IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2023, accepted 15 January 2024, date of publication 18 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355749

RESEARCH ARTICLE

Efficient Reversible Data Hiding Based on View Synthesis Prediction for Multiview Depth Maps

JIN YOUNG LEE

Department of Intelligent Mechatronics Engineering, Sejong University, Seoul 05006, South Korea

e-mail: jinyounglee@sejong.ac.kr

This work was supported in part by the National Research Foundation of Korea (NRF) funded by the Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2021R1C1C1006459 and Grant RS-2023-00219051, and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (MSIT) under Grant RS-2022-00156345.

ABSTRACT Multiview-plus-depth (MVD), which has been used as a 3D format, consists of a texture image and its corresponding depth map at each viewpoint. This MVD-based 3D format has two strong correlations, which are an inter-component correlation between a texture image and its corresponding depth map and an interview correlation between adjacent texture and depth views. However, conventional reversible data hiding (RDH) methods have been mainly developed for texture images. As a result, high performance could not be achieved in depth maps. In order to solve this problem, an efficient RDH method is proposed for multiview depth maps in this paper. The proposed RDH method checks inter-component and interview correlations, and uses inter-component and interview predictions adaptively to embed secret data in the depth map. In particular, a view synthesis prediction (VSP) method is used for the interview prediction. In addition, an allowable depth distortion range, which guarantees no synthesis distortion in virtual views, is calculated to minimize distortion of the depth map marked with the hidden data, while maintaining the high embedding capacity. Experimental results show that the proposed method achieves much higher performance than conventional methods in terms of the embedding capacity and the depth distortion.

INDEX TERMS Depth map, multiview-plus-depth (MVD), reversible data hiding (RDH), view synthesis prediction (VSP).

I. INTRODUCTION

Multimedia information security and privacy are currently very important, because a lot of personal data is exchanged in communication systems and it can be easily accessible from third parties [1]. In order to tackle this problem, many solutions for the multimedia communication systems have been studied. One of the solutions, reversible data hiding (RDH), inserts important secret data in a cover image or a host document. For instance, doctor diagnosis with patient confidential details can be embedded into a medical image for additional information storage. Payment information and digital signature can be also inserted into a document to protect personal service data against unauthorized parties. Since RDH can recover an original cover without distortion, it is widely used in unlawful alteration copyright protection, integrity authentication, and unauthorized distribution.

In general, when considering an image, RDH embeds secret data into the cover image, and the image marked with the hidden data does not change visually. When the hidden data is extracted, the cover image can be losslessly recovered. In this mechanism, conventional RDH methods have been mainly developed for a texture image [2], [3], [4], [5], [6], [7], [8], [9], [10]. For instance, difference expansion (DS) methods embed the secret data by expanding a difference between neighboring pixels. Visual quality of the image marked with the hidden data depends on the difference magnitude. Histogram shifting (HS) methods embed the secret information by modifying both peak and zero points in the histogram. Finally, prediction error expansion (PEE) methods calculate

The associate editor coordinating the review of this manuscript and approving it for publication was Xinfeng Zhang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

prediction error of the pixel to be embedded by using neighboring pixels, and embed the secret data by expanding the error. For performance improvement, many pixel-value-ordering methods have been actively tested recently [11], [12], [13], [14], [15]. However, because they were developed and optimized only for texture images, new techniques should be considered for other image formats.

Following recent advances in multiview video technologies, the natural 3D scenes can be vividly perceived in living room environments now. However, it is very difficult to send all multiview information for realistic 3D perception, because of transmission bandwidth limitations. In order to significantly reduce the amount of data transmitted, multiview-plus-depth (MVD), which consists of a texture image and its depth map, was introduced as a 3D video format [16]. Fig. 1 represents a texture image and its depth map for the *GT_Fly* MVD image. The texture image refers to brightness of an object, while the depth map is a distance between an object and a camera. Pixels with 255 and 0 mean the nearest depth and the farthest depth from the camera, respectively. As depicted in Fig. 1, the depth map contains sharp edges between objects and background. In addition, it has many homogenous areas. The main role of the depth map is to synthesize virtual texture views, and it is not shown directly through advanced displays. However, a small synthesis distortion in depth maps, such as compression error, may generate significantly annoying artifacts around the sharp edges in the virtual views.

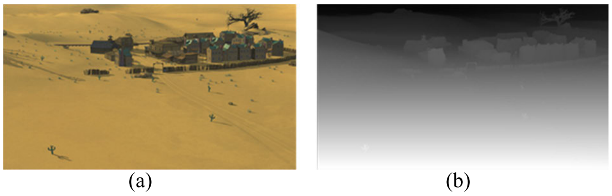


FIGURE 1. MVD format consisting of (a) a texture image and (b) its corresponding depth map for the GT_Fly MVD image.

Thanks to efficient 3D representation of the MVD format, a small amount of multiview information can be transmitted. Fig. 2 shows a 3D image communication system, when three MVD views are used. For example, a sender compresses and transmits three MVD views of v_0, v_1, and v_2 with exiting 3D compression standards, such as 3D-AVC [17] and 3D-HEVC [18]. v_n is MVD at each viewpoint. A receiver decompresses them and synthesizes virtual views at arbitrary viewpoints by using depth image based rendering (DIBR) [19]. The number of output views m is usually greater than the number of input MVD views n. As shown in Fig. 2, if n and m are respectively equal to 3 and 9, the receiver has the output views consisting of three decoded views (solid) and six synthesized virtual views (dashed).

With the advent of 3D, MVD-based RDH methods have emerged. Because the use of PEE methods results in better performance than the use of the other methods, MVD-based methods find pixel positions having small prediction error in depth maps, and hide secret data in there by using a depth no-synthesis-error (D-NOSE) model [20], which introduces an allowable depth distortion range that avoids the synthesis distortion in virtual views and determines the extent to which secret data can be embedded in each pixel. The basic idea of the D-NOSE model has been utilized to enhance 3D coding performance [21], [22]. Similarly, many MVD-based methods applied it into RDH to increase the capacity. For example, both Chung et al.'s method [23] and Shi et al.'s method [24] estimate the depth pixels to be embedded by simply using an average prediction method based on the spatial correlation between neighboring pixels. As shown in Fig. 1, since many values gradually change depending on the distance from the camera, the spatially average prediction is not efficient. Lee et al.'s method [25] proposes an inter-component prediction from a texture image to its depth map for accurate prediction, based on the inter-component correlation between texture and depth images. However, as the MVD views also contain the very high interview correlation between adjacent views, there is still room to improve the prediction accuracy.

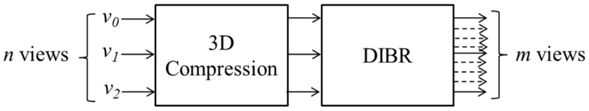


FIGURE 2. 3D image communication system using three MVD views.

