FICNet: An End to End Network for Free-View Image Coding

Chunhui Yang, Jiayu Yang, Yongqi Zhai, and Ronggang Wang, *Member, IEEE*

***Abstract*— Free-view image compression has attracted the gaze of people due to the rapid development of 3D vision applications. However, as far as we know, no end-to-end learned compression model is proposed for free-view image sequences. Most existing learned compression models are limited and only applicable to image sequences with simple horizontal and vertical translations, such as stereo and light field image compression models. In this paper, we first propose an end-to-end network FICNet to improve free-view image compression performance, effectively eliminating the spatial redundancy among multiple views. In our methods, a differentiable depth prediction module is introduced to our model for exploring spatial correlation and achieving end-to-end training. Besides, we demonstrate a strategy of multi-view reference to alleviate the hole problem in depth-based prediction, and a filter network is designed to improve the prediction accuracy further. A residual fusion network with multi-level complementary features is also utilized to enhance the reconstruction quality. Extensive experiments show that our model can perform favorably in generating more refined predictive images and achieves up to a 16.23% BD-rate improvement compared to the state-of-the-art method 3D-HEVC.**

***Index Terms*— Free-view image, image compression, depth-based prediction, multi-view reference, residual fusion network.**

I. Introduction

CURRENTLY, the development of immersive visual systems is unprecedented, bringing people a better visual experience, such as developing and applying IMAX theaters, 3D exhibition halls, and 3D games. However, immersive vision systems are still not widely spread due to storage and transmission speed limitations. For this reason, people pay attention to the compression of free-view images for the purpose of enhancing the practicality of immersive systems by improving their performance of free-view compression. Compared with single-image compression, multi-image compression focuses more on eliminating the spatial redundancy among views to save bit rate. Meanwhile, free-view image compression differs from stereo image and light field image compression because of its complex position transformation. The stereo image sequence is composed of a left-eye view and a right-eye view image from the human. Therefore, the disparity between the two views is mainly caused by the horizontal displacement between the human eyes, and the spatial correlation on the horizontal direction between the stereo images is close. As for the light field images, they consist of a two-dimensional multi-view image array collected by the light field camera or multi-camera system. The disparity between light field images is mainly from the different positions of the camera or camera lens arrays, resulting in the disparity between light field images primarily focusing on horizontal and vertical direction. When it comes to generalized free-view images, they are captured by a 6-degree-of-freedom (6-DOF) camera system, which determines that the disparity in complex directions exists between free-view images and the changes between views are also very flexible compared to the other two. Because the number of views is numerous and the transformation is complex among views, the spatial correlation of free-view image sequences is more difficult to exploit and the images are more laborious to compress at the low bit rate.

At present, some standards have been developed for free-view compression. In traditional codecs, MV-HEVC [1] and 3D-HEVC [2] offered by the Joint Collaborative Team on 3D Video coding development (JCT-3V) are developed based on High Efficiency Video Coding (HEVC) [3] and applied for free-view image and video compression. Since the research on free-view compression mainly focuses on utilizing the spatial correlation among views, MV-HEVC introduces the spatial prediction tool between views in the prediction module compared with HEVC. The codec uses the image information provided by other compressed views to predict the current view, so the codec will perform spatial and temporal predictions when both the spatial and temporal reference frames exist. Then MV-HEVC will select the optimal prediction mode based on the predicted cost for better compression. 3D-HEVC further introduces the tool for depth-based prediction and depth coding based on MV-HEVC, so depth maps will be encoded additionally so that they can be applied for compression and stereo matching. Experiments show that MV-HEVC and 3D-HEVC can achieve satisfactory performance in free-

Manuscript received 30 May 2023; revised 13 January 2024 and 4 March 2024; accepted 2 April 2024. Date of publication 17 April 2024; date of current version 30 September 2024. This work is financially supported for Outstanding Talents Training Fund in Shenzhen, Shenzhen Science and Technology Program-Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents project (Grant No. RCJC20200714114435057), Shenzhen Science and Technology Program-Shenzhen Hong Kong joint funding project (Grant No. SGDX20211123144400001), National Natural Science Foundation of China U21B2012, R24115SG MIGU-PKU META VISION TECHNOLOGY INNOVATION LAB. This article was recommended by Associate Editor L. Zhang. *(Corresponding author: Ronggang Wang.)*

The authors are with the School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China (e-mail: rgwang@pkusz.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3390151.

Digital Object Identifier 10.1109/TCSVT.2024.3390151

view compression. However, the spatial correlation among views cannot be fully utilized and the compression performance is also limited by traditional methods.

Many works on free-view compression also have been proposed. Most of the current research on free-view compression is developed on the framework of the MV-HEVC and 3D-HEVC. Some works focus on improving the basic codec with the traditional methods [4], [5], [6], including optimizing the prediction or transformation method and adding pre-processing or post-processing for free-view sequences. With these methods, the codec can achieve a better compression effect. Some people also introduce the neural network to replace a specific module in the classical codec [7], [8], which improves the compression performance of free-view sequences. These methods can enhance the input images, the predicted images, and the reconstructed images by deep learning so that the enhanced images become more suitable for encoding and reconstruction. In addition, some end-to-end learned frameworks are also proposed for application in stereo image and light field image compression [9], [10], [11]. These frameworks make full use of the spatial correlation among views to realize simple multi-image compression. They can predict simple disparities with neural networks. However, learned free-view compression models are designed for a large number of views that involve extensive spatial transformations and complex spatial correlations. Therefore, the prediction modules with simple neural networks in stereo and light field models struggle with accurate predictions on free-view sequences. Due to this reason, the end-to-end learning framework has not yet been proposed for free-view sequences.

In this paper, FICNet is the first end-to-end learned framework for free-view image compression, as far as we know, to solve the problem of spatial redundancy among views. In our model, we divide the views into key and reference views referring to the traditional codec. For the key views, we design the main coding network to obtain high-quality reconstructed images. Besides, we also devise a residual coding network for reference views to achieve high-quality compression at a low bite rate. Since the prediction method based on optical flow cannot obtain satisfactory results for the compression of free-view sequences, a differentiable prediction based on depth information is introduced into our model to achieve more precise prediction and end-to-end training. However, the problem of holes will appear in depth-based prediction caused by occlusion, so the strategy of multi-view reference is devised to fill holes. With the multi-view reference, we also utilize a filter network to enhance the predicted image and mitigate local distortion issues caused by inaccurate original depth values. In addition, simple additive operation in traditional residual coding also distorts the reconstructed image, and we adopt a residual fusion network to fuse residual and predicted images for better reconstruction adaptively. With the proposed methods, FICNet performs excellently on the peak signal-to-noise ratio (PSNR) and multi-scale structural similarity (MS-SSIM). The contributions of this paper are as follows,

- FICNet is the first learned compression model for free-view image sequences, which combines the advantages of traditional residual coding and neural networks to enhance compression performance.
- We design a novel prediction component with methods of the differentiable depth-based prediction and multi-view reference, which improves the prediction accuracy.
- The residual fusion network with multi-level complementary features is designed to integrate the residual and the prediction images during the residual coding of reference views, which improves the quality of the reconstruction.
- Our model achieves high-quality prediction and outperforms existing state-of-the-art free-view image compression methods.

The rest of this paper is organized as follows: Section II systematically introduces the proposed related work on image compression. Section III describes the overview of the free-view image compression model, including the proposed strategy. Section IV presents comparative experimental results. Finally, the conclusions are given in Section V.

II. Related Work

In general, image compression methods can be divided into two categories: single-image and multi-image methods. Stereo image, light field, and free-view image methods are classified into multi-image methods due to their image properties. In single-image methods, we only need to reconstruct the image itself. However, the compression of multi-image needs to consider the similarity of the key and reference view. We will review both kinds of methods as follows.

A. Single Image Compression

For single image compression, most works mainly aim at improving objective evaluation performance, such as PSNR and MS-SSIM. In [12] and [13], the recurrent neural network (RNN) is introduced into the image compression framework, which generates the stream iteratively. More accurate rate estimation during training is critical for improving image reconstruction performance, so some works pay attention to designing a more reasonable entropy model. In [14], [15], [16], and [17], the fixed-parameter and flexible entropy models are both developed in the image compression framework based on the convolution neural network (CNN) for the rate estimation of entropy codec. In [18] and [19], discretized Gaussian mixture likelihoods and parallelizable checkerboard context models are also applied to the entropy model, which greatly saves bit rate and speeds up the decoding process. Recently, advanced priors are proposed for improving entropy estimation, including the vectorized prior [20], global and local hierarchical priors [21]. An uneven grouping model combined with the context model has also proved effective in [22]. In addition to the entropy model, someone also focuses on other aspects. Invertible neural networks are utilized to mitigate the problem of information loss in [23]. In [24] and [25], the models realize image compression on variable rates by different methods.

B. Stereo Image Compression

For the compression of the stereo image, some works have been proposed to improve the compression performance based

on deep learning. In [9], an end-to-end model is first designed for stereo image compression with the learning ability of neural networks. Based on this model, the compression quality is improved by using a learned regression model to estimate the homography matrix between left and right images in [10]. Lei et al. [26] represent a bi-directional coding structure to reduce spatial redundancy, and the method of stereo attention is introduced into the model in [27]. Moreover, Zhai et al. [28] also employ a stereo-matching model to obtain a pleasant prediction for compression. Besides, the traditional model [29] and end-to-end learned model [30] are both proposed for lossless compression.

