IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 4, 2021, accepted July 18, 2021, date of publication August 3, 2021, date of current version August 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102114

Current Frame Priors Assisted Neural Network for Intra Prediction

HAN ZHANG[1], LI SONG[1,2], (Senior Member, IEEE), YAN HUANG[1], AND RONG XIE[1], (Member, IEEE)

[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Li Song (song_li@sjtu.edu.cn)

This work was supported in part by the National Key Research and Development Project of China under Grant 2019YFB1802701, in part by the MoE-China Mobile Research Fund Project under Grant MCM20180702, in part by NSFC under Grant 61671296 and Grant U1611461, in part by the 111 Project under Grant B07022 and Sheitc 150633, and in part by Shanghai Key Laboratory of Digital Media Processing and Transmissions.

ABSTRACT Intra prediction is the key technology to reduce spatial redundancy in the modern video coding standard. Recently, deep learning based methods that directly generate the intra prediction by neural network achieve superior performance than traditional directional based intra prediction. However, these methods lack the ability to handle complex blocks which contain mixed directional textures or recurrent patterns since they only use the neighboring reference samples of the current one. The other intermediate information denoted as reference priors in this paper generated during the coding process is not exploited. In this paper, a Current Frame Priors assisted Neural Network (CFPNN) is presented to improve the intra prediction efficiency. Specifically, we utilize the local contextual information provided by the neighboring multiple references as the primary inference source. In addition to the neighboring references, we additionally use the other two reference priors within the current frame – the predictor searched by intra block copy (IntraBC) and the corresponding residual component. The IntraBC predictor provides useful nonlocal information to help generate more accurate prediction for complex blocks together with neighboring local information. While the residual component contains unique information that reflects the characteristics of the block to some extent is utilized to reduce the noise contained in the reconstructed reference samples. Moreover, we investigate the best way to integrate the proposed method into the codec. Experimental results demonstrate that compared to HEVC, our proposed CFPNN achieves an average of 4.1% BD-rate reduction for the luma component under the All Intra configuration.

INDEX TERMS Neural network, intra prediction, reference prior, High Efficiency Video Coding.

I. INTRODUCTION

Intra prediction is an efficient way to remove spatial redundancy within a frame by exploiting the similarity of adjacent pixels. It plays an essential role in current video coding standards, like Advanced Video Coding (AVC) [1] and High Efficiency Video Coding (HEVC) [2]. For intra prediction of block based video coding framework, the prediction of the current to-be-coded block is generated by extrapolation of the nearest neighboring single reference line according to the directional prediction mode. There are 35 intra prediction modes for HEVC, while the available prediction modes for AVC are 9. Combined with a more flexible prediction block size, HEVC can achieve about 22.3% bitrate saving compared with AVC on intra prediction [3]. To improve the directional accuracy, the number of directional intra modes is increased to 65 for next generation video coding standard – Versatile Video Coding (VVC) [4]. With the support of a more flexible prediction block shape and size, the coding efficiency is further improved.

In addition to extending the directional modes number, numerous methods have been proposed to improve the intra prediction efficiency of HEVC. Wei *et al.* [5] applied DCT-based interpolation filter to generate fractional reference samples. Chen *et al.* [6] proposed an iterative filtering method to smooth the conventional copy-based prediction samples. Four-tap recursive extrapolation filters based on the Markov model were employed to improve the prediction

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

samples in [7]. These methods still use the nearest neighboring single line as HEVC to generate prediction samples. There are also methods to explore the performance by introducing more reference information or designing a new intra prediction framework. In [8], multiple reference lines were utilized to generate prediction. Qi *et al.* [9] introduced two image inpainting algorithms into intra prediction. In [10], two predicted blocks derived from different prediction directions were weighted combined to generate the final prediction. Chen *et al.* [11] proposed a new intra prediction method which coded half pixels of a block while reconstructed the other half by linear interpolation. Zhang *et al.* [12] proposed a hybrid intra prediction method by jointly exploring nonlocal correlation through template matching prediction and local correlation. Intra Block Copy (IntraBC) which performs motion compensation in the already reconstructed areas within the current frame and has been adopted in the HEVC extension for screen content coding, was also introduced in natural content intra prediction [13]. Li *et al.* [14] proposed a combination of regular intra prediction and IntraBC to generate better prediction.

In the last few years, explorations of applying deep learning to the video coding task have been carried out and have achieved impressive success. Generally speaking, deep learning based video coding can be classified into two categories. The deep neural network of the first category takes the uncompressed video as input and directly outputs the compressed bitstream [15]–[18]. This kind of neural network model is called the end-to-end model, which is out of the scope of this paper. The second category methods still follow the conventional block based hybrid video coding framework [19]. The deep neural network is integrated into the framework to improve the performance of particular module including inter prediction [20]–[22], transform [23], entropy coding [24], [25], rate control [26], [27], in-loop filtering/post processing [28]–[31], and intra prediction [32]–[38] The details of these deep learning based video coding methods are reviewed in Section. II-B.

Although the previous learning based intra prediction methods have achieved remarkable performance, there is still much potential for further improvement. The previous methods use more local contextual information by feeding multiple reference lines or neighboring blocks of the current one into the network, which can generate a better prediction for most blocks. However, for some complex blocks which contain complicated textures like mixed multiple directional textures or recurrent patterns, only using the local correlation of adjacent pixels is not enough to generate an accurate prediction. A possible way to further improve the prediction accuracy of these complex blocks is using additional nonlocal contextual information. Furthermore, previous works take the reconstructed reference as input without utilizing any other compression information generated during the encoding/decoding process, such as partition mode, prediction residual, etc. We name these different kinds of intermediate information as reference priors in this paper. These reference priors, which are used as the elementary component to form the final output of the encoder/decoder, always contain unique characteristics. The effectiveness of reference priors has been demonstrated in post-processing [39] and fractional interpolation [22]. For intra prediction, these reference priors are all within the current frame. Introducing current frame priors is another potential way to improve intra prediction. Moreover, the nonlocal information is also one kind of reference prior since it is derived before reconstructing the current block.

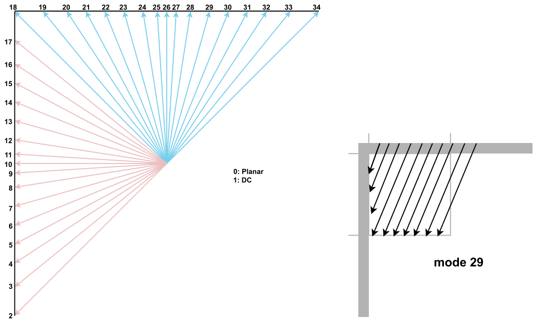
In this paper, we propose a Current Frame Priors assisted Neural Network (CFPNN) to provide more accurate intra prediction for complex blocks. The proposed CFPNN takes three input components to use both local/nonlocal correlation and intermediate reference priors. The neighboring multiple L-shape *reference* lines of the current block are still used as the primary inference source to provide local contextual information. In addition to the L-shape reference lines, we further take advantage of the other two reference priors. On the one hand, the *IntraBC* predictor, which is the best matching block searched in the already reconstructed areas within the current frame, is utilized to explore nonlocal correlation. On the other hand, the corresponding *residual* component of the predictor is used as the third input, since the *residual* prior contains extra information about the characteristics of block texture. Experimental results demonstrate that compared with HM-16.9, the proposed CFPNN scheme achieves an average 4.1% BD-rate reduction. The major contributions of this paper are listed as follows:

- A novel Current Frame Priors assisted Neural Network (CFPNN) based intra prediction is proposed. Apart from the neighboring multiple L-shape reference lines, our proposed method also takes advantage of *IntraBC* predictor and the corresponding *residual* component to explore nonlocal correlation and unique characteristics of intermediate compression information.
- The network architecture is carefully designed to make use of these three input components simultaneously. A channel-wise attention mechanism is applied to combine features from different components efficiently.
- We present two schemes to integrate learning based intra prediction into the codec. For the first scheme, networks with different input components are applied to blocks with different texture characteristics. For the second scheme, all blocks use CFPNN with three inputs uniformly. Comprehensive experiments have been carried out to demonstrate the efficiency.