This paper introduces a high-capacity and low-distortion RDH method for the multiview depth maps in the 3D image communication systems. The proposed RDH method firstly improves prediction accuracy by utilizing inter-component and interview predictions adaptively to achieve high capacity, based on high inter-component and interview correlations of MVD views. It directly uses the inter-component prediction proposed in Lee et al.'s method [25], whereas the interview prediction performs view synthesis prediction (VSP), which has not been used in RDH before. In addition, the proposed method introduces the allowable depth distortion range to achieve lower distortion, based on prediction results, while maintaining the high embedding capacity. The experimental results show that it achieves much higher capacity and lower distortion than the conventional methods.

The remainder of this paper is organized as follows. Section II explains the conventional RDH methods and their problems. In Section III, the proposed method employing the VSP-based interview prediction and the allowable depth distortion range is described in detail. Section IV shows experimental results in terms of the embedding capacity and the depth distortion. Finally, the paper is concluded in Section V.

II. RELATED WORK

Since some depth pixels are converted into the same disparity in DIBR, the D-NOSE model defines the allowable distortion range, which ensures that every depth pixel within this

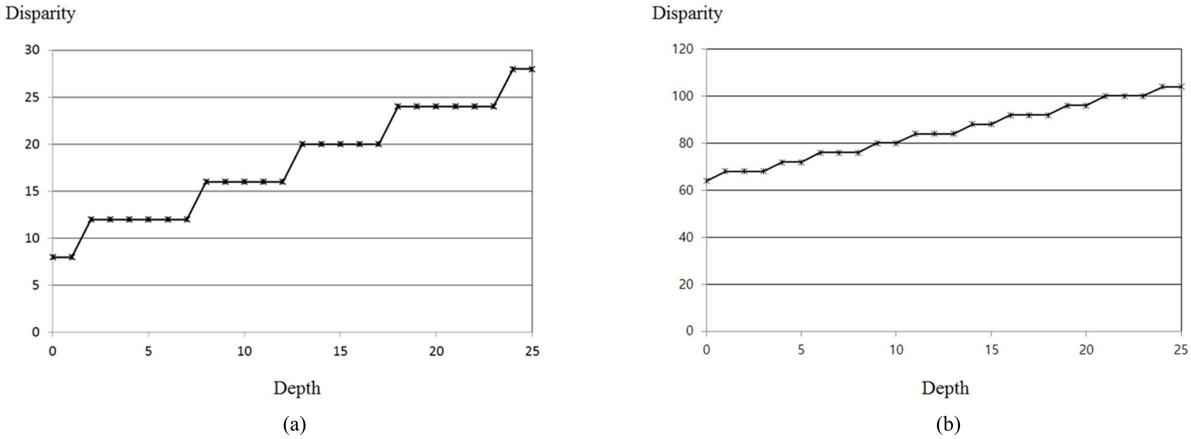


FIGURE 3. Relationship between depth pixels and their corresponding disparities for the (a) Balloons and (b) Poznan_Hall2 MVD images.

range produces the same result in the generation of virtual views. Fig. 3 shows a relationship between depth pixel values (ranging from 0 to 25) and their associated disparities for the *Balloons* and *Poznan_Hall2* MVD images. As shown in Fig. 3, many depth values are mapped into the same disparity. For instance, every depth pixel in [2, 7] has the disparity of 12 in the *Balloons* MVD image. This means that the virtual views will not be distorted, although some values in [2, 7] are modified or distorted to the other values in [2, 7]. Based on this D-NOSE model, the conventional MVD-based RDH methods hide the secret information in depth maps, without any degradation in quality of the virtual views.

A. CHUNG et al.'s METHOD

In Chung et al.'s method [23], the secret data is inserted in a depth pixel where prediction error is equal to 0 or −1. The prediction error e is calculated, as follows,

$$e = c - p \tag{1}$$

where c and p denote a current depth pixel to be embedded and its predicted pixel, respectively, and the prediction is performed by averaging four adjacent pixels, as follows,

$$p = \left\lceil \frac{c(j-1,i)+c(j,i+1)+c(j+1,i)+c(j,i-1)}{4} \right\rceil \tag{2}$$

where *j* and *i* denote a coordinate within a depth map. *c(j-1, i)*, *c(j, i+1)*, *c(j+1, i)*, and *c(j, i–1)* mean above, right, bottom, and left pixels around the current pixel *c(j, i)*, respectively. Since the D-NOSE model is used, original and marked pixels belong to the allowable distortion range $[L_d, U_d]$ with the same disparity d. The marked pixel m is obtained, as follows,

$$m = \begin{cases} p+h, & \textit{if } e=0 \\ p-h-1, & \textit{if } e=-1 \end{cases} \tag{3}$$

where h represents a hidden data value between 0 and 2^b–1, and the maximum number of hidden bits b is calculated, as follows,

$$b = \begin{cases} \lfloor \log_2(U_d-p+1) \rfloor, & \textit{if } e=0 \\ \lfloor \log_2(p-L_d+1) \rfloor - 1, & \textit{if } e=-1 \\ 0, & \textit{otherwise} \end{cases} \tag{4}$$

A location map records the marked pixels where the secret data are hidden. This is losslessly compressed by using the arithmetic coding and inserted with the hidden data. The location map and the hidden data can be extracted in an inverse manner of eq. (3). The current pixel c to be recovered is calculated, as follows,

$$c = \begin{cases} p\,, & \textit{if } p \le m \le U_d \\ p-1\,, & \textit{if } L_d \le m \le p-1 \end{cases} \tag{5}$$

These embedding and extracting are performed without any synthesis distortion. For example, if p and e are respectively 6 and 0, and $[L_d, U_d]$ is [2, 7], then b becomes 1, based on eq. (4). Hence, m can only become 6 or 7, based on eq. (3).

This method utilizes the allowable depth distortion range for the high embedding capacity in the depth maps. However, the range is not fully employed, because it is divided into two sub-ranges. Hence, further improvement could be achieved if the full range were to be applied.

B. SHI et al.'s METHOD

In Shi et al.'s method [24], when prediction error calculated by eq. (1) is equal to 0, secret data is only embedded into a depth pixel. To make full use of the allowable distortion range $[L_d, U_d]$, the marked pixel m is calculated by adding the secret data h directly to the low boundary L_d, as follows,

$$m = L_d + h \tag{6}$$

h has a value between 0 and 2^b-1, and the maximum number of hidden bits b is computed, as follows,

$$b = \begin{cases} \lfloor \log_2(U_d-L_d+1) \rfloor\,, & \textit{if } e=0 \\ 0, & \textit{otherwise} \end{cases} \tag{7}$$

In the extracting, h can be calculated by subtracting L_d from m, based on the location map that indicates if e is equal to 0 or not, as follows,

$$h = m - L_d \tag{8}$$

Because the embedding is performed at the pixels with zero prediction error, the current original pixel c can be directly recovered from the predicted pixel p. For instance, if p and e are 6 and 0, and $[L_d, U_d]$ is [2, 7], then b becomes 2, based on eq. (7). Therefore, m can be 2, 3, 4, or 5, based on eq. (6).