C. Light Field Image Compression

When it comes to the compression of light field images, most of the existing research is developed on traditional codecs, such as HEVC and MV-HEVC. Many works focus on improving algorithms in traditional codecs. In [31], [32], [33], [34], [35], and [36], light field images are arranged into pseudo sequences for compression utilizing HEVC, etc. Light field images are also processed before encoding in various ways and treated as free-view sequences to compress with MV-HEVC in [37], [38], and [39]. Furthermore, some models are proposed with neural networks in components of codec. In [40], [41], and [42], the neural networks are designed for synthesizing virtual views and filtering. The adversarial training mechanism brings gains in the light field image compression model in [43], [44], and [45]. Besides, the graph-based strategy is also adopted for many modules in [46], [47], [48], and [49]. Recently end-to-end works also appear. For example, the 3D-CNN structure is employed to model in [50], which can deal with light field data more efficiently, and Singh and Rameshan [11] generate the disparity information to aid with compression.

D. Free-View Image Compression

For free-view compression, only a few works are published due to the existence of complex disparity among multiple views, and almost all methods are researched on traditional codecs, in which pre-processing is an effective strategy. In [4], the authors design a coding scheme utilizing pre-processing to explore temporal and spatial correlation among multiple views. Flierl et al. [5] also arrange the free-view data and present a new prediction module for compression. Then, the free-view data is partitioned into correlated layers characterized by depth values for compression efficiency in [6]. Moreover, an advanced inpainting method is introduced for filling holes in the inter-view prediction module to improve the prediction quality of codecs in [51] and [52]. Inspired by that, depth information is also applied in later works. In [53] and [54], the depth vector is derived for view prediction to reduce spatial redundancy, and a base-anchored model for compression of free-view images is proposed in [55]. Some existing works also adopt learning methods to improve performance. In [7], authors propose effective networks to compress selected views with the generative adversarial network (GAN), and virtual reference views are generated by neural networks for the prediction of traditional codecs in [8]. However, the above models are not complete end-to-end models without fully developing the learning ability of the neural network.

III. Proposed Framework

In this section, we will elaborate on the design of the FICNet, whose structure is shown in Fig. 1. In our model, differentiable depth-based prediction is used for more efficient prediction during encoding. Meanwhile, the strategy of multi-view reference is introduced to alleviate the hole problem resulting from occlusion in the prediction module. An in-loop filter network is also proposed to compensate for poorly predicted areas in the predicted image. Finally, we design the residual fusion network to enable the adaptive fusion of residual and prediction images to reconstruct reference views.

A. The Overview of FICNet

Our FICNet is a novelty learning model with a well-designed coding structure. As we all know, learning methods can leverage the self-learning and generative capabilities of neural networks to reconstruct more detailed information from multi-view images. Additionally, the residual coding methods are also designed to enhance the compression performance of multi-view data in traditional multi-view codecs. However, the key view and reference view are separately optimized due to the limitations of traditional methods, which restricts the improvement of its performance. Therefore, our main contribution lies in proposing the first learning end-to-end compression model, and we focus on the design of the framework rather than the optimal network structure. This model combines the advantages of traditional residual coding methods and neural networks, providing more possibilities for improving the compression performance of multi-view images.

For free-view images, spatial redundancy among views leads to a significant amount of data repetition. It is crucial to alleviate spatial redundancy between views for compression. Inspired by the traditional codecs, our coding structure also utilizes spatial correlations among views for prediction to reduce redundancy. Subsequently, our model only needs to compress the residuals after prediction for some views. The residual coding structure can eliminate redundant data among views to some extent, achieving a reduction in bit rate.

Based on the residual coding structure, we introduce three codec models for the compression of free-view images, including KV codec, RV codec, and Depth codec as shown in Fig. 1. Initially, the sequence of free-view images is divided into multiple Groups of Pictures (GOP) for compression. Views are categorized into key views v_k and reference views v_r based on their similarities in each GOP. For key views v_k, a high-performance single-image compression model is employed as the Key codec. Key views are encoded, quantized, and decoded to obtain high-quality reconstructed images v_k', which serve as references for the subsequent encoding of reference views. As for reference views v_r, a depth-based residual codec is employed to compress the views as the RV codec. With the decoded views, the depth map of views, camera intrinsics and

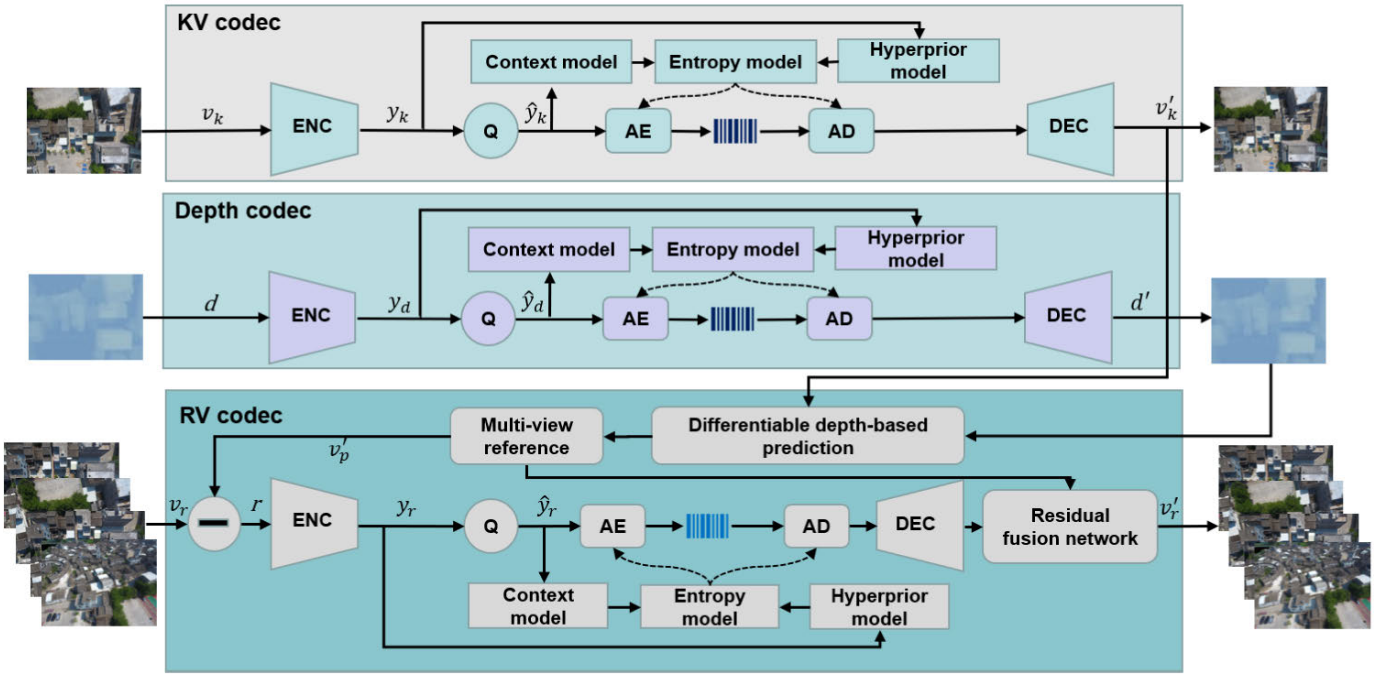

Fig. 1. The architecture of the FICNet. KV codec denotes the codec for key views while RV codec represents the codec for reference views, and the Depth codec is used for depth maps. AE and AD represent arithmetic encoder and arithmetic decoder. Q stands for round quantization. The context model and hyperprior model are used to predict parameters for the entropy model.

extrinsics of views, the prediction of the current reference view can be obtained by differentiable depth prediction. After that, the multi-view reference strategy and the filter network are combined to fill the hole and enhance predicted views. We can acquire the residual map r of the current reference view by a subtraction operation between the original image and the predicted image v'_p, and then the residual coding network is applied to compress the residual map. It is worth mentioning that we use the residual fusion network replacing the traditional additive operation during the final reconstruction of the reference view v'_r. During the encoding of reference views, the depth map d of the key view is required for prediction. Therefore, the Depth codec is utilized for the compression to obtain the reconstructed depth map d'.

Our three codecs are based on the single-image compression network [18], which includes the main branch and entropy branch. The codec network can be flexibly chosen, and we experiment with the commonly used model [18] in this paper. The main branch is composed of four components, consisting of the encoder, decoder, quantizer, and entropy codec as shown in Fig. 1. In the encoder, we utilize the classical residual blocks and convolution layers to downsample the original image and extract image information. The decoder has a symmetric structure with an encoder in order to reconstruct the image better. Furthermore, the quantizer is employed element-wisely to round the latent representation y_k, y_d, y_r to the nearest integer, $\hat{y}_i = round(y_i)$, where y_i and $\hat{y}_i$ represent elements of non-quantized latent representations y_k, y_d, y_r and quantized latent representations $\hat{y_k}, \hat{y_d}, \hat{y_r}$. The entropy branch is composed of the hyperprior network and context model. The hyperprior network also employs the autoencoder structure with two downsampling layers to estimate the hyperprior of encoded latent representation, while the context model is adopted to obtain the autoregressive prior of the current pixel with the aid of the adjacent available information $\hat{y}_{adjs}$. Inspired by [56], we model the latent representations $\hat{y}_i$ as a conditional Gaussian distribution with the mean value μ_i and standard deviation σ_i predicted by hyperprior $\hat{\mathbf{z}}$ and autoregressive prior of $\hat{y}_i$, which can be expressed as

$$p_{\hat{\mathbf{y}}}(\hat{y}_i|\hat{y}_{adjs}, \hat{\mathbf{z}}) = \prod_i (\mathcal{N}(\mu_i, {\sigma_i}^2) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{y}_i). \quad (1)$$

