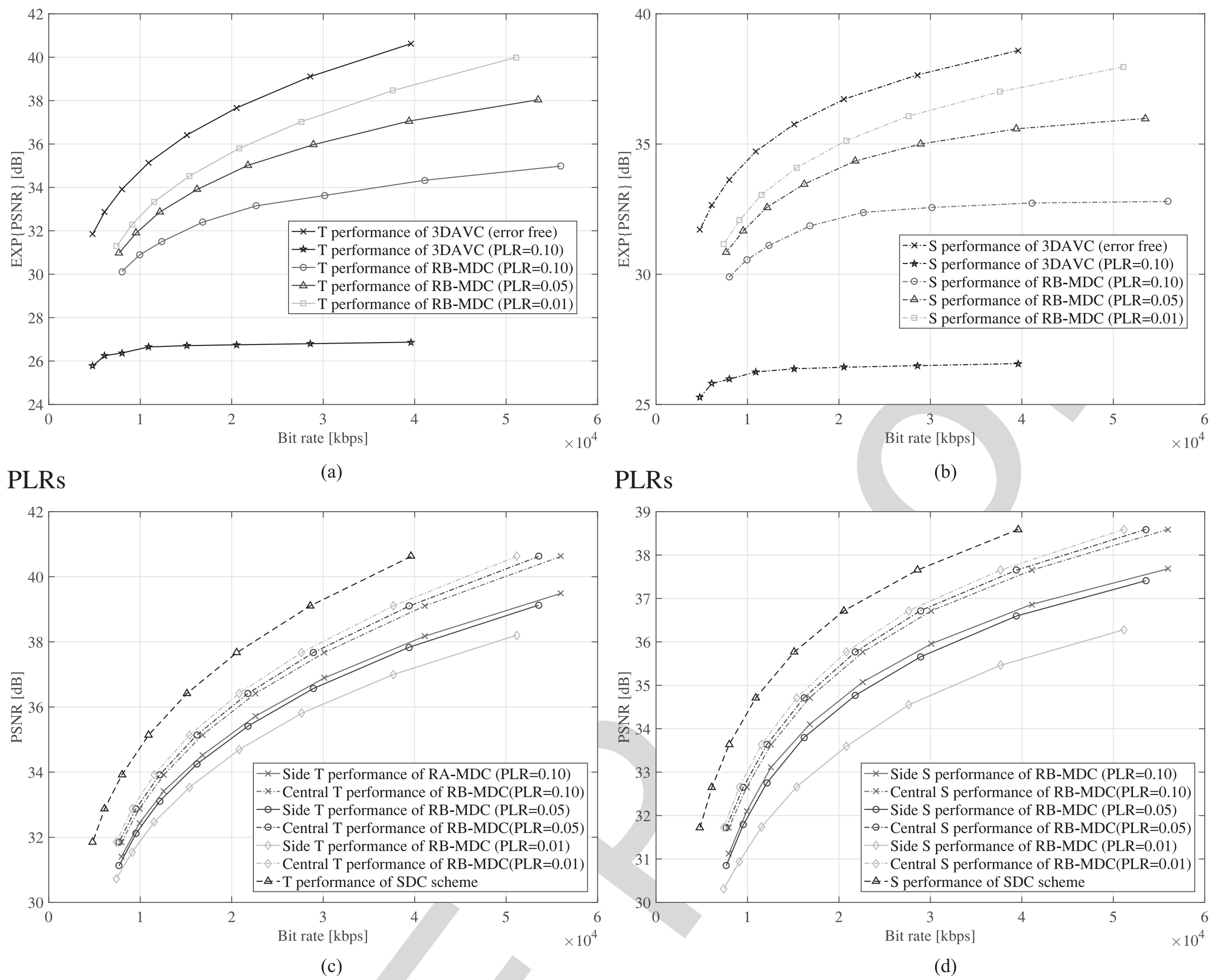

Fig. 12. The rate-PSNR performance of *Undodancer*. (a) Expected texture video performance at different PLRs. (b) Expected syntheisized view performance at different PLRs. (c) Side/central performance of texture videos. (d) Side/central performance of synthesized views.

