Synthesis-Distortion-Aware Hybrid Digital Analog Transmission for 3D Videos

PEIHAO LI1, FENGBAO YANG1, JING ZHANG2, YUN GUAN3,
ANHONG WANG4, (Member, IEEE), AND JIE LIANG5, (Senior Member, IEEE)
1School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
2College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China
3School of Data Science, North University of China, Taiyuan 030051, China
4Institute of Digital Media Communication, Taiyuan University of Science and Technology, Taiyuan 030024, China
5School of Engineering Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
Corresponding authors: Fengbao Yang (yfengb_1968@163.com) and Anhong Wang (wah_ty@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672373, in part by the North University of China Science and Technology Funding Project under Grant 20191630, in part by the Shanxi S&T Major Project under Grant 20191102010, in part by the Shanxi University S&T Achievements Transformation Project under Grant 20191042, in part by the Shanxi S&T Achievements Transformation Project under Grant 201804D131035, in part by the Shanxi S&T Innovation Team under Grant 201705D131025, in part by the Shanxi 1331 Project Key Innovation Research Team, in part by the Shanxi Collaborative Innovation Center of Internet+3D Printing, and in part by the Shanxi One Hundred Talented People.

ABSTRACT The hybrid digital-analog (HDA) video transmission scheme can be used to avoid the cliff effect and saturation effect. However, directly using HDA in 3D video transmission requires too much bandwidth. This paper addresses synthesis-distortion-aware HDA transmission for 3D videos to improve transmission performance. First, a 3D HDA framework that transmits both texture and depth videos in HDA mode is designed. Second, a recursive synthesis distortion estimation model, called RSDE-3D-HDA, is derived, where the transmission errors of both the texture and depth sequences are considered. Third, we optimize the power allocation between digital and analog signals based on the RSDE-3D-HDA. Finally, simulation results show that our model is accurate and the proposed 3D-HDA achieves better performance in terms of synthesis quality than state-of-the-art methods.

INDEX TERMS Hybrid digital-analog video transmission, distortion estimation, 3D video transmission, view synthesis.

I. INTRODUCTION

Three-dimensional (3D) video has drawn extensive attention from both industry and academia in recent years. 3D video contents are usually represented in the multiview video plus depth (MVD) format [1], including texture videos and associated depth videos. By transmitting the MVD data, a virtual view between any two captured views can be rendered on the decoder side by using depth image-based rendering (DIBR) technology [2], which has been adopted in the Moving Picture Experts Group View Synthesis Reference Software (MPEG-VSRS) [3].

Due to the quantization and entropy coding adopted by source coding, conventional digital video transmission schemes, such as H.264/AVC with digital channel coding and modulation for 2D videos, have two main disadvantages: first, when the channel quality is lower than a threshold, the received video quality will degrade dramatically, which is known as the cliff effect [4]; second, when the channel quality is greater than a threshold, even if the channel quality improves further, a better video quality cannot be obtained, this is called the quality saturation effect [5]. The problem is even worse in 3D video transmission because errors in one view can propagate to virtual views.

To overcome these two challenges, an analog-like Softcast method [6], [7] was developed for 2D video transmission. Unlike digital coding, Softcast does not use quantization and entropy coding for transmission; instead, it transmits the video sequence to the wireless channel directly after linear transformation and analog-like amplitude modulation. Therefore, the quality of the received video is linearly related to the channel signal-to-noise ratio (CSNR). However, due to the low compression efficiency of Softcast, the quality of the reconstructed video is worse than that of the reconstructed...
digital video at the same bandwidth. To solve this problem, the hybrid digital-analog (HDA) scheme was proposed in [8], [9] for 2D video transmission, which combined the advantages of digital and analog schemes by dividing the original sequence into digital and analog parts and transmitting the “analog-digital” superposed signal through the wireless channel after encoding and modulating them separately. In the digital coding part, the conventional digital video transmission scheme is adopted, and in the analog coding part, the residual signals between the original video and the reconstructed digital signals are transmitted using the Softcast method.

In [10], Softcast is used for 3D video transmission, which shows that the received video quality varies with the quality of the channel condition. However, it still suffers from low coding efficiency. In [11], a hybrid digital-analog scheme called Swift is proposed for stereo video transmission, and this scheme designs a zigzag coding structure such that both intraview and interview correlations can be explored through prediction among the frames to be analog coded; however, it does not consider the transmission of the depth map, and view synthesis is not considered.

In 3D video systems, it is crucial to develop an accurate algorithm for the encoder to estimate the distortion of the synthesized view of the decoder, which may facilitate the design of optimized encoding schemes at the encoder to improve the quality of the reconstructed 3D video. In 2D video transmission, a recursive optimal per-pixel estimation (ROPE) [12] method was proposed to estimate the decoder-side distortion from the encoder. In [13], a recursive optimal distribution estimation (RODE) approach was developed to obtain the decoder-side distortion of the synthesized virtual views in 3D video transmission. In [14], a depth-bin-based graphical model (DBGM) for view synthesis distortion estimation was developed to improve the speed of the RODE method. A recursive distortion estimation method is proposed in [15] for HDA-based single-view video transmission, which recursively estimates the decoder-side distortion from the encoder and adaptively allocates the transmission power between the digital and analog signals in HDA. Very recently, for 3D Softcast, Laplacian multipliers were used in [16] to find the closed-form expression of the distortion by transforming the power allocation problem into an optimization problem. The mean-removed block-based discrete cosine transform (DCT) and view synthesis distortion were performed to scale the interblock DCT coefficients. However, [16] and [17] focused on Softcast, not HDA, and there has not been any work reported on 3D-HDA.

