ABSTRACT
The introduction of multiple viewpoints inevitably increases the bitrates to store and transmit video scenes. To reduce the compressed bitrates, researchers have developed to skip intermediate viewpoints during compression and delivery, and finally reconstruct them with Side Information (SI). Generally, the depth maps can be utilized to construct SI; however, it shows inferior performance with inaccurate reconstruction or high bitrates. In this paper, we propose a multi-view video coding based on SI of Generative Adversarial Network (GAN). At the encoder, we construct a spatio-temporal Epipolar Plane Image (EPI) and further utilize convolutional network to extract the latent code of GAN as SI; while at the decoder side, we combine the SI and adjacent viewpoints to reconstruct intermediate views by the generator of GAN. In particular, we set a joint encoder constraint of reconstruction cost and SI entropy, in order to achieve an optimal tradeoff between reconstruction quality and bitrate overhead. Experiments show a significantly improved Rate-Distortion (RD) performance compared with the state-of-the-art methods.

KEYWORDS
Multi-view video coding, generative adversarial network, latent code learning, epipolar plane image

1 INTRODUCTION
To provide more immersive experience, multi-view video captures visual information from different positions and angles, and thereby leading a surge in the amount of data. How to reduce the coding bitrates while ensuring the reconstruction quality has become a critical issue. Recent efforts have confirmed the feasibility of deep learning-based video coding [23, 24, 37, 38]. This is benefited from the training of large datasets coupled with the powerful nonlinear modeling capability of neural networks. Unfortunately, little research has been conducted on deep learning-based Multi-view Video Coding (MVC) [30], which is still an open problem.

Traditional MVC methods utilize the hybrid coding framework to encode each viewpoint. To further reduce the output bitrate, a feasible approach is to skip intermediate viewpoints at encoder side and reconstruct them at decoder side. To this aim, small amount of information, which we call Side Information (SI), is introduced to extract the features of intermediate viewpoint for compensation of skipped information. Recently, the depth values of 3D scenes are often employed as SI to synthesize virtual viewpoint images by depth-image-based rendering [31]. This is the currently popular method called Multi-view plus Depth (MVD) [26]. However, the depth information cannot be accurately recovered due to the object occlusion and deformation problems. To improve the reconstruction quality, the low-resolution images are utilized as the SI in [10]. However, the compressed low-resolution images also lead to a high bitrate overhead, which limits its application. To address this issue, we propose to extract the latent codes of intermediate virtual viewpoints with Generative Adversarial Network (GAN) [21], which reduces bitrate overhead while maintaining the reconstruction quality of videos.

It is commonly known that neural networks can extract high-level semantic features, while GAN succeeds in generating images according to prior knowledge of sample datasets. Therefore, we deploy a GAN-based viewpoint reconstruction at decoder side. Meanwhile, an attempt is made to apply the latent vectors of GAN as SI. Previous studies suggest latent codes have contributed to restore original images [4]. However, the conventional GAN approaches do not provide an inverse mapping to project an image back into latent space. Recent years have witnessed a series of methods to extract latent codes, including gradient descent [5] and adversarial feature learning [6]. Despite of these great efforts, they cannot be applied in our task due to lack of tradeoff between SI bitrates and reconstruction quality. As depicted in Fig. 1, we address the above issue with a compact latent code and a bitrate optimization, which are capable of improving reconstruction accuracy with reduced bitrates.

A linear relationship between SI and original signal is usually postulated to simplify the calculation process, but also leads to inaccurate reconstruction in complex scenarios. In our method, we achieve the latent code learning by extracting representative features of GAN, which can recover more details with high-level understanding across viewpoints. The optimal SI is then determined with a hybrid cost function of reconstruction quality and
bitrates. On one hand, the latent code is effectively extracted to accurately reconstruct the intermediate viewpoint; on the other hand, its overhead is eliminated to reduce bitrates as much as possible. In summary, the contributions of this paper are as follows:

- We make the first attempt to introduce GAN coding network to reconstruct the intermediate viewpoint of MVC. The proposed method is able to overcome the shortcomings of conventional methods when processing complex scenarios.
- We propose a multi-view video learning to establish the correlation between latent codes and reconstructed viewpoints. Experimental results show the effectiveness of proposed method.
- We achieve a tradeoff between reconstruct quality and bitrates by a bitrate optimization cost function. Experimental results show an improved Rate-Distortion (RD) performance of MVC.

