A Low Complexity Convolutional Neural Network with Fused CP Decomposition for In-Loop Filtering in Video Coding

Data Compression Conference 2023

Tong Shao[1], Jay N. Shingala[2], Peng Yin[1], Arjun Arora[1], Ajay Shyam[2], and Sean McCarthy[1]

[1]Dolby Laboratories, Inc., USA, [2]Ittiam Systems Pvt. Ltd., India

Outline

- Introduction
  - Background
  - Motivation
- Proposed model: Convolutional Neural Network with Fused CP Decomposition (CP Fused)
- Experimental results
- Conclusions

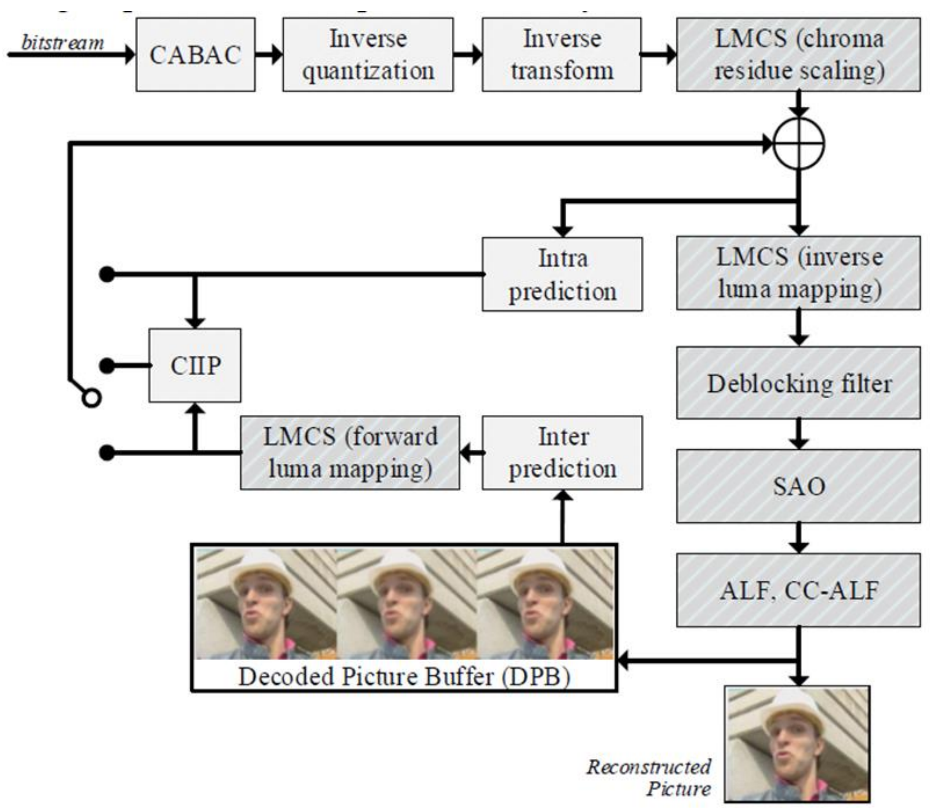
Background

- Block-based coding: prediction, transform, quantization
- Coding artifacts
  - Blocking artifact
  - Ringing artifact
  - . . .