This method obtains the considerably higher capacity than does Chung et al.'s method, because it employs the allowable distortion range fully. However, these two methods only use the spatial correlation between neighboring pixels. Because MVD has the inter-component correlation and the interview correlation, more efficient prediction can be designed for the higher performance.

C. LEE et al.'s METHOD

Since a texture image and its depth map shows same scenes, their correlation is generally very high. In Lee et al.'s method [25], an inter-component prediction from a texture image into a depth map was introduced to improve the prediction performance with high accuracy. It predicts a current depth pixel *c(j, i)* with one pixel in a set of four adjacent pixels, which include *c(j-1, i)*, *c(j, i+1)*, *c(j+1, i)*, and *c(j, i-1)*, by comparing its corresponding texture pixel *t(j, i)* and adjacent pixels *t(j-1, i)*, *t(j, i+1)*, *t(j+1, i)*, and *t(j, i-1)*, as follows,

$$\begin{aligned} dir\,D = \min_{dirT}(&|t(j,i) - t(j-1,i)|, |t(j,i) - t(j,i+1)|, \\ &\times |t(j,i) - t(j+1,i)|, |t(j,i) - t(j,i-1)|) \end{aligned} \tag{9}$$

where *dirT* and *dirD* mean the best directions of the texture and depth predictions, respectively. If the difference between *t(j, i)* and *t(j+1, i)* is smallest in eq. (9), both *dirT* and *dirD* have the direction from the bottom pixel to the center pixel, as shown in Fig. 4. The predicted pixel p can be set into the adjacent pixel along the best direction *adj(dirD)*, as follows,

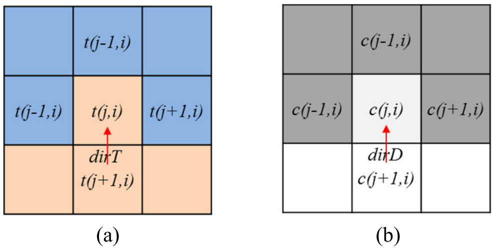
$$p = adj(dirD) \tag{10}$$

As a result, *adj(dirD)* becomes *c(j+1, i)* in Fig. 4. Finally, the data embedding can be performed with eqs. (6) and (7). Same as Shi et al.'s method, if p and e are 6 and 0, and $[L_d, U_d]$ is [2, 7], then b becomes 2 and m can become 2, 3, 4, or 5.

This method improves the embedding capacity with the inter-component prediction based on the strong correlation between texture and depth images. However, it still does not consider the interview correlation between adjacent views. In addition, the allowable depth distortion range $[L_d, U_d]$ in eq. (7) can be more carefully found so as to minimize the depth distortion while retaining the high capacity.

III. PROPOSED RDH METHOD

In this section, an efficient RDH method is proposed for the multiview depth maps. Since MVD includes the high inter-component correlation between texture and depth images as well as the high interview correlation between neighboring views, the proposed method adaptively performs the inter-component and interview prediction methods to achieve the high embedding capacity. Specifically, VSP is employed for the interview prediction. In addition, for the low distortion, the allowable depth distortion range is calculated while still ensuring high embedding capacity.



(a) (b)

FIGURE 4. Inter-component prediction from (a) a texture image to (b) its depth map.

A. VSP-BASED INTERVIEW PREDICTION

In general, virtual views are generated through the operation of view synthesis, which consists of projecting a reference view onto target viewpoints via 3D warping and hole-filling techniques. For instance, during the warping, a texture pixel in a reference image is moved into a virtual view by using its corresponding reference depth map. This involves two steps: First, a texture pixel $t(j_r, i_r)$ in a reference image is projected onto a 3D world coordinate *(u, v, w)*, and second, a resulting pixel *(u, v, w)* is projected onto a local image coordinate $t(j_t, i_t)$ of the target view. However, these steps can be simplified when operating under the 1D camera setting by finding the horizontally shifting disparity d, as follows,

$$d = \frac{f \cdot l}{255}\left(\frac{1}{z_{near}} - \frac{1}{z_{far}}\right) \cdot c(j_r, i_r) \tag{11}$$

where f and l mean the focal length and the baseline distance between neighboring cameras, respectively. z_{near} and z_{far} are the nearest and farthest depths, respectively. $c(j_r, i_r)$ indicates a reference pixel. It is assumed that the calculated disparity d is rounded into an integer value. Finally, the pixel $t(j_t, i_t)$ in the target virtual texture image is filled with the pixel $t(j_r, i_r)$ in the reference image, as follows,

$$t(j_t, i_t) = t(j_r \pm d, i_r) \tag{12}$$

Because 3D warping involves left and right directions, d has two signs. However, during the view synthesis process, some pixels in the occlusion regions become a hole, and cannot be mapped from the reference pixel. In order to solve this issue, advanced hole-filling techniques can be used to fill the hole.

Unlike the texture view synthesis above, we use the depth view synthesis to perform the interview prediction, which we refer to as a VSP-based interview prediction. Let us consider two neighboring views: center and right views. If secret data is embedded into the right view, the center and right views can be treated as the reference and target views, respectively, in the view synthesis. Therefore, when predicting the current pixel $c(j_t, i_t)$ in the right view, its predicted pixel $p(j_t, i_t)$ can be calculated from the reference pixel $c(j_r, i_r)$ in the center view by modifying eq. (12), as follows,

$$p(j_t, i_t) = c(j_r \pm d, i_r) \tag{13}$$

Fig. 5 represents an original depth map and a depth map generated from the VSP-based interview prediction for the *GT_Fly* MVD image. The hole-filling was not performed in the proposed method. Black regions, particularly as seen in the right of Fig. 5 (b), belong to the hole areas, which could not be mapped from the reference view. However, it can be seen that the predicted depth map is still very similar to the original depth map.

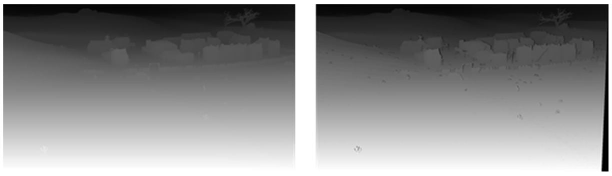

(a) (b)

FIGURE 5. (a) Original depth map and (b) the map predicted from the VSP-based interview prediction for the GT_Fly MVD image.