3) QP Relationship: To obtain the relationship between the quantization parameters (QPs) of the two views, the standard H.264/AVC rate-distortion function is employed as follows

$$\frac{\partial D}{\partial R} = -0.85 * 2^{\frac{QP-12}{3}} \tag{13}$$

By inserting (13) into (12), we can obtain the QP values for the left base view and the right enhancement view:

$$QP'_R = QP_L + 3 * log2 \times \frac{\left((1+\alpha^2)(1-p) + (1+\alpha^2+\alpha_L^2)p\right)}{p(2-\alpha_L^2)} \tag{14}$$

where QP_L and QP'_R are the quantization parameters of the left base view and the right enhancement view, respectively. It can be observed that QP'_R depends on the packet loss rate (PLR) p and on the weight parameters α and α_L. Fig. 5 shows the relationship of $\Delta QP = QP'_R - QP_L$ with PLR, α and α_L. Here, QP_L is set to 21; therefore ΔQP is no larger than 30. We can observe that the larger the values of α and α_L are, the higher ΔQP will be. Moreover, the lower the value of p is, the higher ΔQP will be. These two trends are intuitive. The entire rate-distortion function also applies to the quantization of the depth maps.

The process of determining QP for description 2 is similar to that for description 1. With the assigned QP of the base view, we can calculate the QPs for each region under different network status using (14); thus a redundancy allocation formula is obtained. Note that QP is assigned on the macroblock(MB) level. However, some MBs are likely to contain both disoccluded pixels and general pixels.Hence, we need to calculate the ratios representing the proportions of an MB that are occupied by pixels of each different type. These ratios can be included in the QP calculation because different types of regions have different α and α_L values.

III. Experimental Results and Analysis

In this section, experiments are conducted using the following video sequences:*Newspaper* (1024 × 768) [19], *Lovebird* (1024 × 768) [20], *Balloons* (1024 × 768) [17], *Kendo* (1024 × 768) [17], *BookArrival* (1024 × 768) [21], *Mobile* (720 × 540) [18] and *UndoDancer* (1920 × 1088) [18]. The depth information is estimated versions for *Newspaper*, *Lovebird*, *Balloons*, *Kendo*

TABLE I
DISOCCLUDED RATIO OF EACH SEQUENCE

Sequence	Newspaper	Lovebird	Balloons	Kendo	Bookarrival	Mobile	Undodancer
Resolution	1024×768	1024×768	1024×768	1024×768	1024×768	720×540	1920×1088
disocclusion Ratio	0.088	0.0145	0.044	0.027	0.062	0.039	0.021

and *BookArrival*, whereas computer-generated (CG) depth is used for *Mobile* and *UndoDancer*. The description for the sequences can be found in [22]. For each sequence, two texture plus two depth videos are encoded with 3D-AVC [23] [24] to generate one description. The virtual views are synthesized using view synthesis reference software VSRS-1D-fast due to its fast and good performance [25], [26]. In detail, view 4 and view 6 of *Newspaper* were used to synthesize virtual view 5. View 6 and view 6 of *Lovebird* were used to synthesize virtual view 7. View 1 and view 3 of *Balloons* were used to synthesize virtual view 2. View 1 and view 3 of *Kendo* were used to synthesize virtual view 2. View 8 and view 10 of *BookArrival* were used to synthesize virtual view 9. View 4 and view 6 of *UndoDancer* were used to synthesize virtual view 5. View 1 and view 5 of *UndoDancer* were used to synthesize virtual view 3. The distortions of the virtual views were calculated between the virtual view images synthesized from the original texture plus depth videos and those synthesized from the decoded texture plus depth videos.

The described algorithm was implemented in the 3D-AVC reference software [27], and the important parameters are detailed in the following. The QP values for the base views were chosen from within a range of [22: 36] in step 2 to consider different rate-distortion points, whereas the QP' values for the enhancement views were determined using equation (14). The threshold for the illumination-affected regions is set as JND value frame by frame. Notice probably larger gain can be achieved if this threshold is set frame by frame according to video contents. However, high computation should be introduced to get this threshold. For description 1, the left view and right view were treated as the base view and enhancement view, respectively. The opposite view allocation was applied in description 2. Ultimately, the QP' values lay in the range $[QP_P, 51]$. The IPPP coding structure was used throughout the entire experiment and each row of MBs in each frame was encoded in one slice, which was then carried in one transport packet. This entire configuration was chosen to be similar to that used in the rate-distortion-optimized mode switching method [28] to facilitate a comparison of the results.