In addition, a fully factorized density model [16] based on the cumulative is introduced to model $\hat{\mathbf{z}}$ in hyperprior network, which is shown as

$$p_{\hat{\mathbf{z}}|\boldsymbol{\psi}}(\hat{\mathbf{z}}|\boldsymbol{\psi}) = \prod_i (p_{z_i|\psi^{(i)}}(\psi^{(i)}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{z}_i), \quad (2)$$

where $\psi^{(i)}$ denotes the parameters of each univariate distribution $p_{z_i|\psi^{(i)}}$. $\mathcal{U}$ represents the uniform distribution. Finally, the bit rates of $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ can be estimated as

$$R_{\hat{\mathbf{y}}} = -\sum_i \log_2(p_{\hat{y}_i|\hat{z}_i}(\hat{y}_i|\hat{z}_i)), \quad (3)$$

$$R_{\hat{\mathbf{z}}} = -\sum_i \log_2(p_{\hat{z}_i|\psi^{(i)}}(\hat{z}_i|\psi^{(i)})). \quad (4)$$

Based on the above structure, our KV codec network is designed with four downsampling layers and optimal hyper-parameters from practical experiments [18] to improve the compression of key views. For the RV codec, the information entropy contained in the residual map after prediction is reduced. However, the residual map still contains some important information in cases where the prediction is inaccurate. Therefore, the model with optimal configuration is

still adopted to compress the residual image for the quality of reconstructing the reference view. The residual data needs to be normalized to approximate a Gaussian distribution for the accuracy of probability distribution predictions in the entropy model. For the Depth codec, the depth image is a single-channel image containing less information, so we reduce the number of channels in the codec to meet the compression requirements of the depth. Besides, an accurate depth image is crucial for predicting reference views. For this reason, we design the encoder with only two downsampling layers to retain more depth information. The specific hyper-parameter settings are introduced in Section IV-A.

B. Differentiable Depth-Based Prediction Module

During the encoding of reference views, the prediction module is a significant component and an accurate prediction can reduce the entropy of the residual image. Since the disparity between different views in free-view image sequences is more obvious than that of video sequences, the capability of traditional prediction based on optical flow is limited and residual coding cannot work. In contrast, the prediction method based on depth information can achieve more accurate predictions. Therefore, we innovatively designed a differentiable depth-based prediction module, combining differentiable warping and neural networks to enable end-to-end training for reference view encoding.

The prediction based on depth information can be seen as projecting the pixels on the image plane of the key view onto the image plane of the reference view, known as homography transformation. Consequently, a precise predicted image can be obtained if the depth information is accurate. During the prediction, the projection transformation between two-dimensional and three-dimensional representations is involved. In order to obtain a two-dimensional representation of a three-dimensional scene, the transformation occurs in four coordinate systems, including world coordinate, camera coordinate, image coordinate, and pixel coordinate. Based on the projection as shown in Fig. 2, we can establish the transformation relationship between the three-dimensional world coordinate (X_A, Y_A, Z_A) and two-dimensional pixel coordinate (u_a, t_a) of point A as follows

$$D_a \begin{bmatrix} u_a \\ t_a \\ 1 \end{bmatrix} = H \begin{bmatrix} X_A \\ Y_A \\ Z_A \\ 1 \end{bmatrix}, \tag{5}$$

$$H = P \begin{bmatrix} R_{c\leftarrow w} & T_{c\leftarrow w} \\ 0^T & 1 \end{bmatrix}, \tag{6}$$

where matrix P represents the camera intrinsics, which are only related to the properties of the camera. $R_{c\leftarrow w}$ is a rotation matrix that is used to transform A from the world coordinate to camera coordinate, and $T_{c\leftarrow w}$ is a translation vector. Therefore, the $R_{c\leftarrow w}$ and $T_{c\leftarrow w}$ constitute the extrinsic matrix of the camera. D_a stands for the depth information of point A. Since the prediction process is based on homography to transform the texture image from the key view to the reference view as shown in Fig. 2, we first need to put point

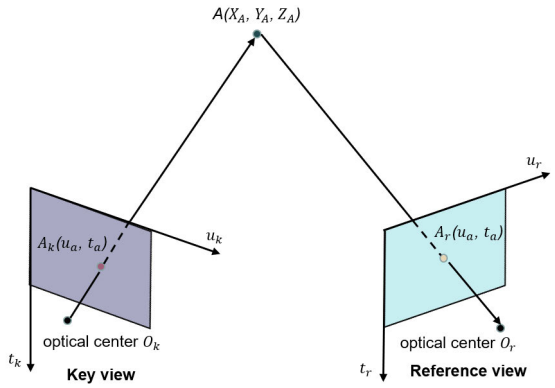

Fig. 2. Differentiable depth-based prediction process. $A_k(u_a, t_a)$ and $A_r(u_a, t_a)$ represent the two-dimensional pixel coordinates of the views, while $A(X_A, Y_A, Z_A)$ represents the three-dimensional world coordinate.

$A_k(u_a, t_a)$ of the key view to the world coordinate shared by multiple views following the inverse process of Eq. (5) with the pixel coordinate and depth in the key view. After we obtain the position of point $A(X_A, Y_A, Z_A)$ in the world coordinate, this point can be projected to the image plane of the specified reference view on the basis of the forward process of Eq. (5) to get $A_r(u_a, t_a)$. The prediction of our model is completed after we obtain the projected view.

Without loss of generality, the projection matrix H itself is a 3×3 matrix, so the homography of the key view is also a 3×3 matrix. The process of warping from the key view to the reference view is similar to the classical plane sweeping stereo [57], which utilizes the differentiable bilinear interpolation to sample pixels from the key view and project them to the reference view. From Eq. (5), it can be seen that the depth D_a is the only variable in the transformation from three-dimensional to two-dimensional representation. Therefore, the homographic transformation in the prediction is differentiable with respect to the unique variable D_a. In other domains, MVSNet [58] also utilizes the similar differentiable approach. As the core step to bridge the encoding of key views and the residual encoding of reference views, the warping operation is implemented in a differentiable way, which enables end-to-end training of the free-view image sequence. This differentiable approach enables leveraging the prediction loss from the reference views to enhance the reconstruction quality of key views during the optimization of the whole network. Therefore, our model can jointly optimize the networks of key and reference views by the reconstruction quality of the views. The method of combining differentiable homography with neural networks is innovative, allowing for joint end-to-end optimization of key views and reference views.

C. Strategy of Multi-View Reference

As we all know, it is difficult to align the reference view well by relying on optical-flow prediction due to the complex spatial transformations. With the inspiration of the view synthesis method, we can acquire the more ideal predicted image with the depth map and camera parameters to reduce the bit rate of the reference view. However, the problem remains that

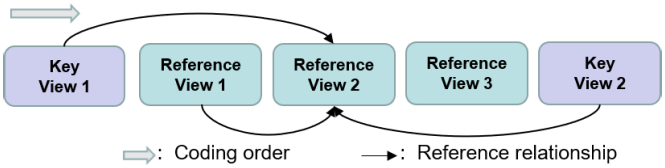

Fig. 3. The structure of the multi-view reference. Key view n represents the key view of the nth GOP. Reference view n represents the nth reference view in one GOP.

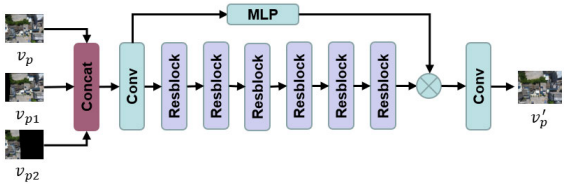

Fig. 4. The structure of the filter network. MLP stands for multilayer perceptron [59]. Resblock denotes residual block [60], and Conv represents a convolution layer with a 3 × 3 kernel.

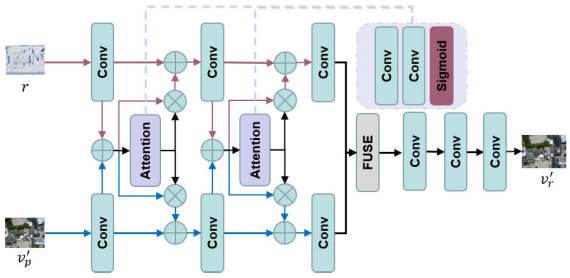

Fig. 5. The structure of residual fusion network. Conv represents a convolution layer with a 3 × 3 kernel. The attention module consists of two convolutional layers and a sigmoid layer. The FUSE represents an addition operation. ⊕ denotes element-wise addition, and ⊗ denotes element-wise multiplication. The red and blue lines represent the process of feature extraction for the residual map and the predicted map respectively.

the predicted image acquired by 3D projection suffers from severe holes created by occlusion and rounding integer error when warping. Therefore, filling holes is crucial for improving compression efficiency.

As for the problems noted above, we demonstrate the multi-view reference strategy to alleviate the issue of the hole in the prediction module. At the pre-processing stage, the free-view images are rearranged based on the similarity with the selected key view, and a reorganized free-view sequence is generated. Accordingly, we can make full use of the texture images and depth maps of three chosen views, which are respectively selected from two neighbour key views and the previous decoded view, to predict the texture image of the current view. Furthermore, we choose the predicted image with the smallest hole as the basic predicted image v_p, where we can reuse other predicted images to fill the holes in the basic map and enhance the quality. With the multi-view reference strategy shown in Fig. 3, predicted images of the reference views are enhanced and the bit rates will be greatly saved.

After the depth-based prediction with multi-view reference, a filter network is proposed to refine and enhance the predicted image. The predicted image after warping often has unsatisfactory areas, such as distorted objects and unusual textures, resulting from the existence of inaccurate original depth information and occlusion. In addition, simple hole-filling methods can result in the generation of boundaries in the predicted images with the multi-view reference strategy. Therefore, the predicted images need to be smoothed and optimized. For this reason, we build a filter network to reduce prediction error for the residual image by adaptive learning. The structure of the filter network is displayed in Fig. 4.