Motivated by the aforementioned work, this paper studies 3D video transmission schemes based on the principles of HDA. The main contributions are as follows. 1) We design a video transmission framework for 3D-HDA, which allows transmission errors, such as packet losses, to occur in both the texture and depth signals. 2) We derive an expression for the recursive synthesis distortion estimation of 3D-HDA, named RSDE-3D-HDA, which can estimate the decoder-side distortion at the encoder and can be used to optimize the resource allocation for 3D-HDA. 3) We give the optimal solution to the power allocation between analog and digital signals on the 3D-HDA model. To the best of our knowledge, this is the first time HDA-based 3D video transmission is considered with power allocation.

The remaining parts of this paper are organized as follows. We first review the related work, including the frameworks of HDA and the view synthesis of 3D video in Section 2. Section 3 presents the proposed 3D-HDA framework, the RSDE-3D-HDA model and the optimal solution. The experimental results are presented in Section 4. Finally, Section 5 concludes this paper.

II. RELATED WORK

A. HYBRID DIGITAL-ANALOG SCHEME

In this section, we briefly review the framework of the hybrid digital-analog scheme and the transmission process of digital and analog signals. Further details can be found in [8], [9] and [18]. Fig. 1 shows the framework of the HDA system, which consists of the digital and analog codec parts. At the digital encoder, the original video sequences are first encoded by H.264/AVC, and the resulting bitstreams are processed by forward error correction (FEC), modulation and power allocation to obtain the digital signal. At the Softcast encoder, the residual between the original sequence and the H.264 reconstructed sequence is processed using DCT, power allocation and whitening to obtain the analog signal. Finally, the digital and analog signals are superimposed and transmitted to the OFDM channel.

At the decoder, the error-free demodulated digital signal is obtained first (by treating the analog signal as noise), and then the modulated signal is subtracted by the received signal to obtain the analog signal. At the Softcast decoder, the linear least squares estimator (LLSE) removes the noise, the inverse discrete cosine transform (IDCT) reconstructs the analog signal, and the digital signal is reconstructed by the H.264 decoder so that the reconstructed video sequence is obtained via these two signals. Note that the standard
deviations of the DCT coefficients are particularly important for analog decoding; therefore, they are transmitted as side information to help LLSE. HDA inherits the advantages of digital and analog schemes, and can achieve improved performance. However, currently, most HDA research focuses on 2D videos, and there are few works applying HDA for the transmission of 3D videos.

B. THE VIEW SYNTHESIS SYSTEM FOR 3D VIDEOS
As shown in Fig. 2, the view synthesis of 3D video is performed by a general framework consisting of a video collector, coding, channel transmission, decoding and synthesis of the virtual view and display. The acquisition process usually involves at least two cameras recording from different angles that can capture depth maps to form a “texture-to-depth” group (GTP) of different views. Virtual view synthesis is the key technology in this framework; it uses decoded reference views to synthesize unknown virtual views, and the synthesis techniques are generally divided into two categories [13]: image-based rendering (IBR) and DIBR. IBR uses the disparity between different views to synthesize virtual views [19], while DIBR uses the geometric information of the depth map to generate virtual views. Therefore, DIBR can achieve better virtual viewpoint quality than IBR [20]. Generally, the DIBR’s framework can be divided into two types, that is, the free-view mode and the one-dimensional parallel mode, according to the placement of the camera in the acquisition process. The former mode needs to consider disparity in multiple directions, which is complicated, whereas the latter mode, which is adopted in this paper, needs to consider only the disparity in the horizontal direction, which is relatively simple.

In the subsequent encoding process, the depth information needs to be processed, such as with subsampling and quantizing [21], [22], to obtain sparse and smooth information, and different distortion scenes need to be evaluated. If only the distortion caused by the encoding is considered, then the GTP needs to be processed by the codec, and the distortion of the view rendering is estimated based on the GTP before and after reconstruction. If only considering the distortion caused by packet loss during the transmission process, then a simulated packet loss operation needs to be performed on the compressed bitstreams and the view synthesis distortion needs to be estimated according to the GTP before and after reconstruction. Finally, according to the results of view synthesis distortion, the appropriate coding scheme is determined after reasonable rate-distortion optimization. The bitstreams are then reconstructed at the decoder after passing through the channel.

The above process represents a digital transmission method, which inevitably causes cliff effects and saturation effects. To overcome these issues, we apply HDA, which divides the GTP into two parts, i.e., a digital signal and an analog signal, and transmits the superposition of these two signals. In addition, we deduce an expression of the synthesis estimation for the 3D video, such as the power allocation factor of HDA in Fig. 1, to achieve better transmission parameter optimization, and the quality of the video received by the audience can be optimized. The reconstructed GTP will then use the DIBR from [2] to render a virtual view video. In the actual transmission scenario, because the distortion of view synthesis is quite necessary, regardless of the number of multi-view videos (MVVs) or the angle of view of the FVV [23], it is difficult to achieve a reasonable transmission cost in the development of 3D video. Thus, a better solution is to selectively transmit several reference views and then use the transmitted reference views to synthesize the final views at the receiver instead of transmitting all views. Regarding the smoothness of the 3D video in the receiver, it is important to consider the relationship between the quality of virtual views and the transmission parameters, and this factor is the main motivation of this paper.

III. THE PROPOSED 3D-HDA VIDEO TRANSMISSION SYSTEM WITH THE RSDE-3D-HDA MODEL
A. THE PROPOSED FRAMEWORK OF 3D-HDA
Fig. 3 shows the overall framework of the proposed 3D-HDA system, where only two views are considered under the assumption that they are transmitted independently.