All experiments were performed in two parts: one to investigate the expected performance and one to investigate the side/central performance at different PLRs. For each part, the results for the left/right views and synthesized virtual views are presented separately. Here, the bit rate includes both descriptions (textures plus depth maps) used in our scheme. For the expected performance assessment, the Bernoulli channel model was adopted, and the performance was measured in terms of the average luminance peak signal-to-noise ratio (PSNR) obtained in 50 independent transmission trials. Side/central curves are presented to represent the performance for the case in which only one channel is working or both channels are working, where the side performance is measured as the average of the two side distortions. Three different packet loss rates of 10%, 5%, and 1% were selected for testing. Error-free results of single description coding are also presented for comparison.

Since the MVD coding structure is still new, few MDC schemes for this format have been introduced. However, several efficient error-resilient algorithms have been proposed for this format and thus can be considered for comparison here [28], [29], [30]. In [28], a rate-distortion-optimized mode switching (RDOMS) scheme was proposed that attempts to optimize the mode decision process considering the end-to-end distortion for error-resilient MVD. Bruno Macchiavello et al. have proposed a loss-resilient coding technique for free-view point videos [30]. The results of these two schemes on the *Newspaper* and *Lovebird* sequences are also reported here. To save room in the figures, our region-based multiple description scheme is abbreviated as RB-MDC, whereas T and S are used to represent a texture view and a synthesized view, respectively. To quantify the impairment caused by the introduced redundancy, we also include the results of the single description scheme (SDC), that is, the results of the classical 3D-AVC method.

In Subfigure a) and Subfigure b) of Figs. 6 and 7, the expected performances for the left/right views and synthesized views, respectively, are presented. It can be observed that our scheme is considerably superior to RDOMS [28], with gains of up to 2 dB on both the texture images and the synthesized views. However, when the bit rate is lower, the gains are relative smaller since our scheme is much more effective at normal and high bir tate cases. Note that the results of RDOMS and Bruno Macchiavello's scheme were obtained from [28]. There are many configuration parameters that can be modified during encoding, and we tried our best to make the configuration as similar as possible to that used in [28]. Furthermore, RDOMS only optimizes the mode selection, whereas our approach introduces MDC. Consequently, it is not truly fair to compare these schemes with ours since MDC has an advantage when the packet loss rate is high. However, this 2 dB gain still demonstrates the effectiveness of the proposed scheme.

In Figs. 6 and 7, the curves for the single description scheme in the error-free case are also included. We note that the gap between the error-free case and the proposed method is small when the bit rate is high because our bit allocation strategy can achieve better performance at higher bit rates. When the bit rate is lower, the three curves at the different PLRs tend to be very close. This is mainly because of the higher QP for a lower bit rate. On the one hand, a large QP will cause many macroblocks to be processed in skip mode, meaning that our bit allocation