The rest of this paper is organized as follows: Section. II reviews the related work. The details of our proposed method are introduced in Section.III. Comprehensive experimental results are presented in Section.IV. Finally, Section.V concludes this paper.

II. RELATED WORK

In this section, we first introduce the original intra prediction of HEVC briefly. Then some recent deep learning based



(a) (b)

FIGURE 1. **HEVC intra prediction modes. (a) 35 modes. (b) Example of intra prediction with mode 29.**

video coding methods, especially deep learning based intra prediction methods, are reviewed.

A. INTRA PREDICTION OF HEVC

In HEVC, there are three new coding structure related concepts – coding unit (CU), prediction unit (PU), and transform unit (TU). A CU with a size varied from 8 × 8 to 64 × 64 can be further divided into PUs, each of which has identical prediction information. For coding the prediction residual, a CU can be split into multiple TUs. In intra prediction, TU is the basic unit that conducts the prediction process. All TUs inside a PU share the same intra prediction information.

For TU with size from 4 × 4 to 32 × 32, there are a total of 35 intra modes, including Planar, DC, and 33 angular modes. DC and Planar are designed for predicting blocks with flattening textures and gradually changing textures, respectively. The remaining angular modes target areas with strong directional textures. For a TU with size NxN, the neighboring 4N+1 samples are used as the single reference line to generate prediction, as shown in Fig.1. DC mode uses the average value of reference samples as the prediction of the current to-be-coded block, while planar mode uses bilinear interpolation to generate the prediction. For angular modes, the prediction is generated by extrapolating the reference samples along a given direction.

B. DEEP LEARNING BASED VIDEO CODING

Recently, deep learning based methods have achieved remarkable improvement in conventional video coding task. These methods are integrated into the hybrid coding framework to replace or improve a particular module.

For inter prediction, Zhao *et al.* [20] used a convolutional neural network (CNN) model to combine two prediction blocks to enhance the bi-directional inter prediction. Lee *et al.* [21] proposed using a video prediction network to generate a virtual reference frame for motion estimation and compensation. In [22], a CNN based fractional interpolation method was presented to improve the inter prediction efficiency. For the transform and entropy coding, Liu *et al.* [23] trained a DCT-like transform network for image coding. References [24], [25] proposed using CNN to predict the probability distribution of the intra prediction related syntax elements. Reinforcement learning has been introduced into the traditional rate control to allocate bitrate and estimate coding parameters in [26], [27]. In order to remove the compression artifacts, He *et al.* [28] proposed to use one kind of reference prior – the partition information to guide the quality enhancement combined with the distorted frame. A dual-domain based artifact reduction neural network was proposed in [29], which can learn representation from both pixel domain and DCT domain. Apart from these post-processing methods that directly improve the quality of decoded frames, Jia *et al.* [30] applied a content-aware CNN model after the SAO as an additional in-loop filter. An attention-based loop filter was proposed to replace originally existed filters in [31], which can process luma and chroma components simultaneously.

Several deep learning based methods have also been proposed to improve the HEVC intra prediction. These methods can be classified into two categories according to the number of types of adopted neural network model (NM). The methods in the first category adopt the same type of NM for all blocks. Cui *et al.* [32] proposed a CNN based intra prediction refinement method. The 8 × 8 prediction block is first generated by the HEVC intra prediction, and then this 8 × 8 block is fed into the CNN model together with its three nearest reconstructed 8 × 8 blocks to get a refined intra prediction block. Instead of using convolutional neural network, Li *et al.* [33] proposed a fully-connected network to learn an end-to-end mapping from the neighboring multiple reference lines to the intra prediction block. Hu *et al.* [36] designed a progressive spatial recurrent neural network (PS-RNN) to conduct intra prediction. A spatial RNN is applied to generate the prediction progressively based on the neighboring content. In addition, they also proposed to use the Sum of Absolute Transformed Difference (SATD) as the loss function. Wang *et al.* [34] proposed a multi-scale CNN (MSCNN) for intra prediction. The MSCNN also uses the predicted block generated by HEVC intra prediction as one of the input components. Different from [32], MSCNN further takes the neighboring multiple reference lines as additional information. With the help of a multi-scale feature extraction subnetwork, the MSCNN achieves much better performance than [32]. In [38], a CNN based intra prediction (CIP) was proposed with the neighboring single reference line and the predictor found by intra block copy as inputs. However, the predictor is directly added to the prediction generated by local information, which is not an efficient way to combine these two kinds of information. For methods in the second category, multiple types of NMs are adopted to improve the intra prediction. Dumans *et al.* [37] proposed a set of neural networks, which is called Prediction Neural Networks Set (PNNS). In PNNS, fully-connected neural networks are used to generate intra prediction for small blocks, while convolutional neural networks are used for large blocks. Sun *et al.* [35] explored two different schemes to integrate multiple types of neural network models into the HEVC

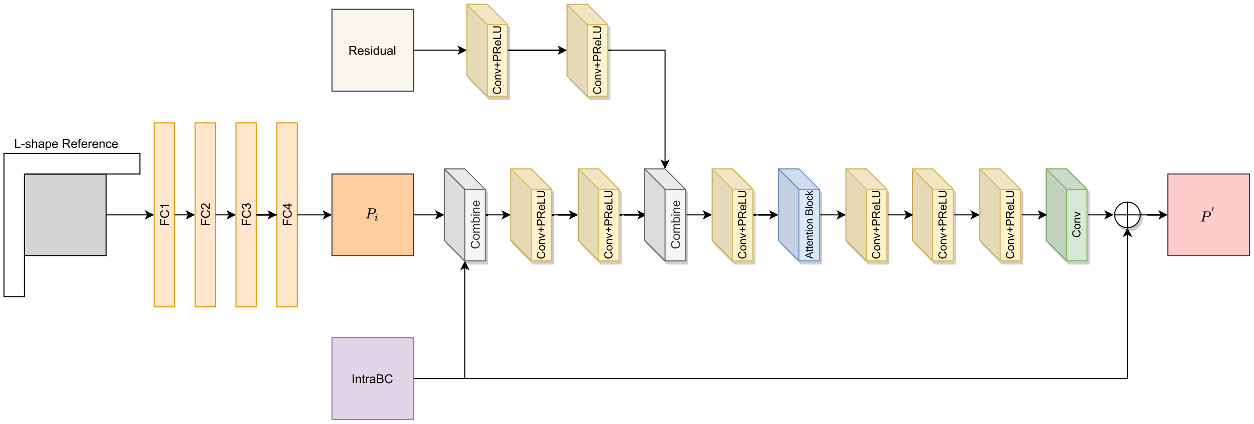


FIGURE 2. Overall Network Structure of Proposed CFPNN, which takes three inputs and directly output intra prediction block.

framework. For the appending scheme, the neural network models are regarded as additional intra prediction modes. For the substitution scheme, the neural network models are integrated by replacing original intra modes. Reference [33] also presented a multiple NMs scheme by applying different neural network modes to angular and non-angular intra direction modes. Generally speaking, methods that adopt multiple types of neural network models perform better than methods with a single type of neural network model.

III. CURRENT FRAME PRIORS ASSISTED NEURAL NETWORK FOR INTRA PREDICTION

In this section, we will introduce the details of our proposed method, including overall framework, network architecture, training data generation, training strategy, and how to integrate into the codec.