2 RELATED WORK

This section focuses on current multi-view coding approaches and analyzes their shortcomings, including the MVC and deep-learning based MVC works.

2.1 Multi-view video coding

Multi-view video coding adds inter-view prediction to the standard of High Efficiency Video Coding (HEVC). It also introduces the concept of depth map, in which each viewpoint has an additional depth video. Therefore, we divide multi-view video coding into two categories: general multi-view video coding and depth map-based multi-view video coding.

General multi-view video coding. MV-HEVC is the state-of-the-art standard, which inspires many improvements on its modules. Hannuksela et al. [8] made a stage summary of multi-view extensions for HEVC and described the standard practices for multi-view video coding, which sets a milestone for the future work. To address the problem of inefficient motion vector prediction, Lee et al. [15] recommended an intra-frame motion vector prediction based on the geometry interrelationship between two adjacent viewpoints derived from epipolar geometry, similarity and affine transformation. Unlike the traditional RD model, Li et al. [18] proposed a multi-view bit allocation method based on the exact target bit relationship between base view and dependent view. To reduce coding complexity, Jiang et al. [13] put forward inter-frame prediction method that reveals the relationship between mode selection and coding distortion threshold by a perceptual distortion threshold model. Li et al. [19] proposed an RD optimization for the dependent viewpoint based on inter-viewpoint dependencies, and greatly improved its performance in MVC. In these general methods, the bitrate may increase sharply with the number of viewpoints, which is because the original video needs to be encoded in each viewpoint rather than the side information.

Depth map-based multi-view video coding. Various coding methods have been proposed for depth map sequences from different aspects such as RD optimization, enhancement, bit allocation, and virtual view synthesis for depth maps. Müller et al. [27] improved motion compensation module to encode depth map sequences, and thus proposed an extended method for depth map HEVC based on inter-view prediction. By synthesizing the intermediate view with depth map and adjacent views, this method greatly saved bitrate and thus set a major milestone in the development of MVC. To address the problem of degraded quality at boundaries of synthesized viewpoints, Rahaman et al. [29] used Gaussian Mixture Models (GMM) to separate the foreground and fills holes in synthesized views. Besides, the amount of data transmitted can be further reduced by frame interpolation. Considering the application of depth map in intermediate view construction, a feasible method to improve MVC is to obtain accurate depth maps. Yang et al. [35] recommended a cross-view multi-lateral filtering scheme, which enhances the quality of depth map using color and depth priors from adjacent views at different time-slots. To reduce the complexity in coding mode selection, Zhang et al. [39] proposed an efficient MVD scheme based on depth histogram projection and allowable depth distortion. Lin et al. [28] proposed to accelerate 3D-HEVC deep intra-frame coding using the characteristics of the human visual system. For the above methods, their qualities are limited by the quality of the depth map.

2.2 Deep learning-based MVC

Deep learning have been introduced into MVC with significantly improved performances. These works include deep learning-based MVC optimization and deep learning-based MVC post-processing. To best of our knowledge, there are no end-to-end deep MVC codec developed and thus they are not discussed in the following.

Deep learning-based MVC optimization. The deep learning-based MVC optimization approach introduces deep learning into specific modules of MVC framework. Jia et al. [12] combined GAN with traditional coding framework to synthesize high-quality view and improve the coding efficiency. By using the neural networks, virtual frames are synthesized as additional reference for the designed hierarchical coding structure. Lei et al. [17] put forward a deep reference frame generation method for MVC, which converts the parallax between different viewpoints through parallax-guided generation network. Lei et al. [16] exploited spatial, temporal, and inter-view correlations and proposed a deep multi-domain prediction for 3D video coding. By employing CNNs to fuse multi-domain references, they achieved significant bitrate savings compared to 3D-HEVC. Liu et al. [22] proposed a CNN-based fast deep intra-frame coding to reduce the complexity of 3D-HEVC, which effectively reduced the intra-frame coding time while guaranteeing the coding performance.