In the filter network, basic image v_p after hole-filling and prediction images v_{p1}, v_{p2} obtained by multiple views prediction are applied as the inputs of the filter network. To better integrate the features of the predicted images, we introduce the attention mechanism to learn the relevant information within the channels of predicted images. The attention mechanism can leverage useful information from multiple channels in the case of multi-view inputs to achieve prediction correction and hole filling. The MLP [59] is utilized to learn the weights of the predicted image channels, leveraging their correlations. Moreover, residual blocks and convolution layers are used to capture features so that the information extracted from multiple predicted images can learn from each other. Finally, we use the learned weights to guide the enhancement of the predicted images to obtain the final predicted images v'_p with high quality. In order to improve the prediction effectiveness without increasing the burden of the network obviously, our filter network adopts a lightweight structure consisting of simple residual blocks [60] and convolution layers.

D. Residual Fusion Network

In FICNet, residual coding is adopted for the compression of the reference image. Nevertheless, the traditional residual coding simply performs an additive operation with the residual and prediction images at the decoder to achieve image reconstruction. So the residual images cannot fit with prediction images in some areas when the reconstruction quality of the residual image is not pleasant. The unsatisfactory fitness will result in distortion in some areas, so our model exhibits a residual fusion network inspired by the image fusion model for the reconstruction of the reference image. The fusion network can achieve additive reconstruction of the prediction and residual image in the feature domain and utilize the generative capabilities of the network to enrich the details of views. Therefore, the multi-level complementary feature fusion module is introduced for the first time to replace the additive reconstruction in residual coding as shown in Fig. 5.

In the residual fusion network, we consider the prediction image v'_p and the residual image r as inputs and extract their image information with the previous three-layer convolutions respectively. However, the features of the prediction image contain basic structural information of the view, while the features of the residual image include detailed texture information. The significant difference in features may affect their fusion. Therefore, we design a multi-level complementary feature (MCF) method to gradually supplement the features of both in the feature extraction process. This method allows both features to preserve the image information of the view

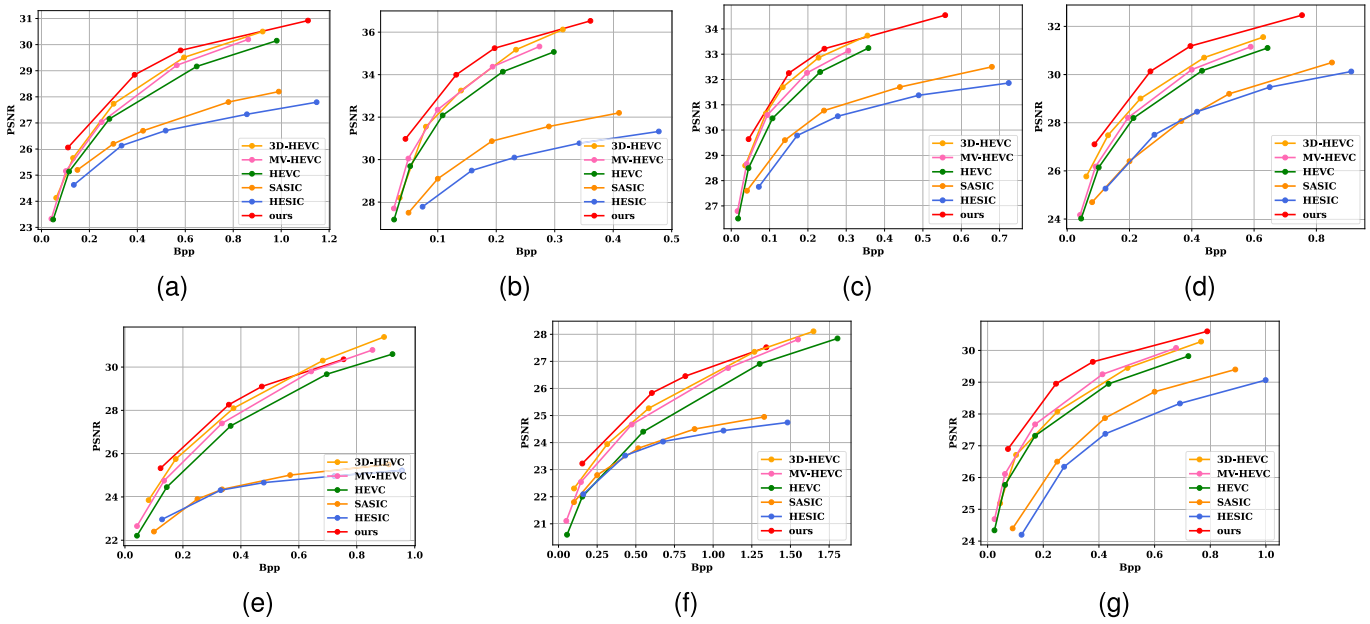

Fig. 6. Comparison of average PSNR performance for two views image compression. (a) - (g) denote different testing sequences. Bits per pixel (bpp) denotes bite rate. Our model achieves higher PSNR than other methods at most bit rates for all sequences.

effectively. Specifically, both features are added and fed into an attention module composed of convolution layers and the sigmoid function to learn spatial weights after each convolution layer. Then, we can obtain the complementary feature at each level with the weight and the fused feature. After feature extraction, we choose the method of pixel-level addition to carry out the fusion operation and obtain the fused features in the fusion module. Finally, the fused features complete the reconstruction of the reference view v_r' with the convolution layers. The reconstruction and the feature extraction part have the same number of convolution layers, for which the symmetric structure helps better recover the view. In our fusion model, we also choose convolution layers to build the network, which does not increase the number of additional parameters significantly. Experiment results indicate that the residual fusion network can achieve adaptive fusion and reconstruction with residual and prediction images, which alleviates the problem of pixel unsuitability.

E. Training Strategy

We train the whole framework in a two-stage manner. In the first stage, the rate-distortion loss is utilized to optimize the entire compression model so that the reconstruction quality of key views and reference views can attain an excellent level. The loss function is like

$$L = \lambda \times (D_k + D_r) + R_k + R_r, \tag{7}$$

$$R = R_{\hat{\mathbf{y}}} + R_{\hat{\mathbf{z}}}, \tag{8}$$

where λ is a trade-off parameter to balance rate-distortion loss, R_k and R_r denote the rate losses of key views and reference views as Eq. (8), and D_k and D_r represent distortion losses on mean square error (MSE).

As for the second stage, the compression performance of the key views is optimal, and we focus on improving the performance of the reference views by reducing the cost of the prediction module. With the loss D_p between the reference view and its predictive image, the bit rate of the reference view can be further reduced. During the training, we fix the parameters of the key view compression network and continue to optimize the parameters of other codecs. λ_2 is also a trade-off parameter. In this state, the loss function changes to

$$L = \lambda \times (D_r + \lambda_2 * D_p) + R_r. \tag{9}$$

IV. Experiment

A. Experiment Details

Following common practices, both our training dataset and test dataset are selected from BlendedMVS [61], which offers 17k high-quality training samples covering a variety of scenes for free-view tasks. We take full-size images for training and testing in the experiments, whose size is 576×768. The batch size is set to 4 because of the limitations of the device. During the training, we introduce the method of variable rate [62] into our model for realizing flexible compression, so λ group is set as 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, and λ_2 is set to 0.05 in Eq. (9). Besides, we use Adam optimization with the initial learning rate of $1e^{-4}$ and the learning rate will decrease from $1e^{-4}$ to $1e^{-5}$ after 100,000 iterations. For the reasonable compression structure, the number of channels in the Depth codec is 64, while KV and RV codecs have 192 channels. We implement the FICNet with the PyTorch framework and train our model using Tesla V100 GPUs.

B. Objective Results

To evaluate the performance of the FICNet, we utilize the PSNR and MS-SSIM to show the effectiveness. We compare our model with the traditional codecs widely used for free-view sequences, including MV-HEVC [1] and 3D-HEVC [2]

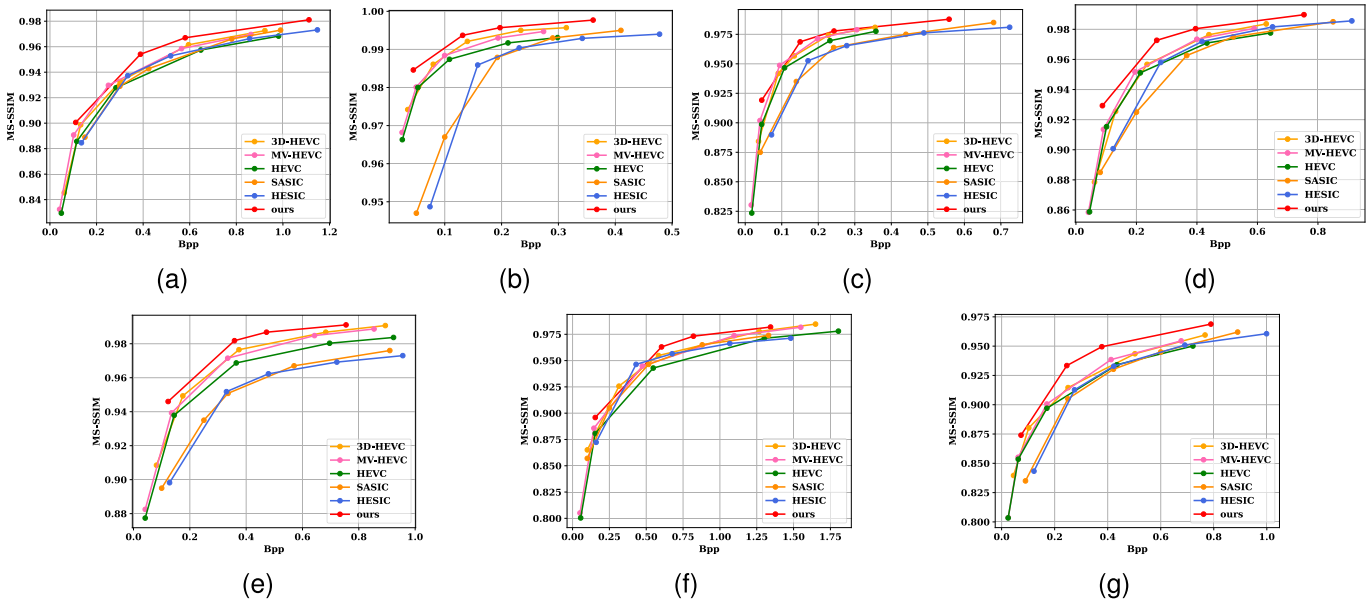

Fig. 7. Comparison of average MS-SSIM performance for two views image compression. (a) - (g) denote different sequences. Bits per pixel (bpp) denotes bite rate. Our model has an advantage in MS-SSIM with the generative capacity of neural networks at most bit rates.