At the encoder, let $T_l$ and $D_l$ denote the original texture and depth maps of the left view, which are first encoded by the H.264/AVC digital encoder to obtain $T_{ld}$ and $D_{ld}$, respectively. Then, $D_{ld}$ and $T_{ld}$ are cascaded and modulated to obtain the digital signal $X_{ld}$. The analog part $T_{lu}$ is from the residual texture between the original texture $T_l$ and the digital reconstructed texture at the encoder. Finally, after DCT and power allocation, the analog $Y_{lu}$ and digital $Y_{ld}$ are superposed to obtain the transmission signal $S_{lc}$, as expressed by Eq. (1).

$$S_{lc} = Y_{lu} + Y_{ld} = \alpha T_{lu} + \beta X_{ld}$$

(1)

where $\alpha$ and $\beta$ are the power allocation factors for the analog and digital signals, respectively.

Then, $S_{lc}$ is transmitted into the OFDM wireless channel with additive white Gaussian noise (AWGN), i.e., $N \sim N(0, \sigma^2)$. At the receiver, the received superposed signal is

$$\tilde{S}_{lc} = S_{lc} + N$$

(2)
\( \tilde{S}_{lc} \) is first demodulated to obtain the digital signal \( \tilde{Y}_{ld} \) by treating the analog signal as noise. The digital signal \( \tilde{X}_{ld} \) is decoded by H.264 with error concealment and power allocation to obtain the recovered depth \( \tilde{D}_{l} \) and the texture \( \tilde{T}_{ld} \). Then, the demodulated digital signal \( \tilde{Y}_{ld} \) is modulated and subtracted from the received superposed signals \( \tilde{S}_{lc} \). After that, the reconstructed analog signal \( \tilde{T}_{la} \) is obtained by \( \tilde{Y}_{ld} \) through power allocation and analog decoding, including inverse DCT and LLSE, which is similar to Softcast. Next, the reconstructed texture \( \tilde{T}_{l} \) and depth \( \tilde{D}_{l} \) are warped to obtain the virtual \( \tilde{S}_{ls} \) of the left view.

For the right view, through the same operation as the left view, we can obtain its virtual \( \tilde{S}_{rs} \). Finally, \( \tilde{S}_{ls} \) and \( \tilde{S}_{rs} \) are used to obtain the synthesized view \( \tilde{S}_{s} \):

\[
\tilde{S}_{i}(i, m, n) = \begin{cases} 
(1 - a)\tilde{S}_{ls}(i, m, n) + a\tilde{S}_{rs}(i, m, n) & \text{if both left and right views are available} \\
\tilde{S}_{ls}(i, m, n) & \text{if only left view is available} \\
\tilde{S}_{rs}(i, m, n) & \text{if only right view is available} \\
\tilde{S}_{s}^h(i, m, n) & \text{if there are some holes}
\end{cases}
\]

where \( \tilde{S}_{ls}(i, m, n) \) and \( \tilde{S}_{rs}(i, m, n) \) denote the \( i \)-th texture pixel in the \( m \)-th block of the synthesized frame \( n \) warped from the left and right reference views, respectively. \( a \in [0, 1] \) is a scaling factor that depends on the location of the virtual view. Additionally, we use an interpolated pixel \( \tilde{S}_{s}^h(i, m, n) \) to represent the hole pixel value when there is no information provided by either the left or right view.

**B. CASCADING AND SUPERPOSITION SCHEME**

In this paper, we assume that the transmission bandwidth matches the encoded bit rate. Each transmitted signal is the superposition of a digital signal and an analog signal. Since the depth images represent the structure information, which is important in 3D warping, we transmit depth images in digital mode. In addition, we explore a cascading mode for digital depth and digital texture. Specifically, the total digital transmission bandwidth is divided into two parts; one part is for the digital texture and the other part is for the digital depth, as shown in Fig. 4, in which \( D_{ld} \), \( T_{ld} \) and \( T_{la} \) denote the digital signal in the depth map, the digital signal and the analog signal of the texture of the left view, respectively. Meanwhile, to easily distinguish these two types of digital signals, a small number of padding bits are added between them, while the transmission bandwidth for the padding bits is ignored for simplification. The power allocation between \( X_{ld} \) and \( T_{la} \) will be discussed in Sec.III.D. Note that, although the rate allocation between \( T_{ld} \) and \( D_{ld} \) should be optimized, experiments show that such optimization is very complicated, and the gain is limited; therefore, for simplification, similar to traditional 3D video transmission, a fixed bit ratio of 4:1 is allocated to them.

**C. THE DEDUCTION OF RSDE-3D-HDA**

In this section, we obtain the distortion expression for the synthesized view \( \tilde{S}_{s} \). In particular, we propose a recursive synthesis distortion estimation model for the 3D-HDA framework of Fig. 3, called RSDE-3D-HDA. The derivation is given below.

We consider only the 1D parallel view synthesis process. According to Eq. (3), \( \tilde{S}_{i}(i, m, n) \) has four cases, depending on whether the left and right views are warping to the current synthesized position. Therefore, the synthesized distortion can be written as the weighted average of four cases:

\[
ED_{i}(i, m, n) = \sum_{k=1}^{4} ED_{ik}(i, m, n) \times p_{k}
\]

where \( p_{k} \) and \( ED_{ik}(i, m, n) \) represent the probability and the distortion of each case, respectively. Next, we find the expression of each \( ED_{ik}(i, m, n) \) and the corresponding \( p_{k} \).
1) SCENARIO 1

Since both the left and right views can warp to the current position, based on Eq. (4), distortion in this scenario can be expressed as

\[
ED_{s1}(i, m, n) = E\{(S_i(i, m, n) - \tilde{S}_i(i, m, n))^2\}
\]

where \(S_i(i, m, n)\) and \(\tilde{S}_i(i, m, n)\) represent the i-th pixel value in the m-th block of the correctly synthesized texture frame n and the reconstructed texture frame n at the decoder when the reconstructed texture frame contains random errors.