Deep learning-based MVC post-processing. Deep learning-based methods are applied to the post-processing stage of the framework, which not only enhances quality of multi-view videos but also effectively removes compression artifacts. Recently, multi-frame quality enhancement approaches [7, 33, 34] have been proposed. They significantly reduce quality fluctuations between compressed video frames by locating peak-quality frames and enhancing low quality frames with adjacent high quality frames. Zhu et al. [41] recommended a view synthesis enhancement for 3D-HEVC, in which artifact removal is considered as an image recovery task to reconstruct the lossless synthesized images. Jammal et al. [11] put forward a multi-view quality enhancement that learns the mapping relationship between low- and high-quality views directly without
To facilitate the use of cross-viewpoint correlation, we employ the Epipolar Plane Image (EPI) method [32]. On epipolar plane, an object projected into different viewpoints will appear on the same straight line of EPI. By using the approach, we aggregate the matching data from different viewpoints to the spatially adjacent positions in EPI, so that it is feasible to use the correlation between viewpoints for further processing. Then, the EPI is temporarily connected as the input of convolutional network to extract latent codes.

We model the latent code extraction as a dimensionality reduction of EPI:

\[ Z = E(X), \quad (1) \]

where \( X \) and \( Z \) represent the original EPI and its low dimensional version, respectively. \( E \) represents an encoder to reduce the dimension of original EPI. To further reduce the bitrate of features, we employ a quantization \( q \) at \( Z \):

\[ \hat{Z} = q(Z). \quad (2) \]

At the decoder side, we recover the EPI information with \( \hat{Z} \) and a generator \( G \).

Our task is to find an \( E \) and and its inverse operation \( G \) to minimize the reconstruction error of EPI:

\[ \hat{E}, \hat{G} = \arg \min_{E, G} \| X - G(q(E(X))) \|. \quad (3) \]

In this work, we achieve latent code extraction and EPI reconstruction by CNN and GAN, respectively, where the two networks are jointly trained. As shown in Fig. 2, the whole framework consists of spatio-temporal EPI construction, latent code extraction and GAN-based EPI reconstruction, which are elaborated as follows.

### 3.2 Spatio-temporal EPI construction

Traditional EPI with pixel rows of images cannot well reflect the spatio-temporal correlations of multi-view images. In this work, we construct a spatio-temporal EPI with the following two steps. Firstly, the multi-view images are decomposed and reassembled based on their spatial locations. Let \( M, N \) and \( K \) denote the width, height and number of views of a multi-view video, respectively.
shown in Fig. 3, images of each view are divided into $8 \times N$ strips with all color channels, where the value 8 is chosen empirically\cite{2} and thus an image consists of $m = M/8$ strips. Then, all strips at the same spatial locations are grouped to formulate a spatial EPI with a dimension of $8K \times N \times 3$. In total, there are $m$ spatial EPIs at a same time. Secondly, a spatio-temporal EPI is obtained by stacking $L$ successive spatial EPIs at time axis, in order to embed the temporal correlations. A spatio-temporal EPI is then with the dimension of $8K \times N \times 3L$. In particular, we set $L = 3$ in this work.

In the following, we refer a spatio-temporal EPI as $EPI_j^m$, where $j = 1, 2, \ldots, m$ and $t$ denotes the temporal index. The spatio-temporal EPI is then fed into convolutional layers to extract latent code.

### 3.3 Latent code extraction

As shown in Fig. 2, the latent code extraction step compresses the spatio-temporal EPI to a latent vector $Z$, which is critical to reconstruct the intermediate viewpoint. In this work, we achieve this with Fully Convolutional Network (FCN) for its high capability of learning data distribution with a compact feature size. With input images at different resolutions, the network can extract various features as the latent code which is further utilized to reconstruct intermediate viewpoint with high visual quality. We also empirically adjust the kernel size and stride of convolution in FCN, in order to extract more accurate latent code to represent the original image.