IEEE Proof

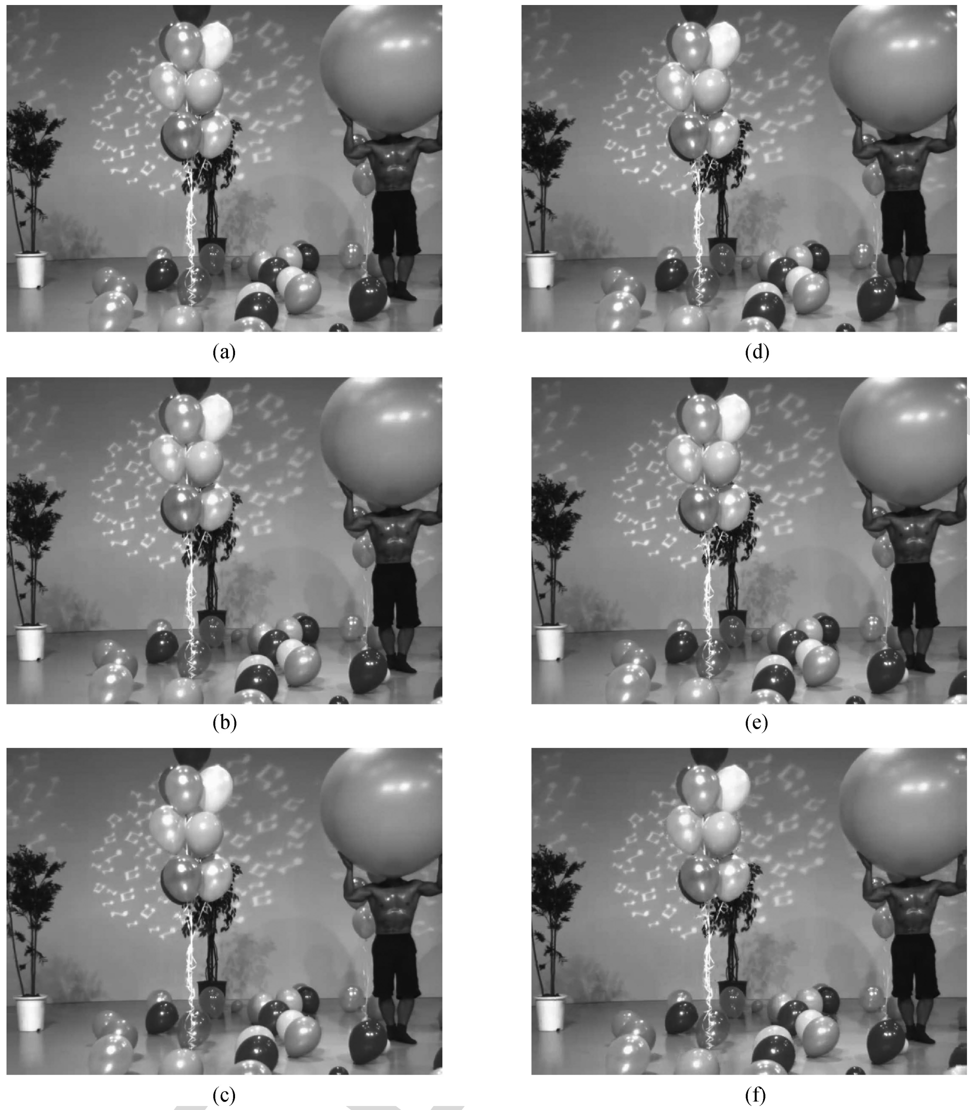

Fig. 13. Subjective visual results for *Balloons*

based on ΔQP will not work. On the other hand, with a large QP, we have less freedom to tune ΔQP because $QP + \Delta QP$ cannot be larger than 51 according to the H.264/AVC standard. For comparison, the results for the SDC scheme with $PLR = 0.10$ and $PLR = 0$ are also included. It can be observed that the proposed scheme is far superior to the SDC scheme in the presence of packet loss, in terms of both texture video performance and synthesized view performance. Moreover, packet loss affects the synthesized view performance more than the texture video performance since the synthesis depends on both the texture and depth images.

Subfigure (c) and Subfigure (d) of Figs. 6 and 7 present the side/central performances for the texture views and the synthesized views, respectively. Here, the performance of left and right views are averaged to provide that of texture view. We can observe that different trade-offs between side and central performance can be achieved under different channel statuses. In addition, the packet loss rate affects the introduced redundancy; a higher PLR corresponds to a higher redundancy. For example, the best side performance is achieved for a high PLR (0.10), whereas the best central performance is observed at a low PLR (0.01). We can determine the additional bit-rate cost for our central description that is required to achieve the same PSNR as that in the error-free SDC case, which is equivalent to the introduced redundancy. The different side description curves represent the performances achieved with different redundancy allocations when only one channel is working. We find that all performances are acceptable, even when one channel is completely nonfunctional. Note that the gain originates from our effective bit-rate allocation strategy for both the color videos and the depth maps.