By adding four auxiliary items \(- (1 - a)S_{ls}(i, m, n), -aS_{rs}(i, m, n), (1 - a)S_{ls}(i, m, n)\) and \(aS_{rs}(i, m, n)\) to Eq. (5) and rearranging it, we can obtain

\[
ED_{s1}(i, m, n) = E\{(S_i(i, m, n) - (1 - a)S_{ls}(i, m, n) - aS_{rs}(i, m, n) + (1 - a)\tilde{S}_i(i, m, n))
+ a(S_{rs}(i, m, n) - \tilde{S}_{rs}(i, m, n))^2\}
\]

This formula shows that the expected distortion is characterized by the known value of \(S_i(i, m, n)\), \(S_{ls}(i, m, n)\), \(S_{rs}(i, m, n)\), and the first and second moments of \(S_{ls}(i, m, n) - \tilde{S}_i(i, m, n)\) and \(S_{rs}(i, m, n) - \tilde{S}_{rs}(i, m, n)\).

Next, we should deduce the second moments of \(S_{ls}(i, m, n) - \tilde{S}_i(i, m, n)\) and \(S_{rs}(i, m, n) - \tilde{S}_{rs}(i, m, n)\). Based on 3D warping, the value of \(\tilde{S}_i(i, m, n)\) is similar to that of \(T_i(i, m, n)\), and because the distortions of \(S_{ls}(i, m, n)\) and \(S_{rs}(i, m, n)\) are the same, we take the distortion of the left view as an example:

\[
ED_l(i, m, n) = E\{(S_{ls}(i, m, n) - \tilde{S}_{ls}(i, m, n))^2\}
\]

where \(S_{ls}(i, m, n)\) and \(\tilde{S}_{ls}(i, m, n)\) represent the same pixel in the left view as an example.

Similarly, the two moments \((k = 1, 2)\) of the expected distortion are

\[
E\{\tilde{Y}_{ld}(i, m, n)^k\} = (1 - p)\tilde{Y}_{ld}(i, m, n)^k + pE\{\tilde{Y}_{ld}(i, m, n - 1)^k\}
\]
For the BER $p$ in Eq. (10) and Eq. (12), since the analog signal is considered as noise to the digital signal, we find an expression for $p$ with BPSK modulation [24].

$$p = \frac{1}{2} * \text{erfc}(\sqrt{\text{CSNR}_c})$$

$$\text{CSNR}_c(i, m, n) = \frac{P_{Yla}(i, m, n)}{P_{Yld}(i, m, n) + \sigma^2}$$

where $P_{Yld}(i, m, n)$ and $P_{Yla}(i, m, n)$, i.e., the powers of the digital and analog signals, respectively, can be expressed as

$$P_{Yld}(i, m, n) = E[Y_{ld}(i, m, n)^2] = \beta^2 X_{ld}(i, m, n)$$

$$P_{Yla}(i, m, n) = E[Y_{la}(i, m, n)^2] = \alpha^2 T_{la}(i, m, n)^2$$

In summary, according to Eq. (8) to Eq. (14), we can recursively estimate $ED_{ld}(i, m, n)$ at the encoder when $\alpha$ and $\beta$ are known.

Then, we can obtain the distortion of the analog signal according to the analog encoding process of Fig. 3. Specifically,

$$ED_{ld}(i, m, n) = E[(T_{ld}(i, m, n) - \tilde{T}_{ld}(i, m, n))^2]$$

$$= E[(T_{ld}(i, m, n) - Y_{ld}(i, m, n) + Y_{la}(i, m, n) - \tilde{T}_{ld}(i, m, n))^2]$$

where $Y_{ld}(i, m, n)$ and $\tilde{T}_{ld}(i, m, n)$ are the analog signal after power allocation at the encoder and the reconstruction signal at the decoder through the channel. Therefore, we can obtain

$$Y_{ld}(i, m, n) = \alpha T_{ld}(i, m, n)$$

$$\tilde{T}_{ld}(i, m, n) = Y_{ld}(i, m, n) + N$$

According to Eq. (15) and Eq. (16), we divide the distortion of the analog signal into two parts: one caused by power allocation and the other caused by noise interference during channel transmission. For the first part, the distortion of the $i$-th pixel in the $m$-th block of the $n$-th frame is denoted as $D^2_{ld}(i, m, n)$; then, the expected distortion is

$$ED^p_{ld}(i, m, n) = E[(T_{ld}(i, m, n) - Y_{ld}(i, m, n))^2]$$

$$= (1 - \alpha^2) T_{ld}(i, m, n)^2 + \sigma^2$$

where $\sigma^2$ is the noise variance.

$$ED_{r}(i, m, n) = E[(D^2_{rd}(i, m, n) + N^2)]$$

$$= (1 - \alpha^2) T_{ld}(i, m, n)^2 + \sigma^2$$

during the warping competition. The edge between $\tilde{V}(j)$ and $\tilde{V}_s(i, m, n)$ is denoted as $b_{ji}$. When the edge $b_{ji}$ is the final winner, all the edges emitted from $\tilde{V}(j)(z = 1, \ldots, j - 1, j + 1, \ldots, n_b)$ to $\tilde{V}_s(i, m, n)$ with the condition should be abandoned. Thus, the winning probability of $b_{ji}$ can be formulated as

$$P_{\text{win}}(b_{ji}) = P(b_{ji}) \times \prod_{z=1}^{j-1} (1 - P(b_{zj})) \times \prod_{z=j+1}^{n_b} (1 - P(b_{zj}))$$

where $P(b_{ji})$ is the probability mass function of the $i$-th decoded pixel, which can be recursively estimated using the decoded depth and texture with the given values of $\alpha$ and $\beta$.