For the encoder network, its input and output are the spatio-temporal EPIs and latent codes of intermediate viewpoints, respectively. The encoder network includes 1 convolutional layer, 4 residual blocks and another 2 convolutional layers. Each residual block includes 2 convolutional layers, 2 Batch Normalization (BN) layers, a Rectified Linear Unit (ReLU) function and an elementwise sum. The residual blocks are connected with skip connections and elementwise sum, in order to ensemble diverse feature information that benefits the visual quality of viewpoint reconstruction. All convolutions except the last employ a convolution stride of 3. As a result, the final output of encoder is with a dimension of $8K/3 \times N/3 \times 3L$.

The bitrates of latent code are further reduced by a uniform quantizer $q$. Let the number of levels be $T$, then $Z$ is quantized into $T$-level $\{c_1, \ldots, c_T\} \subset \mathbb{R}$. Statistical results show a range of [-1,1] of latent code before quantization and experimental results indicates an optimal quantization stepsize of $2/(9 \times 10^5)$. Therefore, we set $T = 9 \times 10^4$ in this work. At the receiver end, the latent code is then reobtained by an inverse quantization. Given a quantized level $c_j$, the reconstructed latent code can be easily calculated as:

$$\hat{Z}_i = q(Z_i) = \arg \min_{Z_i} \|Z_i - c_j\|^2.$$  

(4)

For ease of backpropagation during training, Equation (4) should be differentiable. Here we utilize a relaxation of $q$ by replacing Equation (4) with a differentiable softening formula:

$$\hat{Z}_i \approx \sum_{j=1}^{T} \frac{e^{-\sigma\|Z_i - c_j\|^2}}{\sum_{l=1}^{T} e^{-\sigma\|Z_i - c_j\|^2}} c_j,$$  

(5)

where $\sigma > 0$ indicates a distance weight parameter that a higher $\sigma$ indicates a more accurate approximation of $\hat{Z}$.

### 3.4 GAN-based EPI reconstruction

By using a compact representation of EPI features, we provide multi-view video at a lower bitrate. At the receiver side, we reconstruct the generated EPI $\hat{X}$ from $Z, \hat{X} = G(Z)$, where $G$ denotes the function of reconstruction inverse process. To ensure that the reconstructed EPI $\hat{X}$ is as close as possible to the original EPI $X$, we refer to the GAN framework to introduce the discrimination function $D$. Through the interaction between $D$ and $G$, the EPI generated by $G$ can increasingly approximate the original EPI $X$. As shown in Fig. 2, the EPI reconstruction is composed of three modules: reconstruction, discrimination, and interaction.

The reconstruction module uses a neural network with deconvolution as the generator. We consider the following two distributions: Joint probability density function in latent code extraction $p(Z, X) = p(X)p(Z|X)$; Joint probability density function in reconstruction $p(X, Z) = p(Z)p(X|Z)$. In these distributions, $p(X)$ is the prior probability function of the original EPI, and $p(Z)$ is the probability density function of the latent code. In the latent code extraction process, the encoder network $E$ maps the original EPI $X$ to the latent code $Z: Z = E(X)$, while in the reconstruction process, the generator network $G$ maps the samples of the prior $p(Z)$ to the input space $X = G(Z)$. To accurately reconstruct the EPI, it is necessary to make the conditional probability $p(X|Z)$ coincide with the prior probability $p(X)$ as much as possible.

The discrimination module determines whether the input EPI is the original EPI by a classification network. To better discriminate whether the input image is the original EPI, the training goal is to make the value of $D(X, E(X))$ as large as possible and the value of $D(G(Z), Z)$ as small as possible. Unlike the GAN network which only discriminates the input image, our discriminator network needs to discriminate both the image and the latent code.

The interaction module uses the discrimination results to guide the generator to reconstruct images closer to the original EPI, and also to navigate the discriminator to better identify differences between the generator’s output and the original EPI. To this end, we design this mechanism with an adversarial game in which the discriminator and the generator are trained alternately. The discriminator is trained to distinguish between sample pairs from the encoder ($X, \hat{Z} = E(X)$) and sample pairs from the generator.
where the distance weight parameter defined in Section 3.3.