Fig. 6 depicts difference images generated from the inter-component and interview predictions for the *GT_Fly* MVD image. The black mark means the zero difference, while the white one is that the absolute difference is greater than 0. In this example, it can be observed that the interview correlation between the adjacent views is much stronger than the inter-component correlation. This is because multiview cameras capture the same scene with the same quality simultaneously at multiple viewpoints, resulting in the very strong interview correlation. The view synthesis process can generate target images from given reference images and camera parameters. Because MVD has texture and depth images captured from the cameras and their parameters, the VSP-based interview prediction method effectively leverages the high interview correlation between adjacent views.

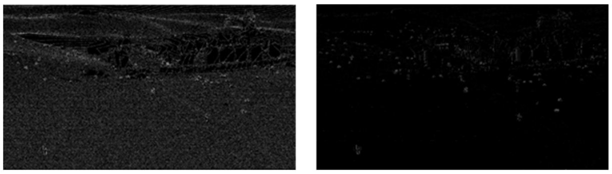

(a) (b)

FIGURE 6. Difference images generated by (a) the inter-component prediction and (b) the VSP-based interview prediction methods for the GT_Fly MVD image.

To improve the prediction accuracy, the proposed method adaptively uses the inter-component prediction and the VSP-based interview prediction. Both the inter-component and interview correlations are computed to select the best method in the inter-component and interview prediction methods at an image level. First, each prediction method generates its predicted depth map. If the difference between the original depth map and the depth map generated through the inter-component prediction of eq. (10) is smaller than that between the original depth map and the depth map predicted through the interview prediction of eq. (13), then the inter-component prediction method is utilized for the embedding. Otherwise, the VSP-based interview prediction is performed. Finally, a bit to indicate which prediction is utilized in the embedding should be recorded for each view. Hence, a sender embeds a prediction type bit *pred* with the hidden data. At a receiver, the best prediction method can be employed in the extracting, based on *pred*.

After selecting the best prediction method, the embedding process starts with the first pixel set and then the second pixel set, which consist of the pixels marked with O and X in Fig. 7, respectively. Hence, when the pixels in the second set are predicted, the pixels in the first set may be distorted with the hidden data. In the average and inter-component predictions, if all the four neighboring pixels around the current pixel are distorted, the prediction accuracy is affected. However, since VSP is always performed on undistorted views, the accuracy of the VSP-based interview prediction is drastically high. If the predicted pixel belongs to the hole, the spatially average prediction or inter-component prediction can be used for that pixel. It is also possible that the hole pixel is skipped without the prediction.

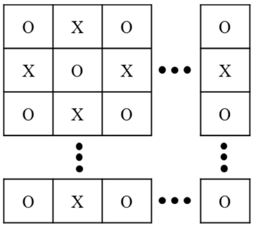


FIGURE 7. First (marked with O) and second (marked with X) pixel sets.

B. PREDICTION-BASED ALLOWABLE DISTORTION RANGE

D-NOSE introduces the allowable depth distortion range $[L_d, U_d]$, which has the same disparity d. For less distortion, the proposed method shifts the allowable distortion range closer to the predicted value p. First, when the prediction error is equal to 0, the number of hidden bits b is calculated, based on eq. (7). Second, if b is greater than 0, a prediction-based range $[L_p, U_p]$ is computed, as follows,

$$[L_p, U_p] = [p - 2^{b-1}, p + 2^{b-1} - 1] \tag{14}$$

Finally, if L_p is less than L_d, L_d^* in the proposed range $[L_d^*, U_d^*]$ is set to L_d, and U_d^* is computed by adding a difference between L_d and L_p into U_p, as follows,

$$[L_d^*, U_d^*] = [L_d, U_p + (L_d - L_p)] \tag{15}$$

If U_p is greater than U_d, U_d^* is set to U_d, and L_d^* is calculated by subtracting a difference between U_p and U_d from L_p, as follows,

$$[L_d^*, U_d^*] = [L_p - (U_p - U_d), U_d] \tag{16}$$

Because the proposed allowable depth distortion range $[L_d^*, U_d^*]$ is always closer to the predicted pixel than the original distortion range $[L_d, U_d]$, the proposed method can obtain the smallest average distortion.

Table 1 shows the embedding information of the methods of Chung et al., Shi et al., Lee et al., and the proposed RDH method when given texture and depth images, as shown in Fig. 8. The depth-to-disparity graph represents the allowable distortion range [2, 7] (red dashed line) of the current pixel (yellow) with the same disparity. The predicted value of all the methods is equal to 6. For instance, the inter-component prediction finds the best prediction direction *dirT* and *dirD* (red solid line) by comparing the corresponding texture pixel (green) and its four neighboring pixels, while the VSP-based interview prediction method uses a depth pixel (gray) shifted by the disparity d (black dashed line). However, the proposed method modifies the conventional allowable distortion range [2, 7] to [4, 7], based on eqs. (15) and (16). Since the number of hidden bits b is 2, the possible h becomes one among 0, 1, 2, and 3, and the marked depth pixel m can be 4, 5, 6, or 7, based on eq. (6). During extraction of h, the proposed method calculates the allowable distortion range in the same manner as during embedding, and then h is extracted from m, based on eq. (8).

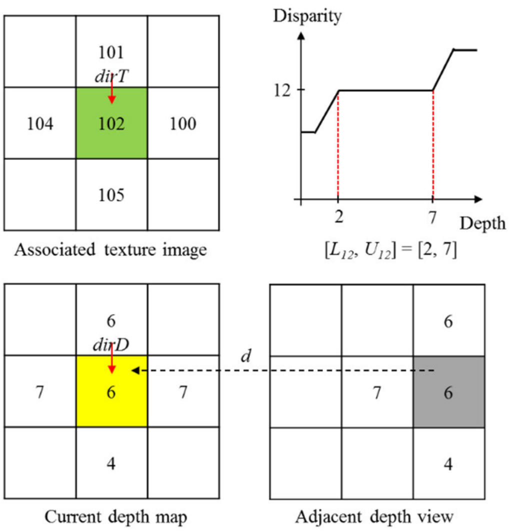


FIGURE 8. Example of texture and depth information.

C. EMBEDDING AND EXTRACTING PROCESSES

This section describes embedding and extracting processes. Some additional information, which is the important data in the extracting process, has to be inserted with the hidden data.