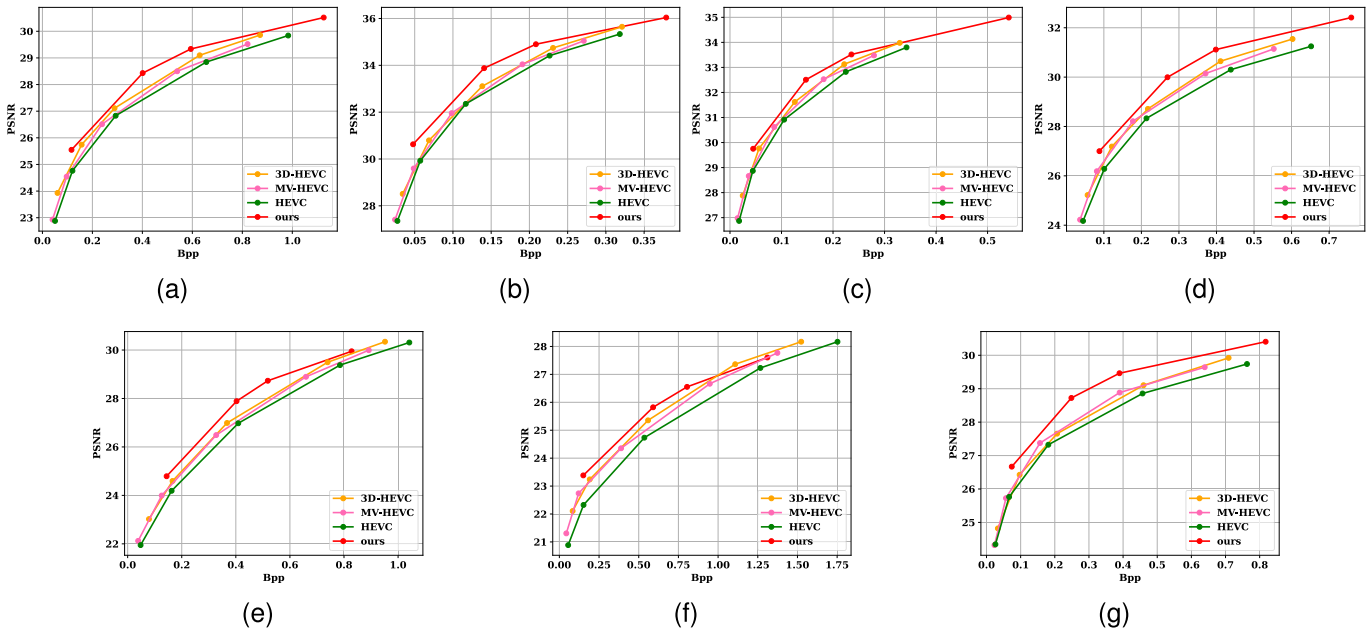

Fig. 8. Comparison of average PSNR performance for three views image compression. (a) - (g) denote different sequences. Bits per pixel (bpp) denotes bite rate. Our model also demonstrates an advantage in PSNR compared to other methods at most bit rates for three views compression.

to verify the excellent compression performance of our model. The HEVC [3] for single image compression is also compared to demonstrate the effectiveness of FICNet. In addition, we also experiment with the stereo image compression models, such as HESIC [10] and SASIC [27]. The objective evaluation results are presented in Fig. 6-9.

In Fig. 6-7, we set the number of views in the compression to two so that we could conduct a broad objective performance comparison with more models. According to the comparison results on PSNR and MS-SSIM, it is obvious that the FICNet is competitive at all bit rates, especially for the low bit rate, which shows that our model can retain the important information of the image and restore texture details within a limited bit rate. For the curve of HEVC, all free-view images are considered as single images and compressed under ALL-INTRA mode when the reference software of HEVC is utilized for image compression. Thus, it can be seen that the compression performance of HEVC is not pleasant because the spatial correlation between free-view images is not fully used for compression. Moreover, we also paint the RD curves of SASIC and HESIC, in which SASIC performs better because of its more precise prediction. However, both models are not doing well on free-view compression, which also indicates the limited generalization of these models.

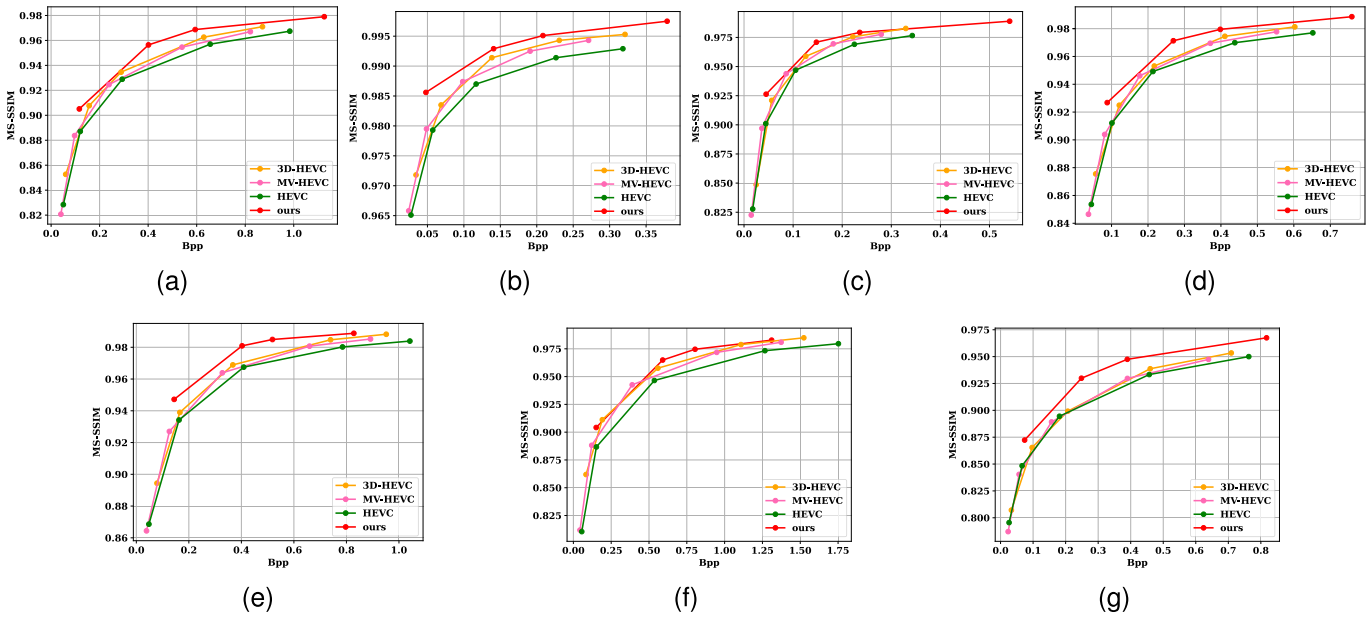

Fig. 9. Comparison of average MS-SSIM performance for three views image compression. (a) - (g) denote different sequences. Bits per pixel (bpp) denotes bite rate. Our model achieves excellent MS-SSIM performance compared to other methods at most bit rates for three views compression.

For stereo image compression, the transformation between images is relatively simple, and the disparity is also implicit, making the prediction maps easier to generate. The complex disparities cannot be directly predicted by neural networks like in stereo models. Therefore, it is not feasible to apply the stereo image compression model for free-view compression. MV-HEVC and 3D-HEVC, which are widely applied for free-view compression, are also compared with our model. The compression effects are satisfactory according to the experimental results. During the compression process of MV-HEVC, the codec explores the spatial correlation between free-view images. For the 3D-HEVC, its reference software incorporates the depth-based prediction method into the prediction module for better prediction performance. In fact, 3D-HEVC is also more suitable for multi-images with horizontal translation, and the bit rate of depth information is included for a fair comparison when calculating the compression performance. Based on the above, the performances of 3D-HEVC and MV-HEVC are comparable. Compared with the noted models, the FICNet achieves an advantageous compression.