Figs. 8–12 present the rate-PSNR performances on the *Balloons*, *Kendo*, *BookArrival*, *Mobile* and *UndoDancer* video sequences. These results confirm that the proposed technique exhibits good behavior regardless of the video content and resolution. Note that we treat holes as disoccluded regions. Hence, for depth maps that contain excessive noise, many holes or disocclusion regions will be generated and the efficiency of the proposed scheme will be affected. In the extreme case in which there are no holes or disoccluded regions, our scheme can achieve the maximum bit-rate savings and is the most effective. In the contrast, if the baseline it too large, many large holes will be generated. Our scheme will cost too many bits to deal with this situation. However, general baseline are not too larger, otherwise we cannot get a good 3D feeling. The disocclusion ratio for each sequence is listed in Table I. For example, *Balloons* contains relatively few disoccluded and illumination-affected regions; thus, its total redundancy is relatively low, and its expected performance is the best among all three sequences with the same resolution. *Newspaper* contains relatively more disoccluded and illumination-affected regions, and consequently, its expected performance is relatively worse. As for *Mobile* and *UndoDancer* with CG depth map, it is not fair to compare these sequences with the other four sequences since they have different resolution and bit-rate ranges. In fact, depth map of *UndoDancer* and *UndoDancer* have few disoccluded and illumination-affected regions, without any noise in depth maps; hence, for a good fixed central performance, its side performance and central performance are relative closer compared with the results of other sequences, due to its low introduced redundancy.

In addition to objective results, some subjective results are provided in Fig. 13. Here, the 10th frame of *Balloon* in view 1, together with the 10th frame in its corresponding synthesized view, are selected to demonstrate the performance. Our RB-MDC are configured at packet loss rate 5%. In order to evaluate the performance, the results of single description (SDC) case are included, in which the total bit rates of our MDC scheme and that of SDC are tuned to be similar as 5000 kbps. Since there are redundancy inserted in RB-MDC scheme, the results of ours is at disadvantage compared with that of SDC at error free case. In fact, there are some distortion around the balloons and trees, however, we cannot notice big visual difference between ours and that of SDC. Particularly, the side visual results that supposes one description is broken down are also very good, which demonstrate the efficiency of our scheme.

IV. Conclusion

In this paper, a region-based multiple description coding scheme for multiview video plus depth is proposed. First, regions are classified into disoccluded, illumination-affected and remaining regions according to their contributions to the virtual view to be synthesized. Second, an optimized expected rate-distortion function is designed based on both the texture video distortions and synthesized view distortions. By assigning different quantization parameters to the three types of regions depending on the channel status, we can minimize the expected distortion. Compared with traditional error-resilient 3D-AVC schemes, the proposed scheme can achieve gains of up to 2 dB in the case of packet loss. In addition, different prioritizations between side and central performance can be applied under different channel conditions, which is a desirable feature of any MDC scheme. An analysis of the experimental results shows that the proposed MDC scheme is a promising approach for the transmission of MVD-format 3D videos over error-prone channels.

It should be noted that our scheme achieves much better performance when the bit rate is higher. This is because our rate allocation strategy is more accurate at higher bit rates by virtue of the larger possible range of ΔQP. In addition, the performance of our scheme is also affected by the quality of the depth maps. If noise is present in the depth maps, such as noise due to depth estimation, irregular holes will be generated and the coding efficiency will consequently deteriorate. Hence, depth maps acquired via time-of-flight sensors must be subjected to noise reduction processing, which may be investigated in our further work.

References

[1] A. Vetro, A. Tourapis, K. Muller, and T. Chen, "3D-TV content storage and transmission," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 384–394, Jun. 2011.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, Sep. 2007, pp. 201–204.

[3] J. Y. Lee *et al.*, "Depth-based texture coding in AVC-Compatible 3D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1347–1361, Aug. 2015.

[4] F. Shao, G. Jiang, W. Lin, M. Yu, and Q. Dai, "Joint bit allocation and rate control for coding multi-view video plus depth based 3D video," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1843–1854, Dec. 2013.

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

[6] C. Zhu, S. Li, J. Zheng, Y. Gao, and L. Yu, "Texture-aware depth prediction in 3D video coding," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 482–486, Jun. 2016.