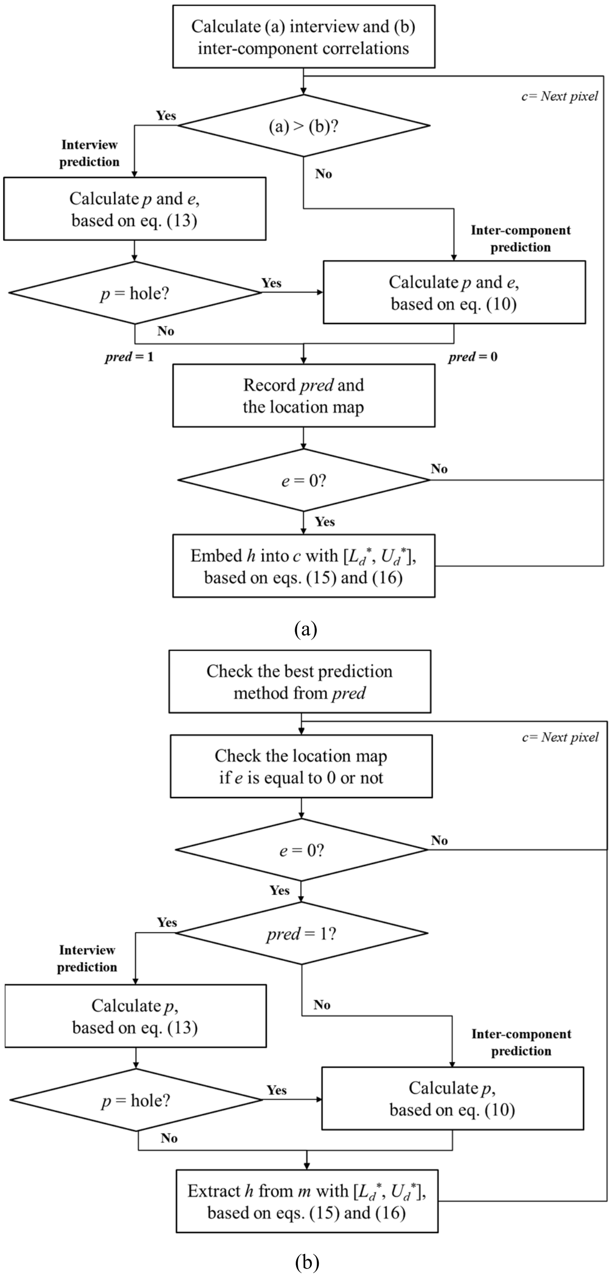

Calculate (a) interview and (b) inter-component correlations
c= Next pixel
Yes
(a) > (b)?
Interview prediction
No
Calculate p and e, based on eq. (13)
Inter-component prediction
p = hole?
Yes
Calculate p and e, based on eq. (10)
No
pred = 1
pred = 0
Record pred and the location map
e = 0?
No
Yes
Embed h into c with $[L_d^*, U_d^*]$, based on eqs. (15) and (16)
(a)
Check the best prediction method from pred
c= Next pixel
Check the location map if e is equal to 0 or not
e = 0?
No
Yes
Yes
pred = 1?
Interview prediction
No
Calculate p, based on eq. (13)
Inter-component prediction
p = hole?
Yes
Calculate p, based on eq. (10)
No
Extract h from m with $[L_d^*, U_d^*]$, based on eqs. (15) and (16)
(b)

FIGURE 9. (a) Embedding and (b) extracting processes of the proposed RDH method.

As with the conventional RDH methods, the location map is losslessly compressed through the arithmetic coding [26], and then extra information, such as the compressed location map, its size, and the prediction type bit, is embedded in the boundary of each depth map under the D-NOSE model. This information is used to losslessly extract hidden data from the marked depth maps and perfectly recover the original depth maps. Fig. 9 shows the embedding and extracting processes of the hidden data.

For example, when multiview depth maps are given with their texture images and camera parameters, the embedding process follows this steps:

TABLE 1. Data embedding information of the conventional methods and the proposed method under Fig. 8.

Method	b	$[L_d, D_d]$	Possible h	m
Chung et al.	1	[2, 7]	[0, 1]	$6 + h$
Shi et al.	2	[2, 7]	[0, 3]	$2 + h$
Lee et al.	2	[2, 7]	[0, 3]	$2 + h$
Proposed	2	[4, 7]	[0, 3]	$4 + h$

1. Predict the pixel to be embedded c with eqs. (10) or (13), based on the best prediction method.
2. Calculate the prediction error e with eq. (1), and then record the best prediction type bit *pred* for the best prediction method and the location map to indicate if e is equal to 0 or not.
3. Calculate the allowable depth distortion range $[L_d^*, U_d^*]$ with eqs. (15) and (16).
4. Embed the secret data to be hidden h into c with eq. (6), if e is equal to 0.
5. Return to the first step to process the next depth pixel.

The extra information is extracted from the boundary of the depth maps. The extraction of hidden data is performed in the similar way to its embedding, as per these steps:

1. Check the prediction type bit *pred* and the location map as to whether e is equal to 0 or not.
2. Predict the pixel to be recovered c with eqs. (10) or (13), based on *pred* only if e is equal to 0. Otherwise, go to the first step to process the next pixel.
3. Calculate the allowable depth distortion range $[L_d^*, U_d^*]$ with eqs. (15) and (16)
4. Extract h from the marked depth map m with eq. (8), and then recover c directly from p.
5. Return to the first step to process the next depth pixel.

IV. EXPERIMENTAL RESULT

Performance in terms of the embedding capacity and depth distortion was evaluated with eight MVD images with a size of 1024 × 768 and 1920 × 1088. These MVD images [27] were used to develop both the 3D-AVC [17] and 3D-HEVC [18] standards in JCT-3V. The embedding capacity was measured in bit per pixel (BPP), as follows,

$$BPP = \frac{\sum_{j=1}^{h}\sum_{i=1}^{w} b(j,i)}{h \times w} \tag{17}$$

where h and w denote a height and a width of each MVD image, respectively. $b(j, i)$ is the number of hidden bits at a coordinate (j, i). Hence, 1 BPP indicates the number of pixels in an image. The distortion was measured in terms of PSNR between the original and marked depth maps in dB.

A. TEST SCENARIO

Three MVD views were employed for the data embedding in the experiments, as displayed in Fig. 2. The center, left, and right views are denoted by v_0, v_1, and v_2, respectively. View numbers of each MVD image follows the JCT-3V condition [27]. Fig. 10 represents the prediction direction. Because the MVD format consists of a texture image and its associated depth map, and three MVD views, that is, v_0, v_1, and v_2 are input, the total number of images evaluated in this three-view configuration is 6, namely T_0, D_0, T_l, D_1, T_2, and D_2 in Fig. 10. T_0 and D_0 belong v_0, T_l and D_1 belong to v_1, and T_2 and D_2 belong to v_2. In the inter-component prediction, a depth pixel is predicted by comparing its corresponding texture pixel and four pixels along the direction (dashed line) from the texture image into the depth map. The interview prediction uses the center depth view as a reference view to predict the adjacent left and right views (solid line). It should be noted that the center depth view D_0 cannot employ the interview prediction because of the embedding and extracting orders. Therefore, D_0 is always processed with inter-component prediction.

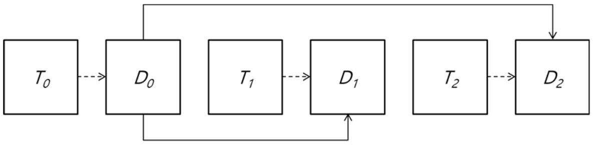


FIGURE 10. Prediction direction in the proposed RDH method.