3.6 The overall algorithm

With the loss function defined in Section 3.5, we are able to train the whole network consisting of the encoder $E$, the generator $G$ and the discriminator $D$. In particular, the training set is generated with 5 typical multi-view sequences including Scence_Door_Flowers, Scence_Leaving_Laptop, Champagne_Tower (1280x960) and Dog (1280x960). Each of them is with 5 views. They are further split into EPI images, and the discriminator $D$ is deployed to identify the reconstruction performance. Inspired by GAN, the objective function of whole network can be expressed as:

$$
\min_{E, G} \max_{D} \left[ \log(D(X, Z)) + \log(1 - D(G(Z), \hat{Z})) \right],
$$

(6)

where $p_X$ and $p_Z$ denote the distributions of spatio-temporal EPI $X$ and latent code $Z$ respectively.

However, this error function does not consider the scenario of MVC, where its objective can be easily achieved by deliver all intermediate information. To facilitate the utility of encoder, we should also introduce an encoder loss $l_{eq}$. In video coding, two performances indexes are critical: the distortion and the bitrate. Correspondingly, we also design the encoder loss as a weighted combination of distortion and bitrate losses, $l_{en} = \alpha l_d + \beta l_r$, where $\alpha$ and $\beta$ are coefficients. By adding the encoder loss, the Equation (6) can be rewritten as:

$$
\min_{E, G} \max_{D} \left[ \log(D(X, Z)) + \log(1 - D(G(Z), \hat{Z})) \right] + \alpha l_d + \beta l_r.
$$

(7)

In this work, we set $\alpha = 1$ and $\beta = 10^{-6}$ empirically.

We calculate the distortion and bitrate losses based on distance and entropy, respectively. The distortion loss is calculated between the original and reconstructed spatio-temporal EPIs,

$$
l_d = d(X, G(q(E(X)))),
$$

(8)

where the distance $d(\cdot)$ is obtained as a combination of pixel-domain Mean Squared Error (MSE) and feature-domain VGG loss,

$$
d(x, y) = l_{MSE}(x, y) + l_{VGG}(x, y)
= \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} (x_{i,j} - y_{i,j})^2 + \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} (\phi(x_{i,j}) - \phi(y_{i,j}))^2.
$$

(9)

Here $w$ and $h$ represent the dimension of EPI, $\phi(\cdot)$ denotes the operation of the VGG network to extract the feature map.

The bitrate loss is calculated as the entropy of quantized latent code that is delivered from sender to receiver end:

$$
l_r = H(\hat{Z}).
$$

(10)

As discussed in Section 3.3, the quantized latent code is represented by $T$ discrete levels $c_j$, $j = 1, 2, \ldots, T$. Therefore, the entropy can be further expressed by

$$
H(\hat{Z}) = H(c) = -\sum_{j=1}^{T} p_{c_j} \log p_{c_j},
$$

(11)

where $p_{c_j}$ represents the posterior probability of all levels. For ease of network training, we calculate $p_{c_j}$ as the following approximations:

$$
p_{c_j} = \frac{1}{n} \sum_{i=1}^{n} \exp(-\sigma \|Z_i - c_j\|^2),
$$

(12)

where $\sigma$ is the distance weight parameter defined in Section 3.3.

4 EXPERIMENTAL RESULTS

4.1 Experimental setup

We examine our method, which is named as MVLL for the sake of simplicity, by the recommendations and popular sequences from the Common Test Conditions (CTC) of MVC [28]. These sequences include the Balloons (1024x768), Book_Arrival (1024x768), Kendo...
There are three other methods implemented for performance comparison. The hybrid low-resolution multi-view video coding method, MRMV [25], uses the low-resolution image as the auxiliary information. The images of even views are downsampled and further reconstructed during the encoding and decoding processes. The multi-view video plus depth standard method, MVD [30], uses the depth maps as the auxiliary information. It does not work when the depth of a view is unavailable. The deep learning-based video quality enhancement method, MFQE 2.0 [7], also introduces an image reconstruction method with deep learning. For fair comparison, we implement all these methods with their pre-trained models and parameters.