[7] V. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74 –93, Sep. 2001.

[8] Y. Wang, A. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proc. IEEE*, vol. 93, no. 1, pp. 57–70, Jan. 2005.

[9] C. Lin, T. Tillo, Y. Zhao, and B. Jeon, "Multiple description coding for H.264/AVC with redundancy allocation at macro block level," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 589–600, May 2011.

[10] H. Karim, A. Sali, S. Worrall, A. Sadka, and A. Kondoz, "Multiple description video coding for stereoscopic 3D," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2048–2056, Nov. 2009.

[11] A. Norkin *et al.*, "Schemes for multiple description coding of stereoscopic video," in *Proc. Int. Conf. Multimedia Content Representation, Classification Security*, 2006, pp. 730–737.

[12] X. Wang and C. Cai, "Mode duplication based multiview multiple description video coding," in *Proc. Data Compression Conf.*, Mar. 2013, pp. 527–527.

[13] C. Lin, Y. Zhao, T. Tillo, and J. Xiao, "Multiple description coding for stereoscopic videos with stagger frame order," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 1016–1025, Jun. 2015.

[14] J. Guo, H. Bai, C. Lin, M. Zhang, and Y. Zhao, "Intra-/inter-view correlation based multiple description coding for multiview transmission," in *Proc. Data Compression Conf.*, Apr. 2015, pp. 446–446.

[15] J. Jin, A. Wang, Y. Zhao, C. Lin, and B. Zeng, "Region-aware 3-D warping for DIBR," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 953–966, Jun. 2016.

[16] J. Wu *et al.*, "Enhanced just noticeable difference model for images with pattern complexity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2682–2693, Jun. 2017.

[17] Nagoya University, "3DV test sequences." [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/ mpegftv/mpeg3dv/CfP/ Q1

[18] Nokia, "3DV test sequences." [Online]. Available: ftp://mpeg3dv.research.nokia.com

[19] Gwangju Institute of Science and Technology, "3DV test sequences." [Online]. Available: ftp://203.253.128.142

IEEE Proof

[20] Electronics and Telecommunications Research Institute, "3DV test sequences." [Online]. Available: ftp://203.253.128.142

[21] Fraunhofer, "3DV test sequences." [Online]. Available: http://sp.cs.tut.fi/mobile3dtv/stereo-video/

[22] *Call for Proposals on 3D Video Coding Technology*, International Organization for Standardization, ISO/IEC JTC1/SC29/WG11 mpeg2011/N12036, 2011.

[23] B. T. Oh and K. J. Oh, "View synthesis distortion estimation for AVC- and HEVC-compatible 3-D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 1006–1015, Jun. 2014.

[24] D. Rusanovskyy, F.-C. Chen, L. Zhang, and T. Suzuki, "3DAVC test model 8," JCT-3V Document JCT3V-F1003m, Nov. 2013.

[25] Y. Chen, G. Tech, K. Wegner, and S. Yea, "Test model 8 of 3D-HEVC and MV-HEVC," JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC1/SC 29/WG 11, Doc. JCT3V-H1003, 2014.

[26] D.-Y. Kim, W.-S. Jang, and Y.-S. Ho, "Analysis of View Synthesis Methods (VSRS 1D fast and VSRS3.5)," Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, 2012.

[27] "3D-AVC reference software ATM 13.1." [Online]. Available: http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/3DV-ATMv1 3.1

[28] P. Gao and W. Xiang, "Rate-distortion optimized mode switching for error-resilient multi-view video plus depth based 3-D video coding," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1797–1808, Nov. 2014.

[29] A. Vosoughi, V. Testoni, P. Cosman, and L. Milstein, "Joint source-channel coding of 3D video using multiview coding," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 2050–2054.

[30] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, and W. T. Tan, "Loss-resilient coding of texture and depth for free-viewpoint video conferencing," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 711–725, Apr. 2014.

Authors' photographs and biographies not available at the time of publication.

IEEE Proof

QUERIES

Q1. Author: Please provide the years in Refs. [17]–[21] and [27].

IEEE Proof