Based on Eq. (19), the probability of $\tilde{S}_s(i, m, n)$ being a hole can be obtained as

$$P_{\tilde{S}_s(i, m, n)}(\phi) = \prod_{j=1}^{n_b} (1 - P_{\text{win}}(b_{ji}))$$

Therefore, the probability of Scenario 1 is

$$p_1 = P(\tilde{S}_l(i, m, n) \neq \phi) \times P(\tilde{S}_{rs}(i, m, n) \neq \phi)$$

$$= (1 - P_{ls(i,m,n)}(\phi)) \times (1 - P_{rs(i,m,n)}(\phi))$$

2) SCENARIO 2

When only the left view can be warped to the current synthesis position, the distortion in this scenario can be expressed as

$$ED_{l2}(i, m, n) = E[(S_{ls}(i, m, n) - \tilde{S}_l(i, m, n))^2]$$

$$= E[(S_{ls}(i, m, n) - S_l(i, m, n) + S_{ls}(i, m, n) - \tilde{S}_l(i, m, n))^2]$$

It is obvious that $ED_{l2}(i, m, n)$ is determined by the distortion of the left view. Similar to Eq. (21), the probability of this scenario is

$$p_2 = P(\tilde{S}_l(i, m, n) \neq \phi) \times P(\tilde{S}_{rs}(i, m, n) = \phi)$$

$$= (1 - P_{ls(i,m,n)}(\phi)) \times P_{rs(i,m,n)}(\phi)$$

3) SCENARIO 3

When only the right view can be warped to the current synthesis position, similar to Eq. (22), this scenario can be expressed as

$$ED_{r3}(i, m, n) = E[(S_{rs}(i, m, n) - \tilde{S}_r(i, m, n))^2]$$

$$= E[(S_{rs}(i, m, n) - S_{rs}(i, m, n) + S_{rs}(i, m, n) - \tilde{S}_r(i, m, n))^2]$$

The probability of this scenario is

$$p_3 = P(\tilde{S}_l(i, m, n) = \phi) \times P(\tilde{S}_{rs}(i, m, n) \neq \phi)$$

$$= P_{ls(i,m,n)}(\phi) \times (1 - P_{rs(i,m,n)}(\phi))$$
4) SCENARIO 4
When neither the left nor the right view can be warped to the current position, a simple hole filling method is used:

\[
\tilde{S}_i(i, m, n) = \begin{cases} 
\tilde{S}_i(i - 1, m, n) & \text{left available} \\
\tilde{S}_i(W - i, m, n) & \text{right available} \\
128 & \text{otherwise}
\end{cases}
\]  

(26)

where \( W \) is the width of the image.

The distortion and probability of this scenario are thus

\[
ED_{\delta i}(i, m, n) = E(\tilde{S}_i(i, m, n) - \tilde{S}_i(i, m, n))
\]  

(27)

\[p_4 = P(\tilde{S}_i(i, m, n) = \phi) * P(\tilde{S}_i(i, m, n) = \phi) = P_{\delta i}(i, m, n)(\phi) * P_{\delta i}(i, m, n)(\phi)
\]  

(28)

Thus, in Eq. (4), when the values of \( \alpha \) and \( \beta \) are given, all the variables are known or can be estimated by the reconstructed pixels on the same frame or the previous frame \( n - 1 \). Consequently, we can recursively estimate the value of \( ED_i \) of the decoder in Eq. (4).

D. OPTIMIZED POWER ALLOCATION
In this subsection, we focus on solving the power allocation between the analog signal \( T_a \) and the digital signal \( X_d \) involved in RSDE-3D-HDA. We use the same power allocation factors \( \alpha \) and \( \beta \) for both the left and right views. Thus, the optimization problem can be expressed as

\[
\arg \min_{\alpha, \beta} ED_i
\]  

s.t. \[P_{Yd}(i, m, n) + P_{Yd}(i, m, n) = P_{Sc}(i, m, n)\]

\[P_{Yd}(i, m, n) = \alpha^2 T_d^2(i, m, n)\]

\[P_{Yd}(i, m, n) = \beta^2 X_d^2(i, m, n)\]

\[CSNR_e(i, m, n) = \frac{P_{Yd}(i, m, n)}{P_{Yd}(i, m, n) + \sigma^2}\]

(29)

where \( ED_i(i, m, n) \) is the distortion expressed in Eq. (4) and \( P_{Sc}(i, m, n) \) is the total transmission power budget of the \( i \)-th texture pixel in the \( m \)-th block in the \( n \)-th frame for the left view or right view, which is divided into the analog transmission power \( P_{Yd}(i, m, n) \) and the digital signal transmission power \( P_{Yd}(i, m, n) \).

We use exhaustive search to solve the optimization formula that we derived in Eq. (29).

First, we should determine the search range of \( \alpha \) and \( \beta \): From their clear relationship seen in Eq. (29), \( \beta \) can be expressed in terms of \( \alpha \) as follows:

\[
\beta = \sqrt{\frac{P_{Sc}(i, m, n) - \alpha^2 T_d^2(i, m, n)}{X_d^2(i, m, n)}}
\]  

(30)

Therefore, we need the range of \( \alpha \), assuming \( \alpha \in (0, \alpha_{\text{max}}) \). From Eq. (30), we can see that the maximum of \( \alpha \) corresponds to the case of \( \beta \) equal to 0, meaning that the total power is allocated to the analog signal. Specifically,

\[
\alpha_{\text{max}}(i, m, n) = \sqrt{\frac{P_{Sc}(i, m, n)}{T_d^2(i, m, n)}}
\]  

(31)

where \( P_{Sc}(i, m, n) \) and \( T_d(i, m, n) \) are given constants.