In Fig. 8-9, we change the number of compressed views to illustrate the effectiveness of FICNet further, and our model is evaluated against other models on three-view image compression for all sequences based on PSNR and MS-SSIM. Based on the experimental results, our model demonstrates excellent performance across a greater number of views and meets the compression needs of free-view image sequences. Besides, the FICNet still performs well on PSNR and MS-SSIM at most bit rates. It can be seen that our model performs better on sequences (b), (d), and (g). These sequences have neat textures and fewer holes generated during prediction, which brings more gains. However, our model still needs to be improved at the high bit rate on some free-view sequences, such as sequences (e) and (f). The image textures of these sequences are too dense, so the prediction of our model does not meet expectations. For this reason, the textures cannot be well reconstructed, and the performance is insufficient at the high bit rate. In summary, the FICNet still performs excellently at most of the bit rates for all sequences and can adapt to the compression requirements of different numbers of views.

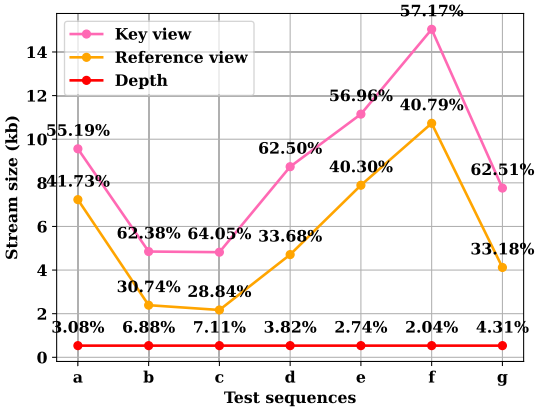

Fig. 10. Bit rate allocation for key view, reference view, and depth map in two-view compression for a - g test sequences. The data labels represent the percentage of different data in the stream.

C. Bit Rate Allocation Results

In the compression of free-view image sequences, we further illustrate how the bit rate is allocated when compressing views of the same scene at a bit rate. Through experiments on bit rate allocation and the compression performance of key and reference views, we can strongly demonstrate the role of our end-to-end model in reducing spatial redundancy. As shown in Fig. 10, we compress two view images on all sequences, and then we calculate the proportions of the bitstream occupied by key views, reference views, and depth

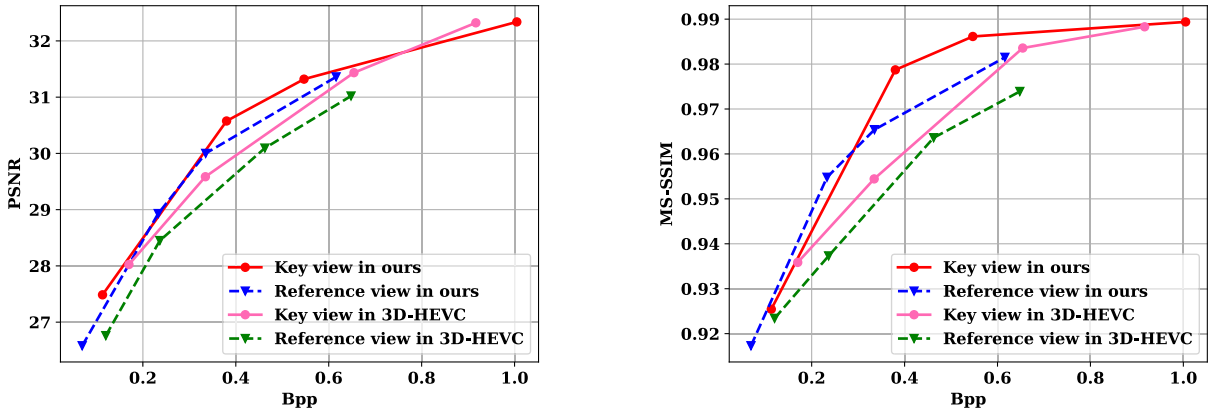

Fig. 11. Comparison of average PSNR and MS-SSIM performance of key views and reference views on our model and 3D-HEVC. The reference view achieves bit rate savings compared to the key view, and our model outperforms 3D-HEVC in terms of performance.

maps respectively. From the compression results, depth maps occupy a small proportion of the compressed bitstream for all sequences due to containing less information, which will not affect compression performance significantly. The size of the bitstream for reference views shows a significant reduction compared to that of key views. The result indicates that our end-to-end residual coding structure plays a crucial role in reducing the entropy and spatial redundancy of reference views. As shown in Fig. 11, we present the average PSNR and MS-SSIM of key views and reference views in two-view compression for all test sequences. Through comparison with 3D-HEVC, we observe a reduction in bit rates for reference views compared to key views. The reduction in bit rates indicates that both models are effective in reducing spatial redundancy. Our model can reconstruct high-quality reference views due to the well-designed prediction model and the utilization of the generative capabilities of neural networks. The performance of the reference view in our model is competitive with the key view, especially at the low bit rate. In summary, our model is effective in enhancing the compression performance of free-view images.

D. Visualization Results

The visualization results of the reconstructed images are provided in Fig. 12. We present the visualizations of the reconstructed images from one view of two free-view test sequences at a bit rate of 0.1 bpp to demonstrate the effectiveness of our model. Additionally, we compare the original image with the reconstructed images of the HEVC, 3D-HEVC, and our proposed model. From the visualization results, it can be observed that the reconstruction results of HEVC and 3D-HEVC are not satisfactory at lower bit rates. They exhibit significant loss of texture information and suffer from local blurring issues in their reconstructed images. In contrast, our proposed model showcases better reconstruction quality in terms of the housing details in the top sequence and ground textures in the bottom sequence. Overall, reconstructed images of our model exhibit richer details and superior subjective quality at low bit rates due to advanced prediction methods and the learning capabilities of neural networks.

E. Model Complexity Results

We compare the average encoding and decoding times for one view of our model with traditional free-view codec models at various bit rates, and the results are shown in Table I. We calculate the average enc/dec time for all test sequences in the three-view compression. Traditional codecs are tested on the CPU to leverage multi-core design for improving speed. Parallel computing units are integrated into many terminal devices such as GPU. Our learning model can take advantage of GPU acceleration, so our model is tested on both CPU and GPU. In terms of encoding time, MV-HEVC takes longer to encode compared to HEVC due to the time-consuming inter-view prediction tools used in MV-HEVC. Additionally, 3D-HEVC has the longest encoding time as it requires additional encoding of depth maps to assist in compressing the free-view images. Our model has a clear advantage in encoding time when tested on GPU with the acceleration design of GPU. However, the encoding time of our model is significantly slower than that of traditional models when tested on the CPU due to the disadvantage of the CPU in image processing. Furthermore, the encoding time of our model remains similar at different bit rates due to the consistent network structure. As for decoding time, the decoding time of traditional codec models is significantly reduced due to their simple designs, which are faster by several orders of magnitude compared to our model. Therefore, our model demonstrates a clear disadvantage in decoding times. Our model will continue to explore methods to accelerate coding in the future.

F. Ablation Study

In this section, the ablation experiments are shown on the components proposed in this paper to demonstrate their effectiveness.

1) Objective Result Analysis: The RD curves are shown in Fig. 13 for the objective comparison results. The experimental results are evaluated on all test sequences for three views compression to obtain the average PSNR and MS-SSIM. Among the RD curves, the curve of *Anchor* represents our baseline model, whose prediction is based on the optical flow method. The curve named *D* denotes the model with the differentiable

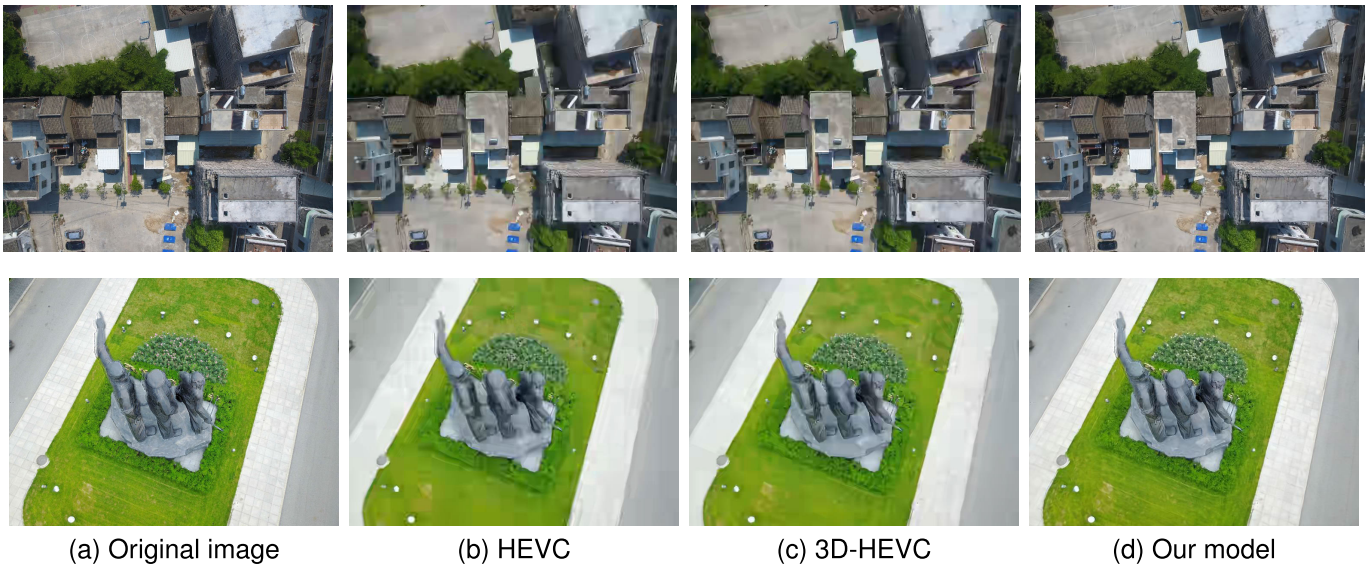

Fig. 12. Visualization of reconstructed images on two sequences. The images from left to right are the original images and reconstructed images of HEVC, 3D-HEVC, and our model.