In the proposed method, the compressing and embedding for three depth views are performed through these steps:

1. Compare the interview correlation between D_2 and D_0, and the inter-component correlation between D_2 and T_2, and select the best prediction type for D_2.
2. Embed h into D_2 by using D_0 for the interview prediction and T_2 for the inter-component prediction in an adaptive way, based on the correlation result.
3. Compare the interview correlation between D_1 and D_0, and the inter-component correlation between D_1 and T_1, and select the best prediction type for D_1.
4. Embed h into D_1 by employing D_0 for the interview prediction and T_1 for the inter-component prediction in an adaptive way, based on the correlation result.
5. Embed h into D_0 by using T_0 for the inter-component prediction.
6. Compress the original texture images, T_0, T_1, and T_2, the marked depth maps, D_0', D_1', and D_2' with 3D-HEVC.

The decompressing and extracting is performed, as follows:

1. Decompress and reconstruct T_0, T_1, T_2, D_0', D_1', and D_2' from the bitstream with 3D-HEVC.

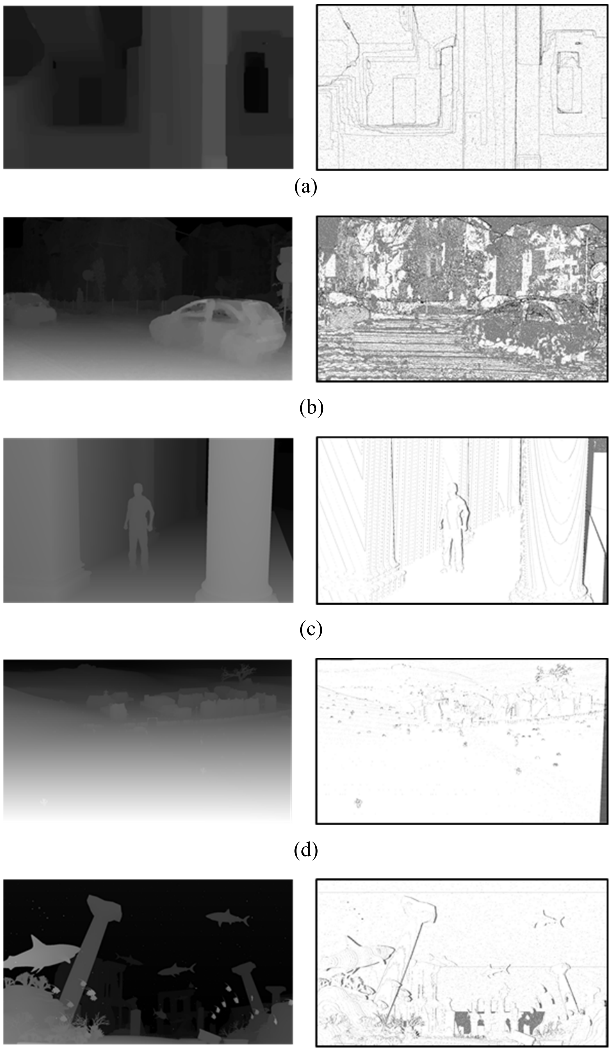
2. Extract h in D'_0 by using T_0 for the inter-component prediction, and then recover D_0.
3. Extract h in D'_1 by employing D_0 for the interview prediction and T_1 for the inter-component prediction in an adaptive way, based on the prediction type, and then recover D_1.
4. Extract h in D'_2 by employing D_0 for the interview prediction and T_2 for the inter-component prediction in an adaptive way, based on the prediction type, and then recover D_2.

Since the embedding order is the right, left, and then center views, and the extracting order is the center, left, and right views, the VSP-based interview prediction method predicts the left and right views by employing the original center view in the embedding and extracting. Note that both the spatially average and inter-component prediction methods may utilize distorted pixels when the adjacent pixels around the current pixel are distorted, that is, marked with hidden data. It might be the case that the first pixel set marked with O is already distorted when predicting the second pixel set marked with X, as displayed in Fig. 7. In this case, the prediction accuracy is usually very poor, but the VSP-based interview prediction is relatively accurate.

B. PERFORMANCE EVALUATION

For performance evaluation, the proposed RDH method with three different processes for the hole pixels was investigated with the conventional methods [23], [24], [25]. Table 2 shows the maximum embedding capacity for multiview depth views. It could be observed that the proposed method achieves higher capacity than the conventional methods. For example, Chung et al.'s method achieves the lowest capacity with an average of 0.6371 BPP, because the allowable distortion range is not fully used. Shi et al.'s method obtains the capacity of 1.0159 BPP by allowing the data to be embedded in the full range. Lee et al.'s method improves the average capacity by 1.1052 BPP through the inter-component prediction. It indicates that inter-component prediction is slightly more accurate than the spatial prediction. The proposed method achieves the highest capacity by determining the inter-component prediction and the VSP-based interview prediction adaptively and using the prediction-based allowable depth distortion range. It should be noted that the spatial and inter-component predictions are not effective for the hole pixels, because the hole generally occurs in the boundaries between the objects and background, which are drastically hard to be predicted accurately. Hence, we skipped the hole pixels without prediction in the proposed method for our experiments.

As illustrated in Table 3, the overall prediction accuracy of the proposed method is improved by 81.01%, but the other methods achieve the accuracy of about 58.96% and 63.52% on average. Since the embedding capacity is proportional to the accuracy, the proposed method can achieve the highest capacity. For the interviews, such as left and right views, the accuracy of the proposed method is much higher than the overall accuracy for the *Undo_Dancer*, *GT_Fly*, and *Shark* MVD images. It means that, since these images have the high interview correlation, the VSP-based interview prediction is very effective. As a result, more secret data can be inserted in the interviews than the center view. Fig. 11 shows binary maps marking the pixels where the prediction error between the original and predicted pixel values is equal to 0 for the *Poznan_Hall2*, *Poznan_Street*, *Undo_Dancer*, *GT_Fly*, and *Shark* MVD images with a size of 1920 × 1088. If the error is equal to 0, white was marked. If the error is occurred, black was marked. As shown in Fig. 11, many pixels are used for the embedding in the proposed method.


(a)

(b)

(c)

(d)

(e)

FIGURE 11. Original depth map (left) and its binary map marking the pixels (right) where the prediction error is 0 for the (a) Poznan_Hall2, (b) Poznan_Street, (c) Undo_Dancer, (d) GT_Fly, and (e) Shark MVD image with a resolution of 1920 × 1088.