Second, we should choose the search step of \( \alpha \). For simplicity, we uniformly compute the search step over the range of \( \alpha \) according to the search number. Specifically,

\[
\Delta_\alpha = \frac{\alpha_{\text{max}}}{N_i}
\]  

(32)

where \( N_i \) is the given value of the search number.

Finally, the optimal \( \alpha \) corresponds to the minimized distortion of Eq. (29). Additionally, we can obtain the optimal \( \beta \) according to Eq. (30).

From the experiments, we find that frame-level power allocation is better satisfied than block-level power allocation. Therefore, our power is computed at the block, but the same power factor \( \alpha \) is used for all blocks in the same frame level:

\[
P(n) = \frac{1}{n_i \times n_m} \sum_{n=1}^{n_m} \sum_{i=1}^{n_i} P(i, m, n)
\]  

(33)

where \( n_i \) is the number of pixels in each block, which is set as 64, \( n_m \) is the number of blocks, and \( P(i, m, n) \) is the power of the \( i \)-th analog signal in the \( m \)-th block in frame \( n \).

In addition, the search complexity and the transmission overhead of the power factor are negligible because, in our experiments, the number of possible values (search range) of \( \alpha \) for each frame is set to 32, and the overhead consists of the 5-bit index of \( \alpha \) and the bitrates for \( \alpha_{\text{max}} \), which costs 10 bits in our scheme. Above all, the proposed RSDE-3D-HDA model is the first to combine 3D video view synthesis with HDA-based power allocation, and it can accurately estimate the transmission error. According to the simulation results in the next section, we can observed that both cliff effect and saturation effect are eliminated.

IV. SIMULATION RESULTS
In this section, we evaluate the performance of our proposed method. We perform four groups of experiments measuring four metrics: 1) the average MSE mismatch ratio (AMMR) performance of the RSDE-3D-HDA, 2) the mean-square error (MSE) performance of the RSDE-3D-HDA, 3) the peak signal-to-noise ratio (PSNR) performance of the proposed 3D-HDA, and 4) the visual quality of the synthesis virtual view. The first two experimental groups are used to verify the RSDE-3D-HDA model performance. The last two experimental groups are used to verify the objective and subjective view quality performance of our 3D-HDA. Additionally, we compare our methods with some popular schemes, such as 3D-Softcast, the 3D-Digital scheme and the distortion estimation model of the RODE method in [13].

The experimental conditions are set as follows. For the digital part, because H.264/AVC is a popular method and our research focuses on the performance of proposed framework, we employed JM 19.0 of the H.264/AVC video codec. For the parameter \( a \), we consider a virtual view in the middle of two reference views; thus, \( a \) is set to a fixed value of \( \frac{1}{4} \). We use a single reference frame for motion estimation. The first frame is coded as the I-frame, and all remaining frames are coded as P-frames. For the analog part, we employ the approach...
FIGURE 5. Performance of the RSDE-3D-HDA model for Kendo and Newspaper sequence. (a), (b), (c) and (d) illustrate MSE performances with CSNR being 0 dB, 5 dB, 10 dB, 20 dB respectively. (e), (f), (g) and (h) illustrate MSE performances with CSNR being 0 dB, 5 dB, 10 dB, 20 dB respectively.
in [6] to process the video sequences. Notably, all the schemes use the same transmission power and bandwidth (the same resolution uses the same bandwidth).

The AMMR and PSNR are defined as

$$\text{AMMR} = \frac{1}{N} \sum_{n=1}^{N} \frac{|E_{\text{est}}(n) - E_{\text{sim}}(n)|}{E_{\text{sim}}(n)}$$ \hspace{1cm} (34)

where $N$ is the number of frames in the sequence, $E_{\text{est}}(n)$ is the encoder-estimated MSE for frame $n$, and $E_{\text{sim}}(n)$ is the actual receiver-side MSE for frame $n$ obtained by averaging over multiple experiments.

$$\text{PSNR} = 10 \log_{10} \left( \frac{(255)^2}{\text{MSE}} \right)$$ \hspace{1cm} (35)

where $\text{MSE}$ is the mean square error between the original video frames and the reconstructed frames.

### A. AMMR PERFORMANCE

In this part, we first test the AMMR performance of our proposed RSDE-3D-HDA model with the testing sequences Kendo, Newspaper (1024*768@30 Hz, 100 frames), Balloons (1920*1088@30 Hz, 100 frames) and Cafe (1920*1080@30 Hz, 100 frames) [25]. The estimated distortion of our RSDE-3D-HDA scheme is applied at the encoder, while the simulated distortion is computed at the decoder. Table 1 presents the AMMR performance with CSNRs of 0 dB, 5 dB, 10 dB and 20 dB. The results show that the proposed method is quite accurate. In addition, we compare the AMMR performance with the referenced methods in [13] for the Balloons, PoznanStreet (1920*1088@30 Hz, 200 frames), Kendo, and Dancer (1024*768@30 Hz, 200 frames) [25] video sequences. For fairness, we adjust the experimental parameters to match the loss rate and group of pictures (GOP) in [13]. Table 2 presents the AMMR performance, with loss rates of 0.02, 0.05, and 0.08, the matched GOP and the total number of frames. “Ours” and “RODE” indicate the results of our method and the results from [13]. The comparison results show that the proposed method is much more accurate than RODE.