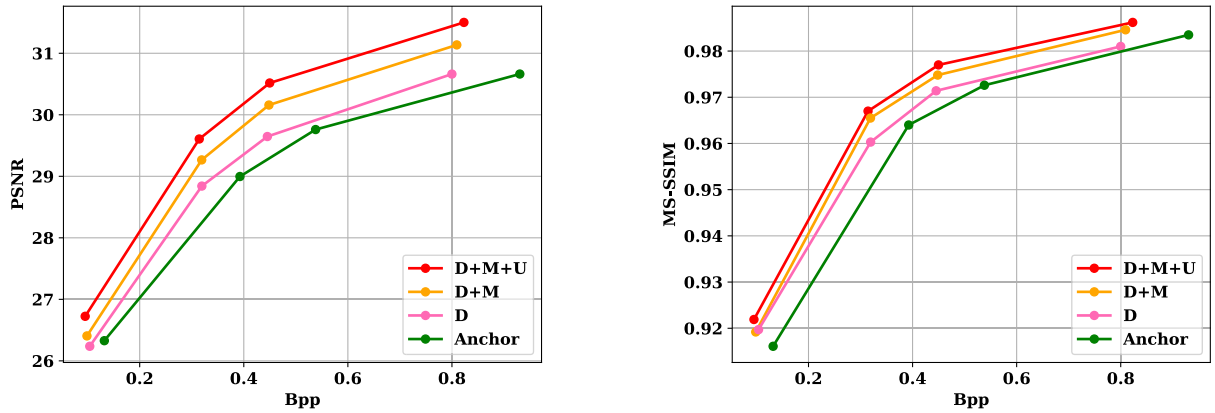

Fig. 13. Results of ablation experiments for our methods on PSNR and MS-SSIM. The experiments are average performance of all sequences. Our methods achieve performance gains according to the results.

depth-based prediction method, and the curve of $D{+}M$ stands for the model utilizing the differentiable depth-based prediction module and multi-view reference. Besides, the curve of $D{+}M{+}U$ describes the model with methods of differentiable depth-based prediction, multi-view reference, and residual fusion. According to the comparison, it can be seen that our differentiable prediction based on depth information has advantages over the traditional prediction with end-to-end training. However, the performance improvement of the model is limited due to the problem of inaccurate original depth data and the trouble of holes. Furthermore, the strategy of multi-view reference realizes improved performance on the model, which demonstrates the essential role of filling holes even if the gain of the multi-view reference strategy depends on the quality of reference views. The filter network eliminates some unreasonable values in prediction images, which also enhances the prediction quality. Finally, our residual fusion network also has an enhancement effect on image reconstruction according to the RD curves. Hence, the utilization of an adaptive additive network is superior to the operation of directly adding the prediction and residual map. In brief, our proposed methods are effective for free-view compression, and all of our methods can achieve gain.

2) Structural Ablation of Filter Network and Residual Fusion Network: In order to explore the rationality of our network design, we conduct the ablation experiment on the filter network using the BD-rate [63]. We test the average performance of all test sequences in the case of three views compression, consistent with the rate control in Fig. 13. The hyper-parameters of our network are inspired by the model [64]. However, the model is designed for the single view, so we introduce the attention mechanism to integrate features from multiple views. Based on this, we need to conduct ablation experiments on the network structure. In Table II, *Res* represents the residual block structure [60], and $\times n$ indicates the number of residual blocks. According to the experimental results, the structure with six layers of residual blocks is reasonable, as it brings a large performance gain compared to the structure with three residual blocks. The structure with nine residual blocks brings less gain and adds a large number of network parameters. Therefore, the structure with six residual blocks is selected with the limitation of the device. In addition,

TABLE I
COMPARISON OF MODEL COMPLEXITY ON CPU AND GPU

Model	Enc time (s/f)	Dec time (s/f)
HEVC	1.112	0.031
MV-HEVC	3.136	0.039
3D-HEVC	11.643	0.069
Our model (CPU)	8.352	5.836
Our model (GPU)	0.423	0.248

TABLE II
STRUCTURAL ABLATION OF FILTER NETWORK ON BD-RATE

Model	BD-rate
Res $\times 6$	0%
Res $\times 6$ w/o mlp	4.23%
Res $\times 3$	7.91%
Res $\times 9$	-1.34%

TABLE III
STRUCTURAL ABLATION OF RESIDUAL FUSION NETWORK ON BD-RATE

Model	BD-rate
Conv $\times 3$	0%
Conv $\times 2$	4.45%
Conv $\times 4$	-0.84%
Conv $\times 3$ with MCF	-4.24%

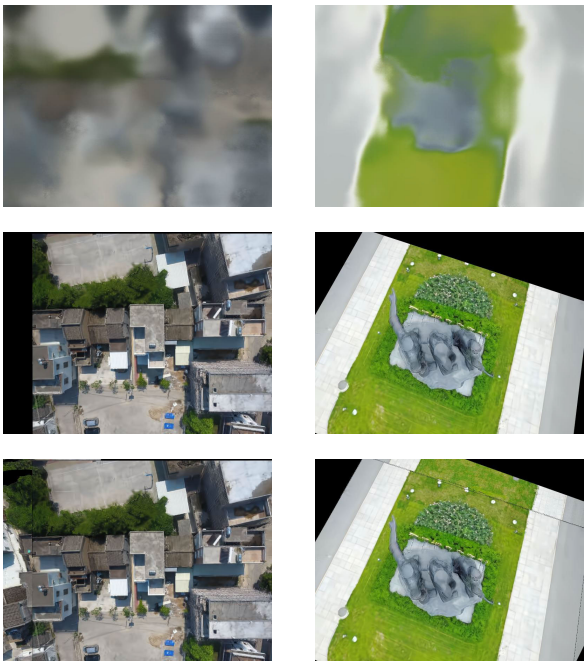

Fig. 14. Visualization of predicted images. The images from top to bottom are the prediction image based on optical flow, the prediction image based on depth and the prediction image based on depth with multi-view reference.

the attention mechanism also brings gains to the filter network. The channel-based attention approach can effectively integrate channel features of multiple prediction images to achieve satisfactory filling effects in the hole regions.

Ablation experiments are also conducted on the residual fusion network to demonstrate the reasonableness and effectiveness of our multi-level complementary feature (MCF) fusion network. In Table III, *Conv* $\times n$ indicates the number of convolution layers in the feature extraction and reconstruction stage. From the comparison, the three-layer convolution structure can achieve good performance without introducing additional network parameters. Besides, it can be seen that the multi-level complementary feature structure brings an additional performance improvement to the fusion network up to 4.24%. Our complementary features can effectively complement the missing information in the features of prediction and residual images for fusion.

3) Visual Analysis: In Fig. 14, we visualize predicted images from different methods on two test sequences, which can intuitively display the effect of our methods. According to the visualization results, the predicted image obtained by the method of optical flow is relatively blurred and cannot be accurately identified, so it cannot match the corresponding area of the original view. For this reason, the entropy of the residual images obtained in the coding of the reference view will increase. On the contrary, the predicted image based on the depth information is clear and accurate, which can attain better prediction results. However, it is obvious that serious holes exist in the depth-prediction image, so we employ the strategy of multi-view reference to fill these holes and repair the predicted image. It can be seen that this approach has some effects towards refining the predicted image even if it still cannot fill in all the holes as shown in the visualization. Therefore, our methods have indeed played an important role in improving prediction accuracy and enhancing reconstruction quality from subjective comparison. However, it is still valuable to explore the enhancement of prediction quality further.

TABLE IV
ABLATION COMPARISON OF MODEL COMPLEXITY ON CPU

Model	Parameters (MB)	Enc time (s/f)	Dec time (s/f)
Anchor	44.8	7.54	4.72
D	48.1	6.59	4.68
D+M	50.2	8.31	5.89
D+M+U	53.1	8.50	7.28

4) Model Complexity: In Table IV, we display the impact of our proposed methods on model complexity, which is evaluated by the size of model parameters, encoding time and decoding time. As for the size of model parameters, many transformation parameters are added and a codec network for compression of the depth map is also introduced in the depth-based prediction module, which increases the parameters effectively. Like this, the residual fusion network introduces additional network parameters and increases complexity. Besides, our multi-view reference strategy also increases the number of parameters since it uses neighbour views to enhance the current predicted images with the filter network.

For the complexity of compression, the encoding and decoding time of our model is also extended after applying our methods. Our model is tested on the CPU, which highlights a pronounced difference. The utilization of the differentiable depth-based prediction module decreases encoding and decoding time compared with optical flow prediction, illustrating its effectiveness. Besides, the model with residual fusion network mainly has an increase in decoding time due to the combination with the additional network during the decoding. The multi-view reference strategy also brings a noticeable increase in encoding and decoding time because of the utilization of the filter network and excessive warping operation.

In conclusion, our proposed methods all increase the model parameters and coding time of images somewhat, but this is acceptable compared with the performance gains they bring.