A. OVERALL FRAMEWORK

In order to keep the consistency of the encoder and decoder, the previously reconstructed boundary samples are used to form an L-shape reference line in the original intra prediction. Meanwhile, neighboring multiple L-shape reference lines are also adopted as the only input in previous learning based intra prediction methods. Although, the neighboring multiple L-shape reference lines can provide more local contextual information than a single reference line and generate a better prediction for most blocks. For some complex blocks which contain mixed directional textures or recurrent patterns, only using neighboring L-shape reference lines and local correlation makes the intra prediction inefficient. To generate a more accurate intra prediction for the current block, especially for complex blocks, we propose a Current Frame Priors assisted Neural Network (CFPNN) to conduct the prediction process. In our proposed CFPNN, in addition to adopting the neighboring multiple L-shape reference lines, we also use the other two reference priors within the current frame – *IntraBC* predictor and corresponding *residual* component to take advantage of nonlocal information and unique characteristics of reference prior to improve the prediction efficiency.

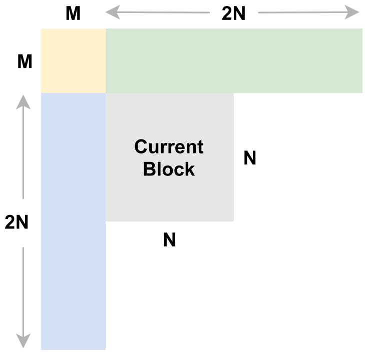


FIGURE 3. Neighboring multiple L-shape reference lines. For a $N \times N$ block, there are $4NM + M^2$ samples in the reference lines.

In intra prediction, the neighboring L-shape reference line, which is composed of the neighboring decoded samples from top-right to bottom-left of the current to-be-coded block, is shown in Fig.3. Similar to previous work, we also utilize multiple L-shape reference lines to extract more local correlations. The width of the reference lines is donated as M. Therefore, for a block with size $N \times N$, there are a total of $4NM + M^2$ samples in the multiple L-shape reference lines. Compared with the original intra prediction process in HEVC, which uses $4N + 1$ samples as reference, multiple reference lines can provide more local contextual information.

In video coding, a fundamental assumption is that the patterns of objects in a frame are similar to corresponding objects in other frames or similar to other patterns within the current frame. To make use of this nonlocal correlation to generate an accurate prediction for complex blocks, We use the best matching predictor of the current one within the current frame as an additional information source. Specifically, we use the intra block copy (IntraBC) to find the best matching predictor [40]. IntraBC is an efficient coding tool in HEVC extensions on Screen Content Coding (HEVC SCC) [41]. In IntraBC, the best matching block of the current to-be-coded one is searched in the already reconstructed regions within the same frame and used as the predictor. As illustrated in Fig.4,

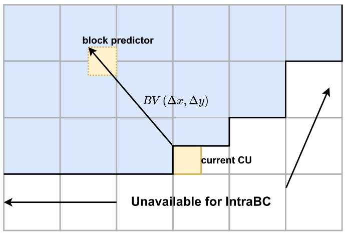


FIGURE 4. Example of IntraBC search range. The light blue area is the available search range, and the block vector (BV) indicates the relative positional relationship.

a block vector (BV) is used to indicate the relative position relationship of the best matching predictor and the current one. The IntraBC predictor introduces a nonlocal correlation that is not contained in the neighboring reference lines to improve the prediction accuracy.

In the block based hybrid video coding framework, both the neighboring L-shape reference lines and the best matching nonlocal predictor searched by IntraBC are decoded reconstruction samples before the in-loop filtering process. These reconstructed samples are accompanied by mixed distortions due to the block based coding process and Quantization step, making the network hard to learn efficient representations from them. To help the network generate a more accurate prediction, we propose to use the corresponding *residual* component to enhance the prediction quality. The *residual* component is the difference between original and corresponding prediction samples. Generally, areas containing complex textures, large motion objects, or sharpening edges are hard to be predicted and accompanied by large residuals, which also means the *residual* component can reflect the complexity of the corresponding block to some extent. Therefore, the *residual* contains extra information about texture complexity, which can help to reduce the noise contained in the reference samples, especially on those complexity areas, and improve the prediction accuracy.

In summary, our proposed CFPNN takes two reference priors – the *IntraBC* predictor and corresponding *residual* component to assist the neighboring multiple L-shape *reference* lines to directly generate intra prediction block.

B. NETWORK ARCHITECTURE OF PROPOSED CFPNN

The overall architecture of our proposed CFPNN is illustrated in Fig.2. As we mentioned in III-A, the inputs of CFPNN are the neighboring multiple L-shape reference lines, the best matching predictor searched by IntraBC, and the corresponding residual component. In order to utilize these three inputs simultaneously, we carefully design the network architecture to generate a better prediction.

Since the neighboring multiple L-shape reference lines are utilized as the primary inference source to provide more local contextual information. We also use a fully-connected (FC) structure to extract features from this input and generate a primitive prediction P_i. For a block with size $N \times N$, the input of the FC structure is the neighboring $4NM + M^2$ samples as shown in Fig.3. The multiple lines could provide more local correlations, but the correlation would decrease if large width lines were used. In our implementation, the width of reference lines M is set as $min(N, 8)$. The multiple reference lines are first flattened to a $4NM + M^2$-dimension vector as the FC layers input. For the ith FC layer, its output is a K-dimension vector calculated as:

$$F_i^{FC}(X) = W_i^{FC} X_i^{FC} + b_i^{FC} \tag{1}$$

where X_i^{FC} is the input of the ith FC layer, which is the flattened reference samples for the first layer and the non-linear activated output of the last layer for other FC layers. W_i^{FC} and b_i^{FC} are the corresponding weight and bias of each layer. The final output of the FC structure is a N^2-dimension vector which will be reshaped to a $N \times N$ block as the primitive prediction.

To make use of the nonlocal correlation, the best matching predictor searched by IntraBC is combined with the primitive prediction P_i generated by the neighboring reference lines. The following convolutional layers are used to extract both local and nonlocal features. The *residual* prior is introduced as complementary information to help to reduce noise contained in the reconstructed reference samples and enhance the generated prediction. We design a light structure that contains two convolutional layers followed by an activation function to extract corresponding features. The extracted residual features are combined with the joint local and nonlocal features extracted from the other two inputs to form the input of the following layers.

In order to make full use of these three inputs, we further propose to use channel-wise attention to combine these feature maps. Specifically, the Squeeze-and-Excitation block (SE block) is used as the channel-wise attention unit [42]. Feature recalibration is performed by two steps named squeeze and excitation in SE block. A 1 × 1 descriptor is generated at the squeeze step for a set of input feature maps with channel size C. This descriptor is used as the input of the following excitation step to derive the channel-wise weights for all channels. These channel-wise weights are applied to scale the original input feature maps and generate the output weighted feature maps. Compared with directly concatenated feature maps extracted from three inputs of our proposed CFPNN, using the channel-wise attention unit fuse the feature maps in a more efficient way. These different kinds of feature maps are combined based on their relative importance on the intra prediction.

These aggregated feature maps are directly fed into the subsequent convolutional layers to generate the final accurate prediction block. In our proposed architecture, each convolutional layer except the last one contains 64 filters with kernel size 3 × 3. The output of each convolutional layer is formulated as:

$$F_i^{Conv}(X) = W_i^{Conv} X_i^{Conv} + b_i^{Conv} \tag{2}$$

where X_i^{Conv} is the input of each convolutional layer. W_i^{Conv} and b_i^{Conv} are the learnable parameters. In our network, we take the Parametric ReLU (PReLU) as the activation function, where a scale parameter needs to be learned during the training phase [43]. We also adopt the residual learning strategy to enable faster converge. In residual learning, the network just learns the difference between output and input. Considering that the best matching block searched by IntraBC is as a predictor for the current to-be-coded one, we add this input component with the output of the last convolutional layer to generate the accurate prediction.