Table 4 shows the performance of the proposed method, when using two different allowable depth distortion ranges, measured in terms of BPP, PSNR, and SSIM of the marked depth maps. In Table 4, $[L_d, U_d]$ represents the allowable

TABLE 2. Maximum embedding capacity of the conventional methods and the proposed method, when (a) the hole pixel is skipped without the prediction and (b) the spatially average prediction and (c) the inter-component prediction are instead used, respectively.

MVD	Chung et al.	Shi et al.	Lee et al.	Proposed		
				(a)	(b)	(c)
Poznan_Hall2	0.0061	0.7202	0.7344	0.9463	0.9463	0.9463
Poznan_Street	0.4058	0.7609	0.8254	0.8860	0.8860	0.8860
Undo_Dancer	0.4218	0.7703	0.8106	1.0926	1.0994	1.1003
GT_Fly	0.7956	1.0368	1.2598	1.7638	1.7702	1.7714
Kendo	0.9235	1.1265	1.2006	1.5551	1.5551	1.5551
Balloons	0.9042	1.0895	1.1803	1.5395	1.5395	1.5395
Newspaper	0.6793	1.1028	1.2029	1.2329	1.2329	1.2329
Shark	0.9067	1.5200	1.6272	2.3990	2.4056	2.4069
Average	**0.6371**	**1.0159**	**1.1052**	**1.4269**	**1.4294**	**1.4298**

TABLE 3. Prediction accuracy of the conventional methods and the proposed method for (a) all views and (b) left and right views only.

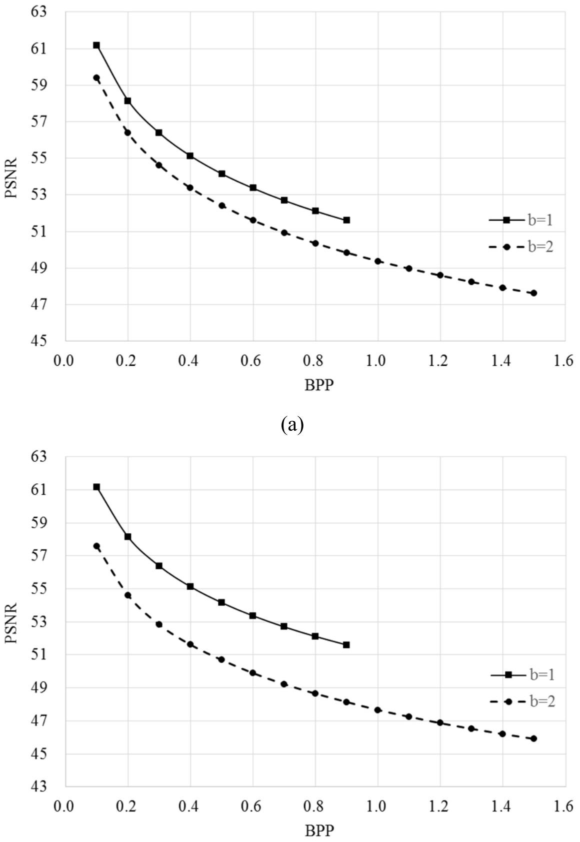
MVD	Shi et al.		Lee et al.		Proposed	
	All	Interview	All	Interview	All	Interview
Poznan_Hall2	72.05	70.24	73.47	71.66	94.66	94.66
Poznan_Street	58.08	57.40	62.80	62.19	68.86	69.01
Undo_Dancer	63.35	63.35	66.60	66.60	89.88	96.64
GT_Fly	51.84	51.86	62.99	62.98	88.19	97.94
Kendo	56.33	55.78	60.03	59.81	77.76	77.48
Balloons	54.47	54.68	59.02	59.44	76.97	76.66
Newspaper	61.98	60.29	65.94	62.85	68.23	64.52
Shark	53.55	53.56	57.32	57.32	83.52	96.33
Average	**58.96**	**58.40**	**63.52**	**62.86**	**81.01**	**84.16**

TABLE 4. Performance of the proposed method in terms of BPP, PSNR, and SSIM, when two different ranges are used, respectively.

MVD	[L_d, U_d]			[L_d^*, U_d^*]		
	BPP	PSNR	SSIM	BPP	PSNR	SSIM
Poznan_Hall2	0.7344	48.7947	0.9919	0.9463	51.3810	0.9957
Poznan_Street	0.8254	48.6437	0.9886	0.8860	49.6802	0.9906
Undo_Dancer	1.0602	47.4192	0.9896	1.0926	49.0731	0.9913
GT_Fly	1.7259	42.1596	0.9639	1.7638	45.3130	0.9746
Kendo	1.2006	45.3954	0.9805	1.5551	47.4616	0.9839
Balloons	1.1803	45.3544	0.9808	1.5395	47.5032	0.9843
Newspaper	1.2029	48.5187	0.9886	1.2329	48.9053	0.9892
Shark	2.3953	39.0669	0.9045	2.3990	39.1745	0.9051
Average	**1.2906**	**45.6691**	**0.9736**	**1.4269**	**47.3115**	**0.9768**

distortion range used in Shi et al.'s and Lee et al.'s methods. [L_d^*, U_d^*] means the proposed prediction-based range, which is calculated from eqs. (15) and (16). As shown in Table 4, both PSNR and SSIM from the proposed range are higher than those from the conventional range. In addition, higher capacity and less distortion are shown across all the MVD images. By making the marked pixel as close as possible to the predicted pixel, the proposed range distorts the pixel less than the conventional range. For example, as depicted in Fig. 7, if the inter-component prediction is determined as the best prediction in the proposed method, the second pixel set can be predicted from the first pixel set with the less distortion. Therefore, the proposed method can minimize the distortion while maintaining the high embedding capacity through the use of the prediction-based range.

The proposed method can constrain the maximum number of hidden bits b in eq. (14). Because b can directly affect the allowable distortion range, the embedding capacity and the depth distortion can be controlled. For instance, as depicted in Fig. 3 (a), all depth pixels in $[2, 7]$ have the disparity of 12 in the *Balloons* MVD image. Based on eqs. (15) and (16), if p and e are equal to 6 and 0, respectively, then $[L_{12}, U_{12}]$ are calculated into [4] and [7], as illustrated in Table 1. However, if b is limited to 1, the range becomes [5] and [6], and the number of possible h is reduced from 4 to 2. Therefore, the embedding capacity decreases, but the depth map can be less distorted. Fig. 12 depicts the depth quality of the proposed method for the right view of the *Balloons* and *GT_Fly* MVD images, when the embedding capacity increases from 0.1 to 1.5 BPP. Because the maximum b in these MVD images is equal to 2 under D-NOSE, it can be constrained with 1 or 2. It could be observed that the depth PSNR when b is equal to 1 is higher than that when b is equal to 2, but the maximum embedding capacity is lower.



(a)

(b)

FIGURE 12. Depth quality of the proposed RDH method under the different b values, when the capacity increases from 0.1 to 1.5 BPP, for the (a) Balloons and (b) GT_Fly MVD images.