V. Conclusion

In this paper, we are the first to propose the FICNet, a novel end-to-end learned network for free-view image compression. In our model, we divide the free-view images into key views and reference views for direct encoding and residual encoding respectively. We mainly focus on improving the prediction module to reduce the bit rate and enhance reconstruction quality. In the prediction module, the depth information is utilized to perform the differentiable warp operation and obtain the predicted images, which effectively improves the prediction accuracy and reduces the bit rate of the reference views. Besides, we employ a strategy of multi-view references to fill the holes generated during the prediction. Furthermore, we introduce a filter network to enhance the obtained predictive images and alleviate the problem of area distortion caused by unreasonable depth values. Finally, a residual fusion network with multi-level complementary features is designed to enhance the reconstruction process in residual coding adaptively. Experimental results show that the FICNet outperforms other models on all sequences, both on average PSNR and MS-SSIM, and the more delicate prediction images can be acquired with the proposed components. In conclusion, we provide a 6-DOF image compression framework suitable for a variety of scenes, which is of great significance for the promotion of 3D vision applications. In our future work, we will explore the subjective quality and more robust compression model, conducting tests on various metrics, such as UBQI [65].

References

[1] G. Tech, K. Wegner, Y. Chen, M. M. Hannuksela, and J. Boyce, *MV-HEVC Draft Text 9*, document JCT3V-I1002, Sapporo, Japan, 2014.

[2] G. Tech, K. Wegner, Y. Chen, and S. Yea, *3D-HEVC Draft Text 7*, document JCT3V-K1001, Geneva, Switzerland, 2015.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[4] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1717–1720.

[5] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for multiview video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1474–1484, Nov. 2007.

[6] A. Gelman, P. L. Dragotti, and V. Velisavljevic, "Multiview image compression using a layer-based representation," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 493–496.

[7] C. Lan, H. Yan, C. Luo, and T. Zhao, "GAN-based multi-view video coding with spatio-temporal EPI reconstruction," 2022, *arXiv:2205.03599*.

[8] J. Lei et al., "Disparity-aware reference frame generation network for multiview video coding," *IEEE Trans. Image Process.*, vol. 31, pp. 4515–4526, 2022.

[9] J. Liu, S. Wang, and R. Urtasun, "DSIC: Deep stereo image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3136–3145.

[10] X. Deng et al., "Deep homography for efficient stereo image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1492–1501.

[11] M. Singh and R. M. Rameshan, "Learning-based practical light field image compression using a disparity-aware model," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[12] G. Toderici, "Variable rate image compression with recurrent neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–12.

[13] G. Toderici et al., "Full resolution image compression with recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5435–5443.

[14] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 2961–2987.

[15] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 736–754.

[16] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 4961–5007.

[17] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive entropy model for end-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 8204–8223.

[18] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7939–7948.

[19] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14771–14780.

[20] X. Zhu, J. Song, L. Gao, F. Zheng, and H. T. Shen, "Unified multivariate Gaussian mixture for efficient neural image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17612–17621.

[21] J.-H. Kim, B. Heo, and J.-S. Lee, "Joint global and local hierarchical priors for learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5992–6001.

[22] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5727.

[23] Y. Xie, K. L. Cheng, and Q. Chen, "Enhanced invertible encoding for learned image compression," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 162–170.

[24] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10532–10541.

[25] F. Yang, L. Herranz, Y. Cheng, and M. G. Mozerov, "Slimmable compressive autoencoders for practical neural image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4998–5007.

[26] J. Lei, X. Liu, B. Peng, D. Jin, W. Li, and J. Gu, "Deep stereo image compression via bi-directional coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19637–19646.

[27] M. Wödlinger, J. Kotera, J. Xu, and R. Sablatnig, "SASIC: Stereo image compression with latent shifts and stereo attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 651–660.

[28] Y. Zhai, L. Tang, Y. Ma, R. Peng, and R. Wang, "Disparity-based stereo image compression with aligned cross-view priors," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2351–2360.

[29] H.-C. Feng, M. W. Marcellin, and A. Bilgin, "A methodology for visually lossless JPEG2000 compression of monochrome stereo images," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 560–572, Feb. 2015.

[30] Z. Huang, Z. Sun, F. Duan, A. Cichocki, P. Ruan, and C. Li, "L3C-stereo: Lossless compression for stereo images," 2021, *arXiv:2108.09422*.

[31] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.

[32] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–4.

[33] N. Bakir, W. Hamidouche, O. Déforges, K. Samrouth, and M. Khalil, "Light field image compression based on convolutional neural networks and linear approximation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1128–1132.

[34] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, "Content-based light field image compression method with Gaussian process regression," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 846–859, Apr. 2020.

[35] Y. Chen, P. An, X. Huang, C. Yang, D. Liu, and Q. Wu, "Light field compression using global multiplane representation and two-step prediction," *IEEE Signal Process. Lett.*, vol. 27, pp. 1135–1139, 2020.

[36] H. Jung, H.-J. Lee, and C. E. Rhee, "Re-ordered micro image based high efficient residual coding in light field compression," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3195–3204.
[37] W. Ahmad, R. Olsson, and M. Sjöström, "Interpreting plenoptic images as multi-view sequences for improved compression," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4557–4561.
[38] M. Sharma and G. Ragavan, "A novel randomize hierarchical extension of MV-HEVC for improved light field compression," in *Proc. Int. Conf. 3D Immersion (IC3D)*, Dec. 2019, pp. 1–8.
[39] W. Ahmad, S. Vagharshakyan, M. Sjöström, A. Gotchev, R. Bregovic, and R. Olsson, "Shearlet transform-based light field compression under low bitrates," *IEEE Trans. Image Process.*, vol. 29, pp. 4269–4280, 2020.
[40] J. Gu, B. Guo, and J. Wen, "High efficiency light field compression via virtual reference and hierarchical MV-HEVC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 344–349.
[41] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, and J. Yang, "Light field image compression based on deep learning," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2018, pp. 1–6.
[42] J. Hou, J. Chen, and L.-P. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 517–530, Feb. 2019.
[43] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 177–189, Mar. 2019.
[44] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrouth, and O. Déforges, "Light field image coding using dual discriminator generative adversarial network and VVC temporal scalability," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[45] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrouth, and O. Déforges, "Light field image coding using VVC standard and view synthesis based on dual discriminator GAN," *IEEE Trans. Multimedia*, vol. 23, pp. 2972–2985, 2021.
[46] X. Hu, J. Shan, Y. Liu, L. Zhang, and S. Shirmohammadi, "An adaptive two-layer light field compression scheme using GNN-based reconstruction," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2s, pp. 1–23, Apr. 2020.
[47] J. Shan, Y. Liu, and Y. Wang, "Light field images compression based on graph convolution networks," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.
[48] Y. Zhang et al., "Light field compression with graph learning and dictionary-guided sparse coding," *IEEE Trans. Multimedia*, vol. 25, no. 1, pp. 3059–3072, Feb. 2022.
[49] Y.-H. Chao, H. Hong, G. Cheung, and A. Ortega, "Pre-demosaic graph-based light field image compression," *IEEE Trans. Image Process.*, vol. 31, pp. 1816–1829, 2022.
[50] T. Zhong, X. Jin, and K. Tong, "3D-CNN autoencoder for plenoptic image compression," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2020, pp. 209–212.
[51] Y. Gao, G. Cheung, T. Maugey, P. Frossard, and J. Liang, "Encoder-driven inpainting strategy in multiview video compression," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 134–149, Jan. 2016.
[52] D. M. M. Rahaman and M. Paul, "Virtual view synthesis for free viewpoint video and multiview video compression using Gaussian mixture modelling," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1190–1201, Mar. 2018.
[53] Y. Chen, X. Zhao, L. Zhang, and J.-W. Kang, "Multiview and 3D video compression using neighboring block based disparity vectors," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 576–589, Apr. 2016.
[54] Y. Zhang, L. Zhu, R. Hamzaoui, S. Kwong, and Y.-S. Ho, "Highly efficient multiview depth coding based on histogram projection and allowable depth distortion," *IEEE Trans. Image Process.*, vol. 30, pp. 402–417, 2020.
[55] D. Rüefenacht, A. T. Naman, R. Mathew, and D. Taubman, "Base-anchored model for highly scalable and accessible compression of multiview imagery," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3205–3218, Jul. 2019.
[56] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 10771–10780.
[57] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1996, pp. 358–363.
[58] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.
[59] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[61] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1790–1799.
[62] Z. Cui, J. Wang, S. Gao, B. Bai, T. Guo, and Y. Feng, "Asymmetric gained deep image compression with continuous rate adaptation," 2020, *arXiv:2003.02012*.
[63] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 2001.
[64] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
[65] K. Gu, X. Xu, J. Qiao, Q. Jiang, W. Lin, and D. Thalmann, "Learning a unified blind image quality metric via on-line and off-line big training instances," *IEEE Trans. Big Data*, vol. 6, no. 4, pp. 780–791, Dec. 2020.

Chunhui Yang received the B.S. degree in computer science and technology from Sichuan University, China, in 2019. He is currently pursuing the Ph.D. degree in computer application technology with Peking University. His research interests include image coding and immersive media coding.

Jiayu Yang received the B.E. degree from Hefei University of Technology, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Electronic and Computer Engineering, Peking University, China. His research interests include video coding and immersive media coding.

Yongqi Zhai received the B.S. degree in electronic information engineering from Jilin University, China, in 2020. He is currently pursuing the Ph.D. degree in computer application technology with Peking University. His research interests include traditional and learning-based image/video coding.

Ronggang Wang (Member IEEE) was born in Hailin, Heilongiiang, China, in 1976. He received the Ph.D. degree in computer engineering from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. From 2006 to 2010, he was a Researcher with the France Telecom Research and Development Laboratory. He is currently a Professor with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University. He is the author of more than 200 articles and more than 100 inventions. His research interests include video coding and ultra-high-definition immersive media processing. He has led the development of the ISO/EC MPEG Internet Video Coding. He is also the Co-Chair of the ISO/EC MPEG Internet Video Coding Ad Hoc Group and the Chair of the lEEE 1857.9 Subgroup on Immersive Visual Content Coding.