C. TRAINING DATA GENERATION

The first ten frames of all sequences from the SJTU 4K video dataset [44] and images with 2K resolution from the super resolution dataset DIV2K [45] are used to form the training sequence set and generate the training data pair. Since the best coding parameters need to be determined by RDO at encoder side, the training data pair $\left(x^{Ref}, x^{IntraBC}, x^{Resi}, y\right)$ – multiple L-shape *reference*, *IntraBC* predictor and corresponding *residual* prior are extracted from the decoder side, while the *label* y is extracted from the original sequence. To generate the training data, all images from the training sequence set are encoded under the All Intra (AI) configuration recommended by HEVC. Four different QPs – {22, 27, 32, 37} are used to encode these images. In order to collect the *IntraBC* at the decoder side, the best matching block is searched when the original intra prediction process is finished, and the related information is transmitted to decoder side as additional side information.

In HEVC, although the intra prediction related information is derived at the PU level, the actual prediction process is conducted for each TU. The training data pairs are also extracted at the TU level. Our proposed method tends to generate the intra prediction block as close to the current TU as possible. Thus the current to-be-coded TU is extracted as *label*. For each to-be-coded block, the neighboring multiple reconstructed samples as shown in Fig.3 are collected as the multiple L-shape *reference*. If some parts of the reference lines have not been reconstructed yet, the same padding manner in HEVC intra prediction is applied to generate the reference lines. In our implementation, the IntraBC best matching block is searched for each CU, and all TUs in the CU share the same IntraBC related motion information. As for the collection of the *IntraBC* predictor and corresponding *residual* component, we first retrieval the best matching block for the entire CU at the already reconstructed regions according to the block vector. Then the *IntraBC* and corresponding *residual* prior can be easily extracted based on the relative position of the current TU within its corresponding CU. Not all the blocks coded with intra mode are collected as training data. The best matching block for IntraBC is searched based on the Sum of Absolute Difference (SAD) between the candidate predictors and the current block. For some blocks with overly complex textures, it is difficult to find accurate similar predictor. Using these rare blocks is not conducive to training a general model. The training data are further refined by the following condition:

$$\frac{MSE^{i}_{IntraBC}}{Ave(MSE_{IntraBC})} < 1/2 \tag{3}$$

where $MSE^{i}_{IntraBC}$ is the Mean Square Error (MSE) between the ith block and its corresponding IntraBC predictor, $Ave(MSE_{IntraBC})$ is the average MSE of all blocks in a frame. Only the blocks whose IntraBC predictor has small MSE are kept as training data. To make the network easier to train, the training data pairs are further normalized to range [0, 1].

Since the available TU size for intra prediction varies from 4 × 4 to 32 × 32, we train individual networks for each possible TU size. In addition, the training data extracted from the bitstream coded with different QPs are mixed together. When conduct the real video coding test, all test QPs share the same model. In this way, the number of required neural network models can be significantly reduced.

D. TRAINING STRATEGY

In order to get the end-to-end mapping from the triple input components to the output intra prediction, the network parameters need to be estimated according to the training data with predefined loss function. To keep the consistency with the distortion metric used in the RDO of video coding, we adopt MSE as the loss function. We also introduce a regularization term in the loss function to avoid over-fitting during the training procedure. Given a training data set contains N training data pairs $\left\{x_i^{input}, y_i\right\}$, the training loss is:

$$L(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|F(x_i^{input};\Theta) - y_i\right\|_2^2 + \frac{\lambda}{2}\|\Theta\|_2^2 \tag{4}$$

where, x_i^{input} is the input triple components which are $(x_i^{Ref}, x_i^{IntraBC}, x_i^{Resi})$. $F(x_i^{input};\Theta)$ is the generated intra prediction, y_i is the corresponding label. Θ is the learnable network parameters, including weights, biases of the convolutional layer and fully-connected layer, and the scale factor of PReLu.

During the training phase, the regularization term weight λ is set as 10^{-5}. The weight of each layer is initialized by Xavier, and the bias is initialized as 0. The parameter set Θ is optimized with Adam in [46]. The base learning rate is initialized as 10^{-4}, and decayed to 10^{-8} by a factor of 10^{-1}. The proposed neural network model is trained on Caffe.

E. INTEGRATION WITH HEVC

The proposed method is integrated into HEVC framework to conduct real video coding test and evaluate the coding performance. Since our proposed method utilizes the best matching predictor searched by IntraBC, we implement a preliminary IntraBC technique in the original HEVC framework. In our implementation, IntraBC is conducting on the CU level. For a CU, the best predictor is searched within the already reconstructed areas, and a Block Vector (BV) is used to indicate

TABLE 1. Overall BD-rate performance of our proposed method compared to HEVC.

Class	Sequence	CFPNN-F			CFPNN-U		
		BD-Rate			BD-Rate		
		Y	U	V	Y	U	V
ClassA1	Tango	-6.8%	0.2%	-3.2%	-4.5%	1.7%	-1.2%
	Drums100	-3.6%	-0.8%	-1.6%	-3.0%	-0.5%	-1.2%
	CampfireParty	-3.1%	-2.3%	-2.3%	-3.0%	-2.1%	-1.7%
	ToddlerFountain	-3.0%	1.8%	-1.2%	-2.7%	1.5%	-1.2%
ClassA2	CatRobot	-5.1%	-2.8%	-2.1%	-4.6%	-2.6%	-1.9%
	TrafficFlow	-6.8%	-2.1%	-2.8%	-6.1%	-2.3%	-3.1%
	DaylightRoad	-5.5%	-1.7%	-2.3%	-5.4%	-0.8%	-1.9%
	Rollercoaster	-5.0%	-2.9%	-2.0%	-3.7%	-1.9%	-1.6%
ClassA	Traffic	-4.1%	-2.2%	-1.9%	-3.3%	-2.5%	-2.0%
	PeopleOnStreet	-5.0%	-2.4%	-2.1%	-4.3%	-3.6%	-3.6%
ClassB	Kimono	-2.7%	-1.5%	-0.8%	-0.2%	-0.6%	-0.8%
	ParkScene	-3.1%	-2.4%	-1.8%	-2.5%	-1.8%	-2.2%
	Cactus	-5.0%	-2.2%	-2.9%	-3.9%	-2.5%	-3.3%
	BasketballDrive	-5.2%	-2.5%	-3.1%	-2.9%	-3.1%	-2.5%
	BQTerrace	-5.2%	-2.2%	-2.2%	-4.7%	-3.3%	-2.1%
ClassC	BasketballDrill	-3.4%	-2.5%	-1.5%	-3.6%	-2.2%	-1.7%
	BQMall	-2.2%	-1.9%	-1.4%	-2.0%	-2.2%	-2.3%
	PartyScene	-1.7%	-0.7%	-0.9%	-1.5%	-0.6%	-1.0%
	RaceHorsesC	-2.9%	-1.6%	-1.8%	-2.0%	-1.4%	-2.3%
ClassD	BasketballPass	-1.7%	-1.0%	0.9%	-0.7%	-0.8%	0.4%
	BQSquare	-1.2%	1.3%	-1.2%	-1.2%	1.3%	0.0%
	BlowingBubbles	-1.8%	-2.6%	-2.6%	-1.2%	-1.1%	-2.6%
	RaceHorses	-3.3%	-2.6%	0.0%	-2.4%	-1.9%	-1.1%
ClassE	FourPeople	-5.6%	-3.4%	-3.7%	-4.6%	-4.5%	-4.4%
	Johnny	-7.5%	-6.6%	-5.2%	-5.6%	-5.7%	-4.9%
	KristenAndSara	-6.1%	-3.5%	-3.3%	-4.8%	-3.7%	-2.5%
Average	ClassA1	-4.1%	-0.3%	-2.1%	-3.3%	0.2%	-1.3%
	ClassA2	-5.6%	-2.3%	-2.3%	-4.9%	-1.9%	-2.1%
	ClassA	-4.6%	-2.3%	-2.0%	-3.8%	-3.1%	-2.8%
	ClassB	-4.3%	-2.2%	-2.1%	-2.8%	-2.3%	-2.2%
	ClassC	-2.6%	-1.7%	-1.4%	-2.3%	-1.6%	-1.8%
	ClassD	-2.0%	-1.2%	-0.8%	-1.4%	-0.6%	-0.8%
	ClassE	-6.4%	-4.5%	-4.1%	-5.0%	-4.6%	-3.9%
Overall		-4.1%	-2.0%	-2.0%	-3.2%	-1.8%	-2.0%

the relative positional relationship as illustrated in Fig.4. To balance the searching complexity and the predictor accuracy, we further constrain the IntraBC search range by a resolution dependent parameter s:

$$|BV_x| + |BC_y| < s \tag{5}$$

where, (BV_x, BV_y) is the horizontal and vertical displacement for the current CU, respectively. The total search displacement of both horizontal and vertical directions cannot exceed the limited range s. s is determined based on the width and height of the current frame:

$$s = max(128, 128 \cdot round(\sqrt{\frac{W \cdot H}{1920 \cdot 1080}})) \tag{6}$$

Compared with searching at the full available ranges, applying a limited search range has negligible influence on the performance.