### B. MSE PERFORMANCE

We test the MSE performance of the proposed RSDE-3D-HDA model with the Kendo and Newspaper (1024*768@30 Hz, 100 frames) video sequences with only the luminance component to verify the correctness of our methods. Fig. 5 shows the MSE performances of the Kendo and Newspaper sequences for channel signal-to-noise ratios (CSNRs) of 0 dB, 5 dB, 10 dB and 20 dB. The AMMRs for the Kendo sequences are only 6.67%, 0.66%, 1.54%, and 6.84%, and those for the Newspaper sequences are 0.52%, 1.05%, 1.38%, and 1.79%, respectively. The estimated distortion at the encoder matches the actual decoder-side distortion quite well. Additionally, it can be seen that the MSE decreases

---

**TABLE 1. AMMR (%) of the proposed RSDE-3D-HDA model.**

| Sequence   | CSNR | 0 dB | 5 dB | 10 dB | 20 dB |
|------------|------|------|------|-------|-------|
| Kendo      | 6.67 | 0.66 | 1.54 | 6.84  |
| Newspaper  | 0.52 | 1.05 | 1.38 | 1.79  |
| Balloons   | 2.14 | 0.69 | 1.33 | 4.14  |
| cafe       | 0.05 | 0.36 | 0.13 | 1.11  |
| Average    | 2.35 | 0.69 | 1.01 | 3.47  |

**TABLE 2. AMMR(%) comparation.**

| Sequence   | GOP Size | 0.02 | 0.05 | 0.08 |
|------------|----------|------|------|------|
| Balloons   | 60       | RODE | Ours | RODE | Ours |
| Kendo      | 30       | 3.56 | 0.83 | 6.03 | 4.04 |
| Dancer     | 30       | 6.53 | 1.97 | 5.93 | 2.44 |
| PoznanStreet | 60   | 8.79 | 0.22 | 7.84 | 0.31 |

---
with increasing CSNR, which means that the received video quality significantly improves.

C. PSNR PERFORMANCE

We test the PSNR performance of the proposed 3D-HDA with other standard methods, such as the 3D-Digital method and 3D-Softcast method, and the Kendo and Newspaper (1024*768@30 Hz, 100 frames) video sequences are taken as examples. For fairness, we use the same bandwidth for all schemes: for the 3D digital scheme, the H.264/AVC method with 1/2 FEC coding is used to guarantee error-free digital transmission, and the bitstreams of the texture and depth map are cascaded. For the 3D-Softcast method, both texture and depth are encoded by Softcast, and the packet loss guarantees the same bandwidth as other methods. Additionally, the ratio between the texture and depth maps is set to 4:1 in these methods.

We can see from Fig. 6 that the digital scheme cannot overcome the cliff effect or saturation effect because of the fixed quantization parameter. Although the performance of Softcast tends to be linear, it cannot achieve an acceptable performance. The proposed 3D-HDA without optimization (meaning $\alpha = 1$) can obtain linear performance, while the
PSNR performance is at a lower level than that of the digital scheme when the CSNR is poor. Only the proposed 3D-HDA with the optimized power allocation (meaning α is optimized by RSDE-3D-HDA) can achieve superior performance at both low and high CSNRs and overcome the cliff effect and saturation effect perfectly.

D. VISUAL PERFORMANCES
Furthermore, we show the visual comparison of the 10th frame of the Newspaper sequence and the 15th frame of the Kendo sequence with the local enlargement of each frame. It can be seen from Fig. 7 and Fig. 8 that Softcast could not obtain acceptable performance. Our 3D-HDA method with optimization power allocation obtains the best performance in terms of visual quality, and the results are the closest to the ground truth of the synthesis view.

V. CONCLUSION
In this paper, we develop a framework for HDA-based 3D video transmission and deduce a recursive view synthesis distortion model. The experimental results show that the accuracy of the proposed distortion estimation scheme is satisfactory and that the framework outperforms state-of-the-art schemes. There are three main advantages of our scheme. First, the framework can tolerate the depth and texture errors during wireless transmission. Second, the scheme can estimate the synthesis distortion of the decoder from the encoder side due to the deduced distortion model. Third, due to the exploration of HDA transmission for 3D video, the quality of the synthesized virtual view is linear with the CSNR, overcoming the cliff effect and saturation effect in digital transmission methods.