The popular video coding criteria are employed to evaluate and compare all the above four methods. These criteria include the PSNR-aware measurements (rate-PSNR curves, the BDPSNR and BDBR [1]) and the SSIM-aware measurements (rate-SSIM curves, the ADSSIM and ADBR [40]). Four Quantization parameters (Qps), namely 27, 32, 37 and 42, are employed in original encoder to formulate the curves.

### 4.2 Objective Comparison

Fig. 4 shows the rate-PSNR curves of all compared methods: MRMV, MVD, MFQE 2.0 and MVLL. In each curve, the four data samples correspond to the four Qp settings; while in each data sample, the rate-PSNR performances are obtained as the averaged results of all views. For sequences (g)-(l), the results of MVD are not provided because the depth maps of these sequences are unfortunately inaccessible. Nevertheless, the compression performances of the other sequences can also reveal the effectiveness of this method to some extent. Fig. 5 shows the rate-SSIM curves of these methods, where the SSIM index is considered is more consistent with human vision system than PSNR.

From Figs. 4-5 we can get several conclusions. Firstly, all methods achieve good performances in terms of R-Q curves, by retaining a high compression ratios at the acceptable visual quality. Generally speaking, they are still superior to the original encoder without optimizations. Secondly, the MRMV methods achieve superior performance than MVD in most sequences, which may be due to the extra bitrates for depth maps in MVD. In case of multi-view plus depth coding, the MVD method is still preferred. Thirdly, the MFQE 2.0 method achieves acceptable performance but is still inferior to our MVLL. This method was designed for general video coding and thus do not exploit the inter-view correlations between multi-view videos. Fourthly, our MVLL method outperforms the compared...
methods in most sequences, which is attributed to its GAN-based inter-view prediction. An exception is the 
Shark sequence, where our method achieves comparable performance in terms of SSIM and inferior performance in terms of PSNR. The sequence is an animation video, while as indicated in Section 3.6, our MVLL is trained with natural scenes. The failure case indicates the limitation of learning-based methods. However, considering the natural video sequences, our method still outperforms the state-of-the-art performance.

Tables 1 and 2 presents the quantitative comparisons of R-Q performances, where the efficiency of our MVLL is evaluated with MRMV, MVD and MFQE 2.0 as the benchmarks. The terms BDBR and ADBR indicate the average bitrate savings at the same PSNR and SSIM, respectively. The terms BDPNSR and ADSSIM indicate the average quality increments at the same bitrate, respectively. From the tables, we can see the quantitative comparison results are consistent with those in Figs. 4-5. Compared with the other methods, our MVLL reduces 26.93%, 54.56% and 52.65% bitrates at
the same PSNR, or 45.42%, 47.01% and 52.91% bitrates at the same SSIM. It also significantly improves the video quality in terms of PSNR and SSIM at the same bitrate. These results demonstrate the efficiency and effectiveness of our proposed method.

4.3 Subjective Comparison

To show the generated videos more intuitively, we present the typical images of Balloons, Book_Arrival, Kendo, Lovebird1 and Newspaper in Fig. 6. For each video sequence, all methods generate the video frames at similar bitrates for fair comparison.

From Fig. 6, our synthesized images are more visually pleasing since the images remain more details and contain less noises. As a comparison, the blurred regions with artifacts or noticeable pixel errors can be found in the reconstructed images of other methods. We highlight these local blocks with colored boxes and separate patches for a clear vision of these differences. The above results show that the MVLL leads to better synthesized view quality than other compared approaches at the same bitrate, which is consistent with the objective comparison shown in Section 4.2. This also reveals the superiority of our method from another point of view.

5 CONCLUSION

Nowadays, there exists a bottleneck to further compress video streams with the traditional hybrid model. Researchers have been contributing to deep-learning-based video compression, where the motion prediction/compensation, RD optimization or entropy coding is realized by deep network. In this paper, we made the first attempt to combine deep GAN model with multi-view video coding. We utilized the latent code of GAN as SI in an RD-optimal manner. The latent codes are generated with a deep network and further utilized to reconstruct the intermediate views, thereby saving the streaming bitrates of multi-view videos. Experimental results show a significant performance gain over the state-of-the-art schemes using either depth map or low-resolution images as SI. We hope this work can provide an innovative methodology to deep-learning-based multi-view video coding.