There are two kinds of schemes to integrate the proposed learning based intra prediction method into HEVC. The first scheme is applying neural networks with different inputs to blocks with different texture characteristics. Our proposed CFPNN utilizes both local and nonlocal correlations to handle complex blocks containing complexity textures like mixed directional textures and recurrent patterns. Nevertheless, for those blocks which contain smooth content or single directional texture, accurate prediction can be generated by only using the neighboring multiple reference lines. There is no need to use nonlocal IntraBC predictor as

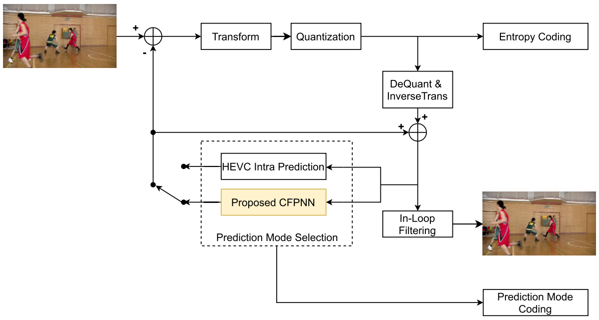


FIGURE 5. A brief diagram of integrating CFPNN into HEVC.

input for these blocks since there is additional overhead for singling the IntraBC related BV information. In this scheme, a lightweight neural network that only takes the neighboring multiple L-shape reference lines and consists of the related structure in Fig.2 is applied to generate intra prediction for those blocks with simple texture. This scheme is referred to as CFPNN-F with *F* represents Full. The other scheme is using only the proposed neural network with three inputs for all blocks without considering the block texture. This scheme is referred to as CFPNN-U with *U* represents Uniform. The proposed neural network based intra prediction method is selected by the rate-distortion optimization (RDO) against the traditional intra prediction. A brief diagram of integrated CFPNN in HEVC is shown in Fig.5. If the RD cost of the neural network based intra prediction is smaller than that of the traditional intra prediction, the CFPNN scheme is enabled. A CU level flag is adopted to indicate whether CFPNN is used. For the CFPNN-F scheme, an additional flag is used to distinguish which type of network is used. Since TU is the basic unit to conduct actual intra prediction, all TUs in a CU share the same intra prediction method. We train individual models for different TU sizes with luma component. When CFPNN is enabled, the corresponding model is applied to the TU luma component. Furthermore, the chroma components will also reuse the corresponding model trained with the luma component.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETTING

To conduct a comprehensive evaluation of our proposed method, the CFPNN has been integrated into the HEVC reference software HM-16.9 [47]. We use the All-Intra (AI) configuration defined by the HEVC Common Test Condition (CTC) as the test condition [48]. The sequences in CTC and 4K sequences in [49] are used as test sequences. Since we target improving the intra prediction efficiency of natural video content, the screen content video sequences are excluded from the test sequences. These test sequences contain a wide range of contents with various resolutions. There is no overlap between the test sequences and the training sequences. The first frame of these test sequences is coded with AI configuration under 4 QPs – {22, 27, 32, 37}. We use BD-rate to evaluate the coding performance [50]. For BD-rate, a negative value indicates performance improvement, and a positive number indicates performance loss.

TABLE 2. Comparison with previous work with only 8 × 8 intra prediction.

Sequence	IPFCN [33]	PSRNN [36]	App7 [28]	CFPNN
Traffic	-3.4%	-3.8%	-5.2%	-4.8%
PeopleOnStreet	-3.4%	-3.8%	-5.4%	-5.3%
Kimono	-7.8%	-6.6%	-10.9%	-7.5%
ParkScene	-3.3%	-3.4%	-4.4%	-3.9%
Cactus	-3.2%	-3.3%	-4.3%	-5.8%
BasketballDrive	-8.3%	-7.8%	-9.9%	-8.8%
BQTerrace	-1.5%	-2.6%	-3.2%	-4.6%
BasketballDrill	-1.0%	-2.9%	-1.6%	-5.1%
BQMall	-1.4%	-2.9%	-3.5%	-2.5%
PartyScene	-1.2%	-2.3%	-2.4%	-2.5%
RaceHorsesC	-2.6%	-2.8%	-3.1%	-3.2%
BasketballPass	-1.7%	-2.5%	-2.7%	-2.1%
BQSquare	-0.8%	-1.8%	-1.6%	-2.6%
BlowingBubbles	-1.5%	-2.3%	-2.7%	-1.7%
RaceHorses	-1.9%	-2.6%	-3.1%	-2.1%
FourPeople	-3.9%	-6.8%	-6.0%	-6.6%
Johnny	-7.6%	-5.6%	-8.6%	-9.7%
KristenAndSara	-5.7%	-6.6%	-6.7%	-10.0%
ClassA_Ave	-3.4%	-3.8%	-5.3%	-5.1%
ClassB_Ave	-4.8%	-4.7%	-6.5%	-6.1%
ClassC_Ave	-1.5%	-2.7%	-2.7%	-3.3%
ClassD_Ave	-1.5%	-2.3%	-2.5%	-2.1%
ClassE_Ave	-5.7%	-6.3%	-7.1%	-8.8%
Overall	-3.4%	-3.9%	-4.7%	-4.9%

B. COMPARISON WITH HEVC

The overall performance of our proposed CFPNN method compared with HEVC is summarized in Table.1. There are two integration schemes of our method as described in Section.III-E. Compared to HM-16.9, both of these schemes can significantly improve the intra prediction performance. For the luma component, CFPNN-F can achieve 4.1% BD-rate reduction on average, and the maximum BD-rate reduction can be up to 6.8%. The average BD-rate reduction of CFPNN-U is 3.2%, which is also remarkable. In order to keep the consistency of the encoder and decode side, the IntraBC related motion information BV also needs to be transmitted to the decoder side. The joint consideration of bitrate and distortion during RDO process will make some blocks not choosing the learning based prediction method when just applying uniform type of neural network in CFPNN-U. However, more blocks tend to use learning based prediction in the CFPNN-F scheme, where only neighboring reference lines are used for blocks with relatively simple texture while local/nonlocal correlations and unique characteristics of residual prior are used for complex blocks. Since the CFPNN-F scheme achieves better performance, we use this scheme to represent our proposed method and conduct other analyses. It can be observed that our proposed scheme works for all test sequences with varied resolutions

TABLE 3. Comparison with state-of-the-art methods under AI configuration[1].