REFERENCES
[1] C. Fehn, “Depth-image-based rendering (DBR), compression, and transmission for a new approach on 3D-TV,” Proc. SPIE, vol. 5291, pp. 93–104, May 2004.
[2] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, “View synthesis techniques for 3D video,” Proc. SPIE, vol. 7443, pp. 1–11, Sep. 2009.
[3] View Synthesis Reference Software (VSRS) Version 3.5. Accessed: Feb. 19, 2013. [Online]. Available: http://wg11.sc29.org/svnrepos/MPEG-4/test/tags/3D/viewsynthesis/VQRS-3-5
[4] M. Skoglund, N. Phamdo, and F. Alajaji, “Hybrid digital–analogue source–coding for bandwidth compression/expansion,” IEEE Trans. Inf. Theory, vol. 52, no. 8, pp. 3757–3763, Aug. 2006.
[5] T. Kratochvíl and R. Štukavec, “DVB-T digital terrestrial television transmission over fading channels,” Radioengineering, vol. 17, no. 3, pp. 96–102, Sep. 2008.
[6] S. Jakubczak and D. Katabi, “A cross-layer design for scalable mobile video,” in Proc. 17th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom), Las Vegas, NV, USA, Sep. 2011, pp. 289–300.
[7] S. Jakubczak, H. Rahul, and D. Katabi, “One-size-fits-all wireless video,” in Proc. 8th ACM SIGCOMM HotNets Workshop, New York, NY, USA, Oct. 2009, pp. 449–450.
[8] L. Yu, H. Li, and W. Li, “Wireless scalable video coding using a hybrid digital-analog scheme,” IEEE Trans. Circuits Syst. Video Technol., vol. 24, no. 2, pp. 331–345, Feb. 2014.
[9] L. Yu, H. Li, and W. Li, “Hybrid digital-analog scheme for video transmission over wireless,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), Beijing, China, May 2013, pp. 1163–1166.
[10] T. Fujishashi, T. Koike-Akino, T. Watanabe, and P. V. Orlik, “Soft video delivery for free viewpoint video,” in Proc. IEEE Int. Conf. Commun. (ICC), Paris, France, May 2017, pp. 1–7.
[11] D. He, C. Luo, F. Wu, and W. Zeng, “Swift: A hybrid digital-analog scheme for low-delay transmission of mobile stereo video,” in Proc. 18th ACM Int. Conf. Modeling, Anal. Simulation Wireless Mobile Syst. (MSWiM), Cancun, Mexico, Nov. 2015, pp. 327–336.
[12] R. Zhang, S. L. Regunathan, and K. Rose, “Video coding with optimal inter/intra-mode switching for packet loss resilience,” IEEE J. Sel. Areas Commun., vol. 18, no. 6, pp. 966–976, Jun. 2000.
[13] D. Zhang and J. Liang, “View synthesis distortion estimation with a graphical model and recursive calculation of probability distribution,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 5, pp. 827–840, May 2015.
[14] J. Jin, J. Liang, Y. Zhao, C. Lin, C. Yao, and A. Wang, “A depth-bin-based graphical model for fast view synthesis distortion estimation,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 6, pp. 1754–1766, Jun. 2019.
[15] J. Zhang, A. Wang, J. Liang, H. Wang, S. Li, and X. Zhang, “Distortion estimation-based adaptive power allocation for hybrid digital–analogue video transmission,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 6, pp. 1806–1818, Jun. 2019.
[16] L. Luo, T. Yang, C. Zhu, Z. Jin, and S. Tang, “Joint texture/depth/power allocation for 3-D video SoftCast,” IEEE Trans. Multimedia, vol. 21, no. 12, pp. 2973–2984, Dec. 2019.
[17] T. Yang, L. Luo, C. Zhu, and S. Tang, “Block-DCT based optimization for wireless SoftCast of depth map,” IEEE Access, vol. 7, pp. 29484–29494, 2019.
[18] Y. Hu, H. Li, and W. Li, “Wireless cooperative video coding using a hybrid digital–analogue scheme,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 3, pp. 436–450, Mar. 2015.
[19] H. Shum and S. B. Kang, “Review of image-based rendering techniques,” Proc. SPIE, vol. 4067, pp. 2–13, May 2000.
[20] P. L. D. Tian, P.-L. Lai, and C. Gomila, “View synthesis techniques for 3D video,” Proc. SPIE, vol. 7443, Sep. 2009, Art. no. 74430T.
[21] C. Yao, J. Xiao, T. Tillo, Y. Zhao, C. Lin, and H. Bai, “Depth map down-sampling and coding based on synthesized view distortion,” IEEE Trans. Multimedia, vol. 18, no. 10, pp. 2015–2022, Oct. 2016.
[22] J. Jin, Y. Zhao, C. Lin, and A. Wang, “An accurate and efficient nonlinear depth quantization scheme,” in Proc. Pacific Rim Conf. Multimedia. New York, NY, USA: Springer, 2015, pp. 390–399.
[23] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, “3D video and free viewpoint Video–technologies, applications and MPEG standards,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2006, pp. 2161–2164.
[24] P. G. I. Glover, Digital Communications. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.
[25] X. Media. xiph.org. Accessed: Jan. 5, 2020. [Online]. Available: http://media.xiph.org/video/derf/
JING ZHANG received the B.S. degree from the North University of China, in 2014, and the M.E. degree from the Taiyuan University of Science and Technology, Taiyuan, China, in 2018, where he is currently pursuing the Ph.D. degree with the Institute of Digital Media and Communication. His research interests include affective computing, machine learning, pattern recognition, and image processing.

JIE LIANG (Senior Member, IEEE) received the B.E. and M.E. degrees from Xi’an Jiaotong University, China, in 1992 and 1995, respectively, the M.E. degree from the National University of Singapore (NUS), in 1998, and the Ph.D. degree from Johns Hopkins University, Baltimore, MA, USA, in 2003.

Since May 2004, he has been with the School of Engineering Science, Simon Fraser University, Canada, where he is currently a Professor. He has served as the Associate Director of the School, from 2015 to 2018. In 2012, he visited the University of Erlangen-Nuremberg, Germany, as an Alexander von Humboldt Research Fellow. From 2003 to 2004, he worked with the Video Codec Group of Microsoft Digital Media Division. From 1997 to 1999, he was with Hewlett-Packard Singapore and the Center for Wireless Communications (now part of the Institute for Infocomm Research), NUS. His research interests include image and video processing, computer vision, and machine learning. He has also served as a member of the IEEE Multimedia Systems and Applications (MSA) Technical Committee and the IEEE Multimedia Signal Processing (MMSP) Technical Committee. He received the 2014 IEEE TCSVT Best Associate Editor Award, the 2014 SFU Dean of Graduate Studies Award for Excellence in Leadership, and the 2015 Canada NSERC Discovery Accelerator Supplements (DAS) Award. He is a registered Professional Engineer in British Columbia. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP) and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT).

ANHONG WANG (Member, IEEE) received the B.S. and M.S. degrees from the Taiyuan University of Science and Technology, China, in 1994 and 2002, respectively, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), in 2009. She became an Associate Professor with TYUST, in 2005, where she became a Professor, in 2009. She is currently the Director of the Institute of Digital Media and Communication, Taiyuan University of Science and Technology. Her research interests include image and video coding and secret image sharing.

YUN GUAN received the B.S. degree from the Taiyuan University of Science and Technology, Taiyuan, China, in 2018. She is currently pursuing the M.E. degree with the North University of China (NUC), Taiyuan.

Her research interests include deep learning and medical image.