Finally, we studied a portion of the location map size over the capacity, when the maximum number of hidden bit b is equal to 1 and 2, respectively. Table 5 shows PSNR and the portion at given BPPs (ranging from 0.1 to 0.5) for the *Shark* MVD image. As b increases from 1 to 2, both PSNR and the portion become lower at every BPP. This is because b with equal to 1 allows each pixel to be embedded by one bit only. Note that the low portion indicates high efficiency in terms of the embedding capacity, because data to be embedded includes secrete and side information, such as a compressed location map, its size, and a prediction type. On the contrary, the low PNSR of marked depth maps can represent unnatural quality, which is very undesirable in terms of data hiding. Hence, b should be carefully selected.

TABLE 5. Portion of the location map size over the embedding capacity.

BPP	b=1		b=2	
	PSNR	Portion	PSNR	Portion
0.1	61.1507	20.43	56.4057	10.48
0.2	58.1332	20.08	54.5183	10.13
0.3	56.3711	20.11	53.2836	10.12
0.4	55.1230	20.10	52.2888	9.99
0.5	54.1515	20.19	51.4057	10.01

V. CONCLUSION

Multiview depth maps were employed to embed the secret data. In order to obtain both the embedding capacity and low distortion, the proposed method checks the inter-component and interview correlations. Based on the correlation result, the optimum prediction method for the inter-component and interview predictions was adaptively determined to improve the prediction performance in embedding. Notably, VSP was used for interview prediction. The allowable depth distortion range could be adjusted around the predicted pixel value to minimize the distortion of the marked depth map while maintaining the high capacity. Experimental results showed that the proposed method achieved the higher capacity and the lower distortion than other conventional methods.

For further works, we are planning to optimize previous methods, which have been mainly developed for the texture images, on top of the depth maps. To ensure no synthesis distortion in virtual views, the concept of the allowable depth distortion range should be well-integrated into conventional methods.

REFERENCES

[1] W. Z. Khan, M. Y. Aalsalem, and M. K. Khan, "Communal acts of IoT consumers: A potential threat to security and privacy," *IEEE Trans. Consum. Electron.*, vol. 65, no. 1, pp. 64–72, Feb. 2019.

[2] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 8, pp. 890–896, Aug. 2003.

[3] A. M. Alattar, "Reversible watermark using the difference expansion of a generalized integer transform," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1147–1156, Aug. 2004.

[4] L. Kamstra and H. J. A. M. Heijmans, "Reversible data embedding into images using wavelet techniques and sorting," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2082–2090, Dec. 2005.

[5] Z. Ni, Y.-Q. Shi, N. Ansari, and W. Su, "Reversible data hiding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 3, pp. 354–362, Mar. 2006.

[6] D. M. Thodi and J. J. Rodriguez, "Expansion embedding techniques for reversible watermarking," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 721–730, Mar. 2007.

[7] Y. Hu, H.-K. Lee, and J. Li, "DE-based reversible data hiding with improved overflow location map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 250–260, Feb. 2009.

[8] X. Li, B. Yang, and T. Zeng, "Efficient reversible watermarking based on adaptive prediction-error expansion and pixel selection," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3524–3533, Dec. 2011.

[9] O. M. Al-Qershi and B. Ee Khoo, "Two-dimensional difference expansion (2D-DE) scheme with a characteristics-based threshold," *Signal Process.*, vol. 93, no. 1, pp. 154–162, Jan. 2013.

[10] X. Li, W. Zhang, X. Gui, and B. Yang, "Efficient reversible data hiding based on multiple histograms modification," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 9, pp. 2016–2027, Sep. 2015.

[11] X. Li, J. Li, B. Li, and B. Yang, "High-fidelity reversible data hiding scheme based on pixel-value-ordering and prediction-error expansion," *Signal Process.*, vol. 93, no. 1, pp. 198–205, Jan. 2013.

[12] B. Ou, X. Li, and J. Wang, "Improved PVO-based reversible data hiding: A new implementation based on multiple histograms modification," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 328–339, Jul. 2016.

[13] W. He, K. Zhou, J. Cai, L. Wang, and G. Xiong, "Reversible data hiding using multi-pass pixel value ordering and prediction-error expansion," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 351–360, Nov. 2017.

[14] T. Zhang, X. Li, W. Qi, and Z. Guo, "Location-based PVO and adaptive pairwise modification for efficient reversible data hiding," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2306–2319, 2020.

[15] S. Xiang and G. Ruan, "Efficient PVO-based reversible data hiding by selecting blocks with full-enclosing context," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2868–2880, May 2022.

[16] P. Merkle, K. Müller, and T. Wiegand, "3D video: Acquisition, coding, and display," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 946–950, May 2010.

[17] J. Y. Lee, J.-L. Lin, Y.-W. Chen, Y.-L. Chang, I. Kovliga, A. Fartukov, M. Mishurovskiy, H.-C. Wey, Y.-W. Huang, and S.-M. Lei, "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.

[18] G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, "Standardized extensions of high efficiency video coding (HEVC)," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 1001–1016, Dec. 2013.

[19] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, May 2004.

[20] Y. Zhao, C. Zhu, Z. Chen, and L. Yu, "Depth no-synthesis-error model for view synthesis in 3-D video," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2221–2228, Aug. 2011.

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

[22] J. Y. Lee and H. W. Park, "HEVC-based three-layer texture and depth coding for lossless synthesis in 3D video coding," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 20929–20945, Aug. 2020.

[23] K.-L. Chung, W.-J. Yang, and W.-N. Yang, "Reversible data hiding for depth maps using the depth no-synthesis-error model," *Inf. Sci.*, vol. 269, pp. 159–175, Jun. 2014.

[24] X. Shi, B. Ou, and Z. Qin, "Tailoring reversible data hiding for 3D synthetic images," *Signal Process., Image Commun.*, vol. 64, pp. 46–58, May 2018.

[25] J. Y. Lee, C. Kim, and C.-N. Yang, "Reversible data hiding using inter-component prediction in multiview video plus depth," *Electron.*, vol. 8, pp. 1–17, May 2019.

[26] K. Sayood, *Introduction to Data Compression*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2000.

[27] D. Rusanovskyy, K. Müller, and A. Vetro. *Common Test Conditions of 3DV Core Experiments*, document JCT3V-E1100, Aug. 2013.

JIN YOUNG LEE received the B.S. degree in information and communication engineering from Sungkyunkwan University, Suwon, South Korea, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2008 and 2018, respectively. He was with Samsung Electronics, Seoul, South Korea, from 2008 to 2018, where he was involved in the development of various video coding standards. Since 2018, he has been a Professor with the Department of Intelligent Mechatronics Engineering, Sejong University, Seoul. His research interests include image processing and video coding.