Class	Cui *et al* [32]	Zhang *et al* [38]	IPFCN-S [33]	IPFCN-D [33]	PSRNN [36]	MSCNN [34]	CFPNN
ClassA	-1.1%	-1.1%	-3.7%	-4.4%	-3.6%	-4.0%	-4.6%
ClassB	-0.6%	-1.8%	-2.7%	-3.2%	-2.0%	-2.6%	-4.3%
ClassC	-0.6%	-1.2%	-1.9%	-2.1%	-2.3%	-3.4%	-2.6%
ClassD	-0.5%	-0.2%	-1.5%	-1.8%	-2.6%	-3.6%	-2.0%
ClassE	-0.7%	-2.0%	-4.2%	-4.5%	-3.5%	-4.2%	-6.4%
Overall	-0.7%	-1.3%	-2.6%	-3.0%	-2.8%	-3.4%	-3.8%

[1] Some results are copied from [34]

and contents. We apply the model trained for luma component to chroma components directly. The average BD-rate reduction for two chroma components is 2.0% and 2.0%, respectively. The coding performance of chroma components can be further improved by using models explicitly trained for these components.

Some typical Rate-Distortion (RD) curves are illustrated in Fig.6 to intuitively demonstrate the superior performance of our proposed scheme. It can be observed that CFPNN achieves better performance than original HEVC in the entire bitrate range. Specifically, the coding performance improvement at low bitrate range is higher than that at high bitrate range. The reason of this difference is the reconstructed reference lines at low bitrate always accompanied by severe distortion and the prediction generated by original HEVC intra prediction is inefficient. In this case, more blocks would choose our proposed learning based prediction method and the performance improvement would more significant.

C. COMPARISON WITH STATE-OF-THE-ART

In order to evaluate the efficiency of our proposed method, we also compare CFPNN with other learning based intra prediction methods [32]–[38]. These previous methods are either applied to fixed block size intra prediction or applied to default configuration with variable block size intra prediction. PSRNN [36] and App7 [28] are only applied to the 8 × 8 intra prediction. For a fair comparison, we also investigate the performance of our proposed scheme with only 8 × 8 intra prediction, where the intra prediction is limited to 8 × 8 TU. The detailed BD-rate comparison of 8 × 8 intra prediction is shown in Table.2. We also reproduce the IPFCN in [33] with our training data and test on the fixed-size block. IPCFN, PSRNN, and App7 can achieve 3.4%, 3.9% and 4.7% average BD-rate reduction for the luma component, respectively, while our method outperforms these methods and achieves 4.9% BD-rate reduction.

We also compare CFPNN with other methods under default AI configuration with variable block size intra prediction. The BD-rate reduction comparison on ClassA-ClassE is summarized in Table.3. The method in [32], method in [38], IPFCN-S, IPFCN-D, PSRNN and MSCNN [34] achieve average 0.7%, 1.3%, 2.6%, 3.0%, 2.8% and 3.4% BD-rate reduction, respectively. By contrast, the coding performance improvement of our scheme is 3.8%. The most related work of CFPNN is the method in [38], which also utilizes the IntraBC information. However, they just directly add the IntraBC predictor and the prediction generated by neighboring reference as a primary prediction, which is not an efficient way to utilize the nonlocal information. Our scheme outperforms these methods from two aspects. First, we introduce the other two reference priors to the intra prediction task, and these input priors are combined more efficiently with neighboring references by channel-wise attention mechanism. Second, we apply neural networks with different inputs to blocks with different texture characteristics.

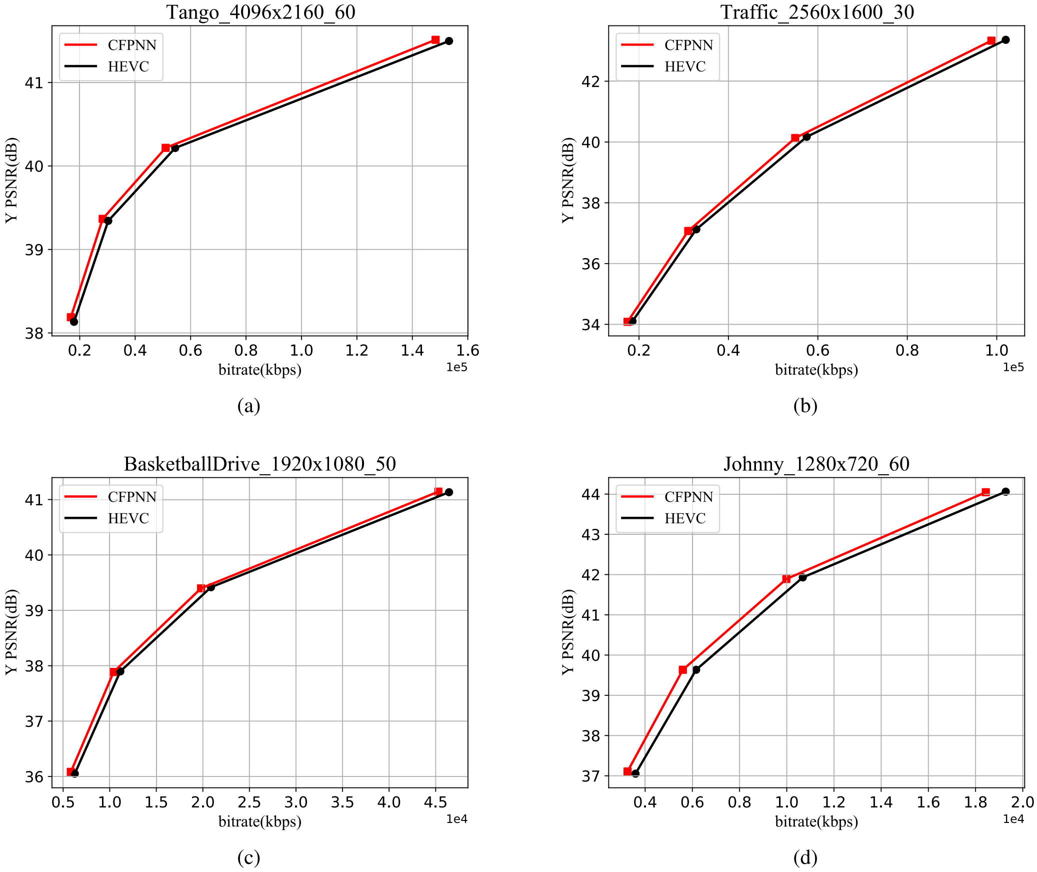
TABLE 4. BD-rate results of different input components combinations.

Class	CFPNN only Rec	CFPNN w/o Resi	CFPNN
ClassA	-3.2%	-4.2%	-4.6%
ClassB	-2.1%	-3.2%	-4.3%
ClassC	-2.3%	-3.0%	-2.6%
ClassD	-1.5%	-1.7%	-2.0%
ClassE	-4.5%	-5.0%	-6.4%
Overall	-2.5%	-3.4%	-3.8%

D. ANALYSIS OF DIFFERENT INPUT COMPONENTS

In our proposed CFPNN scheme, in addition to the neighboring multiple L-shape reference lines, we also employ the other two reference priors – *IntraBC* predictor and corresponding *residual* component to generate better intra prediction for complex blocks. To verify the effectiveness of these two priors, we evaluate the performance of different input components combinations. Specifically, we test the network with only neighboring references as input and network without using the *residual* component. The architecture of these two networks is composed of the corresponding branch in Fig.2. It should be noted that when using only neighboring references as input, the IntraBC related BV information is no longer transmitted.

Table.4 shows the BD-rate comparison of different input components combinations. The network with only neighboring reference lines achieves an average 2.5% BD-rate



(a)

(b)

(c)

(d)

FIGURE 6. Rate-distortion curves of some typical sequences.

reduction, which is basically the same as IPFCN-S in [33], in which the neighboring reference lines are combined with fully-connected layer instead of convolutional layer. The network using *IntraBC* prior helps to save an additional 0.9% BD-rate compared with only using neighboring reference lines. This is because the IntraBC predictor can introduce additional nonlocal information to generate a better prediction for complex blocks together with local information. The CFPNN with both priors achieves the best performance, which demonstrates that the unique characteristics contained in the *residual* prior is also helpful to enhance the quality of generated prediction.

E. COMPUTATIONAL COMPLEXITY

We summarize the computational complexity of our proposed scheme compared with HEVC anchor in Table.5, which is indicated by the encoding and decoding time. Both the neural network forward operation and the other operations of the codec are conducted on CPU. The encoding time increasing of our method is 4972% on average, and decoding time increasing is 19204%. At the encoder, our proposed neural network intra prediction is selected by the RDO against the original HEVC intra prediction. The time-consuming forward operation of the neural network is the main reason for encoding time increasing. Although the input *IntraBC* prior needed to be searched at the encoder, we have taken a trade-off between the IntraBC prediction accuracy and the computational complexity by limiting the search range of IntraBC based on (5). At the decoder, a CU level flag is first decoded to indicate whether CFPNN is selected as the intra prediction method. If CFPNN is used, the corresponding neural network model will be deployed. The time increasing of the decoder mainly depends on the ratio of CUs that selected CFPNN as intra prediction method.

TABLE 5. Computational complexity and BD-rate comparison.

Method	BD-Rate of AI			Time Complexity	
	Y	U	V	EncT	DecT
IPFCN	-3.0%	-1.5%	-1.6%	1745%	12347%
CFPNN-U	-3.2%	-1.8%	-2.0%	2617%	16003%
CFPNN	-4.1%	-2.0%	-2.0%	4972%	19204%

Another reason for the increased complexity is that we apply different types of neural networks to blocks with different characteristics. Two RDO processes are conducted at encoder to determine whether learning based intra prediction is used and which type of neural network is used. In the CFPNN-U scenario, all blocks use the uniform type of neural

TABLE 6. BD-rate comparison under different QP settings.

Class	Sequence	SmallQP	standardQP	LargeQP
ClassA1	Tango	-2.1%	-6.8%	-7.3%
	Drums100	-1.6%	-3.6%	-5.5%
	CampfireParty	-1.3%	-3.1%	-4.6%
	ToddlerFountain	-2.1%	-3.0%	-3.6%
ClassA2	CatRobot	-2.1%	-5.1%	-5.5%
	TrafficFlow	-2.0%	-6.8%	-7.6%
	DaylightRoad	-2.0%	-5.5%	-7.2%
	Rollercoaster	-2.9%	-5.0%	-5.1%
ClassA	Traffic	-1.8%	-4.1%	-5.3%
	PeopleOnStreet	-2.2%	-5.0%	-6.7%
ClassB	Kimono	-1.7%	-2.7%	-2.6%
	ParkScene	-1.5%	-3.1%	-3.2%
	Cactus	-1.7%	-5.0%	-7.5%
	BasketballDrive	-2.2%	-5.2%	-6.9%
	BQTerrace	-2.0%	-5.2%	-9.5%
ClassC	BasketballDrill	-1.3%	-3.4%	-4.6%
	BQMall	-1.0%	-2.2%	-4.3%
	PartyScene	-0.6%	-1.7%	-4.3%
	RaceHorsesC	-1.7%	-2.9%	-4.4%
ClassD	BasketballPass	-1.0%	-1.7%	-4.2%
	BQSquare	-0.8%	-1.2%	-3.2%
	BlowingBubbles	-0.6%	-1.8%	-2.7%
	RaceHorses	-1.6%	-3.3%	-5.2%
ClassE	FourPeople	-2.6%	-5.6%	-6.9%
	Johnny	-2.1%	-7.5%	-7.9%
	KristenAndSara	-2.1%	-6.1%	-8.8%
Average	ClassA1	-1.8%	-4.1%	-5.3%
	ClassA2	-2.3%	-5.6%	-6.4%
	ClassA	-2.0%	-4.6%	-6.0%
	ClassB	-1.8%	-4.3%	-6.0%
	ClassC	-1.1%	-2.6%	-4.4%
	ClassD	-1.0%	-2.0%	-3.8%
	ClassE	-2.3%	-6.4%	-7.9%
Overall		-1.7%	-4.1%	-5.6%

network, and only take one RDO process to select the best intra prediction method. The average time complexity of this scheme is also summarized in Table.5. It can be observed that the time complexity is reduced at the cost of reduction of performance increasing. In addition, the computational complexity of CFPNN-U is higher than IPFCN in [33], since the network structure is more complicated.

F. EXPLORATION ON DIFFERENT QP SETTINGS

In our proposed scheme, the codec with different QPs shares the same CFPNN model trained with mixed training data generated under standard QP setting – {22, 27, 32, 37}. To validate the generalization ability of our method, we further evaluate the coding performance under small QP setting – {11, 16, 21, 26} and large QP setting – {33, 38, 43, 48}. The coding performance of the luma component is summarized in Table. 6. The average BD-rate reduction under small QP setting and large QP setting is 1.7% and 5.6%, respectively.

Generally, small QP setting results in high bitrate while large QP setting results in low bitrate. It can be observed that the performance improvement at low bitrate case is more significant than the BD-rate reduction at high bitrate case. In the case of low bitrate, the single neighboring reconstructed reference line used by the original intra prediction method always suffer severe distortions, resulting in inaccurate prediction. By contrast, our method uses more local information and nonlocal information, which are helpful to generate a better prediction. The percentage of CUs choosing the proposed learning-based method is much higher at low bitrate case.

TABLE 7. Usage ratio of CFPNN for different QPs.

Class	QP22	QP27	QP32	QP37
ClassA1	48.05%	48.65%	50.85%	50.38%
ClassA2	46.75%	51.93%	53.12%	52.95%
ClassA	42.71%	52.13%	57.93%	59.17%
ClassB	42.97%	51.15%	55.82%	56.31%
ClassC	29.93%	36.29%	43.37%	46.79%
ClassD	28.57%	33.17%	35.13%	42.82%
ClassE	40.22%	46.59%	49.41%	50.45%
Overall	39.89%	45.70%	49.38%	51.27%

G. EVALUATION OF RDO USAGE RATIO

In our proposed scheme, the CU level flag is enabled to indicate whether the learning based intra prediction method is used for all TUs in this CU. We calculate the usage ratio of CFPNN to further clarify the effectiveness of our method. The usage ratio is defined as the ratio of total areas of all CUs that choosing learning based intra prediction to the area of the entire frame as (7):

$$r = \frac{\sum_{i=1}^{N} w_i \times h_i}{W \times H} \tag{7}$$

where W and H are the width and height of the frame. N is the total number of CUs choosing proposed scheme, w_i and h_i is the corresponding width and height of CU, which is 8, 16, 32.

Table.7 summarizes the average usage ratio of each class for 4 different QPs under AI configuration. With the increase of QP, the average usage ratio also increases. This is consistent with previous BD-rate performance. The BD-rate reduction at large QP range (low bitrate) is relatively high, while the average usage ratio of the proposed learning based method is also high.

V. CONCLUSION

In this paper, we propose a Current Frame Priors assisted Neural Network (CFPNN) for intra prediction of video coding. In our proposed method, we also use the neighboring multiple L-shape reference lines of the current block, which containing rich local contextual information as the primitive inference source. In addition to the neighboring reference lines, we introduce the other two reference priors within the current frame – the best matching predictor searched by IntraBC and the corresponding residual component to make use of the nonlocal correlation and the unique texture information contained in the residual component. The network architecture is carefully designed to extract features from these input components simultaneously and fuse them efficiently. We also investigate different schemes when integrating the proposed CFPNN into the codec. Compared with the HEVC reference software, our proposed learning based intra prediction method achieves an average of 4.1% BD-rate reduction under All Intra configuration.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[2] 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.

[3] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1792–1801, Dec. 2012.

[4] B. Bross, J. Chen, S. Liu, and Y. Wang, *Versatile Video Coding Editorial Refinements on Draft 10*, document JVET-T2001, Joint Video Experts Team (JVET), 2020.

[5] R. Wei, R. Xie, L. Song, L. Zhang, and W. Zhang, "Improved intra angular prediction with novel interpolation filter and boundary filter," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.

[6] H. Chen, T. Zhang, M.-T. Sun, A. Saxena, and M. Budagavi, "Improving intra prediction in high-efficiency video coding," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3671–3682, Aug. 2016.

[7] S. Li, Y. Chen, J. Han, T. Nanjundaswamy, and K. Rose, "Rate-distortion optimization and adaptation of intra prediction filter parameters," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 3146–3150.

[8] J. Li, B. Li, J. Xu, and R. Xiong, "Efficient multiple-line-based intra prediction for HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 4, pp. 947–957, Apr. 2018.

[9] X. Qi, T. Zhang, F. Ye, A. Men, and B. Yang, "Intra prediction with enhanced inpainting method and vector predictor for HEVC," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1217–1220.

[10] C.-H. Yeh, T.-Y. Tseng, C.-W. Lee, and C.-Y. Lin, "Predictive texture synthesis-based intra coding scheme for advanced video coding," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1508–1514, Sep. 2015.

[11] C. Chen, S. Zhu, B. Zeng, and M. Gabbouj, "A new block-based method for HEVC intra coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1727–1736, Aug. 2017.

[12] T. Zhang, X. Fan, D. Zhao, R. Xiong, and W. Gao, "Hybrid intraprediction based on local and nonlocal correlations," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1622–1635, Jul. 2018.

[13] H. Chen, Y.-S. Chen, M.-T. Sun, A. Saxena, and M. Budagavi, "Improvements on intra block copy in natural content video coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2015, pp. 2772–2775.

[14] Y. Li, L. Li, D. Liu, H. Li, and F. Wu, "Combining directional intra prediction and intra block copy with block partition for HEVC," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 524–528.

[15] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11006–11015.

[16] J. Lin, D. Liu, H. Li, and F. Wu, "M-LVC: Multiple frames prediction for learned video compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3546–3554.

[17] H. Liu, H. Shen, L. Huang, M. Lu, T. Chen, and Z. Ma, "Learned video compression via joint spatial-temporal correlation exploration," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, hboxpp. 11580–11587.

[18] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6628–6637.

[19] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding: A review and a case study," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–35, 2020.

[20] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, and J. Yang, "Enhanced bi-prediction with convolutional neural network for high-efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3291–3301, Nov. 2019.

[21] J.-K. Lee, N. Kim, S. Cho, and J.-W. Kang, "Deep video prediction network-based inter-frame coding in HEVC," *IEEE Access*, vol. 8, pp. 95906–95917, 2020.

[22] H. Zhang, L. Song, L. Li, Z. Li, and X. Yang, "Compression priors assisted convolutional neural network for fractional interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1953–1967, May 2021.

[23] D. Liu, H. Ma, Z. Xiong, and F. Wu, "CNN-based DCT-like transform for image compression," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2018, pp. 61–72.

[24] C. Ma, D. Liu, X. Peng, L. Li, and F. Wu, "Convolutional neural network-based arithmetic coding for HEVC intra-predicted residues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1901–1916, Jul. 2020.

[25] R. Song, D. Liu, H. Li, and F. Wu, "Neural network-based arithmetic coding of intra prediction modes in HEVC," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[26] M. Zhou, X. Wei, S. Kwong, W. Jia, and B. Fang, "Rate control method based on deep reinforcement learning for dynamic video sequences in HEVC," *IEEE Trans. Multimedia*, vol. 23, pp. 1106–1121, 2021.

[27] J.-H. Hu, W.-H. Peng, and C.-H. Chung, "Reinforcement learning for HEVC/H.265 intra-frame rate control," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.

[28] X. He, Q. Hu, X. Zhang, C. Zhang, W. Lin, and X. Han, "Enhancing HEVC compressed videos with a partition-masked convolutional neural network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 216–220.

[29] J. Guo and H. Chao, "Building dual-domain representations for compression artifacts reduction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 628–644.

[30] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma, "Content-aware convolutional neural network for in-loop filtering in high efficiency video coding," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3343–3356, Jul. 2019.

[31] M.-Z. Wang, S. Wan, H. Gong, and M.-Y. Ma, "Attention-based dual-scale CNN in-loop filter for versatile video coding," *IEEE Access*, vol. 7, pp. 145214–145226, 2019.

[32] W. Cui, T. Zhang, S. Zhang, F. Jiang, W. Zuo, Z. Wan, and D. Zhao, "Convolutional neural networks based intra prediction for HEVC," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, p. 436.

[33] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3236–3247, Jul. 2018.

[34] Y. Wang, X. Fan, S. Liu, D. Zhao, and W. Gao, "Multi-scale convolutional neural network-based intra prediction for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1803–1815, Jul. 2020.

[35] H. Sun, Z. Cheng, M. Takeuchi, and J. Katto, "Enhanced intra prediction for video coding by using multiple neural networks," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2764–2779, Nov. 2020.

[36] Y. Hu, W. Yang, M. Li, and J. Liu, "Progressive spatial recurrent neural network for intra prediction," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3024–3037, Dec. 2019.

[37] T. Dumas, A. Roumy, and C. Guillemot, "Context-adaptive neural network-based prediction for image compression," *IEEE Trans. Image Process.*, vol. 29, pp. 679–693, 2020.

[38] Z. Zhang, Y. Li, L. Li, Z. Li, and S. Liu, "Combining intra block copy and neighboring samples using convolutional neural network for image coding," in *Proc. Vis. Commun. Image Process. (VCIP)*, 2018, pp. 1–4.

[39] L. Ma, Y. Tian, and T. Huang, "Residual-based video restoration for HEVC intra coding," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–7.

[40] X. Xu, S. Liu, T.-D. Chuang, Y.-W. Huang, S.-M. Lei, K. Rapaka, C. Pang, V. Seregin, Y.-K. Wang, and M. Karczewicz, "Intra block copy in HEVC screen content coding extensions," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 6, no. 4, pp. 409–419, Dec. 2016.

[41] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[44] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 34–35.

[45] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 126–135.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] *HMSoftware*. [Online]. Available: https://vcgit.hhi.fraunhofer.de/jvet/HM/-/tree/HM-16.9

[48] F. Bossen, *Common Test Conditions and Software Reference Configurations*, document JCTVC-L1100, 2013, vol. 12.

[49] J. Boyce, K. Suehring, X. Li, and V. Seregin, *JVET Common Test Conditions and Software Reference Configurations*, document JVET-J1010, 2018.

[50] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, 2001, pp. 2–4.

HAN ZHANG received the B.S. degree in information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include image/video coding and processing. He received the Best 10% Paper Award at the 2016 IEEE Visual Communications and Image Processing (VCIP).

LI SONG (Senior Member, IEEE) received the B.E. and M.S. degrees in engineering, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University (SJTU), in 2005. He joined SJTU, as a Faculty Member, where he is currently a Full Professor with the Department of Electronic Engineering. He was a Visiting Professor with Santa Clara University, from 2011 to 2012. He has more than 200 publications and 40 granted patents, and 18 standards technical proposals in field of video coding and image processing. His research interests include video processing and multimedia systems.

YAN HUANG received the B.S. degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. From 2018 to 2019, he was a Research Associate with the MultiMedia & Vision Research Group, Queen Mary University of London. His research interests include video coding/processing and machine learning.

RONG XIE (Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Northeast Dianli University, Jilin, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2002. She joined Shanghai Jiao Tong University, as an Assistant Professor, in 2002. From April 2009 to November 2009, she visited the Information Laboratory (InfoLAB), University of Southern California, USA, as a Visiting Scholar. From December 2009 to December 2010, she worked with the IT Center, Shanghai World Expo Coordination Bureau. She is currently an Associate Professor with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. Her research interests include video coding and processing. She is a member of the IEEE Signal Processing Society.