Blocking artifact

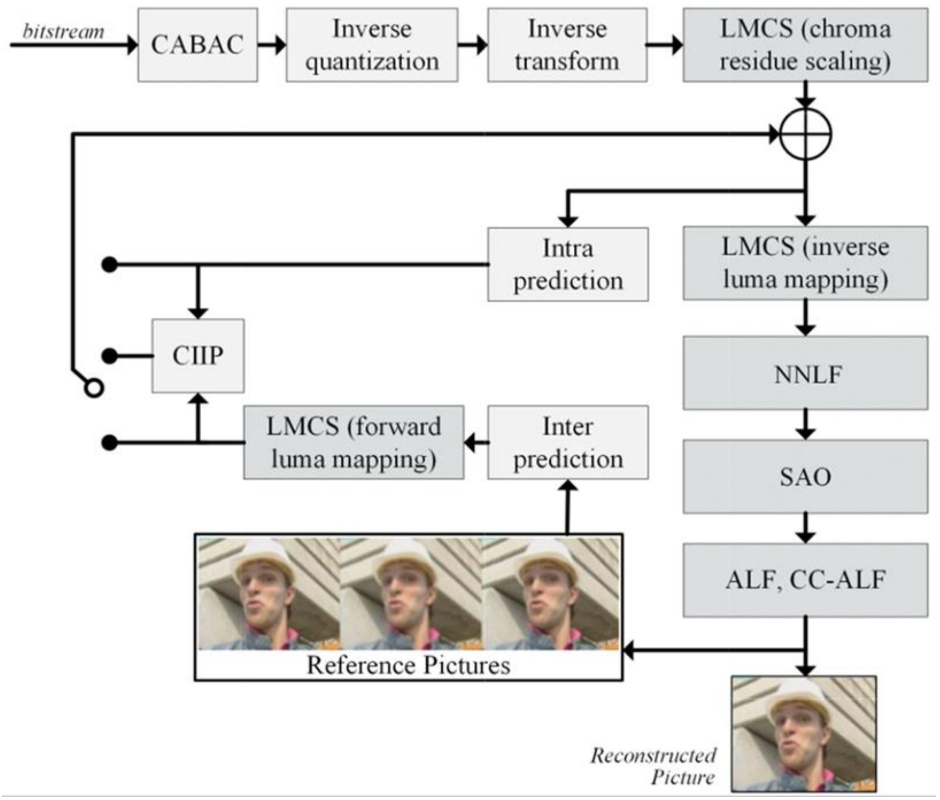
Background

- In-loop filters in VVC (Versatile Video Coding)
- Filters:
  - DBF (Deblocking filter): low-pass filters
  - SAO (Sample adaptive offset)
  - ALF (Adaptive loop filtering): Wiener Filter
  - CC-ALF (Cross-component ALF)



Background

- ❑ NNLF: Neural network-based loop filter
- ❑ Convolutional neural network (CNN)
- ❑ Supervised learning
    - ❑ Input: compressed samples with artifacts
    - ❑ Targeted output: uncompressed original samples
- ❑ Replace the traditional DBF
- ❑ Placed before SAO and ALF/CC-ALF



Background

- Many NNLF publications and contributions in JVET EE1 (Exploration Experiment)
- VVC reference software: VTM-11.0-nnvc (NNVC 1.0)
- NNVC 3.0 (NCS 1.0): two NNLF filters with best performance have been adopted in the ref. software

	Parameters (M)	KMAC/Pixel (K)	BD-Rate (%), RA	BD-Rate (%), AI
NCS#0, JVET-AA0088	1.90	485	-8.71	-6.52
NCS#1, JVET-AA0111	3.12	539	-9.44	-7.26

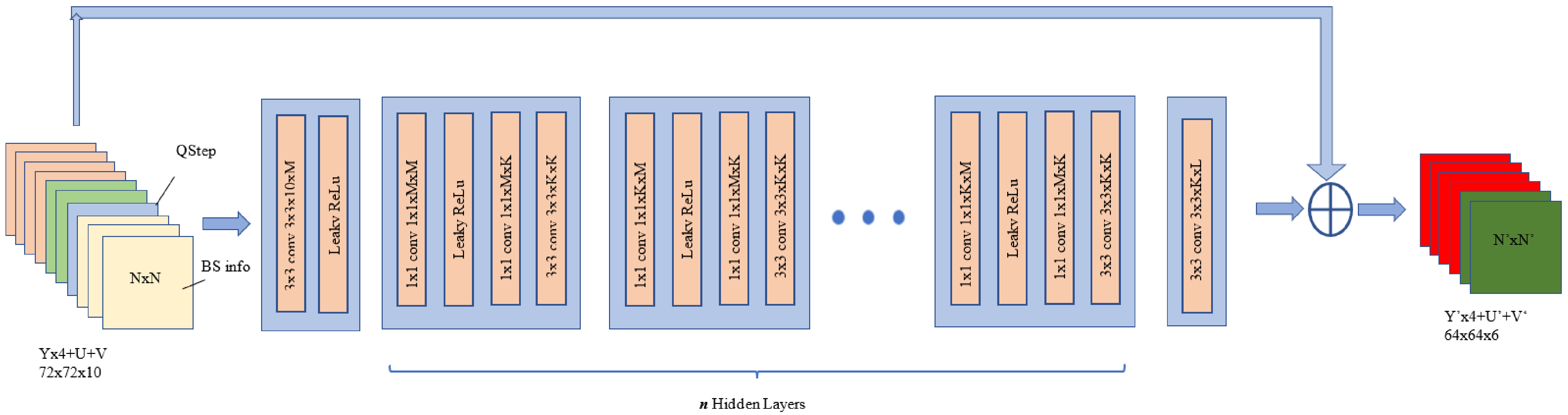
Motivation

- Two filters in NNVC 3.0 (NCS 1.0) are of relevant high complexity, though the gain is around 9%.
- KMAC/Pixel is more than 400, while CPU decoding time is several hundreds times of the VTM anchor.
- Difficult to be deployed in real-world applications.

	Parameters (M)	KMAC/Pixel (K)	BD-Rate (%), RA	BD-Rate (%), AI
NCS#0, JVET-AA0088	1.90	485	-8.71	-6.52
NCS#1, JVET-AA0111	3.12	539	-9.44	-7.26

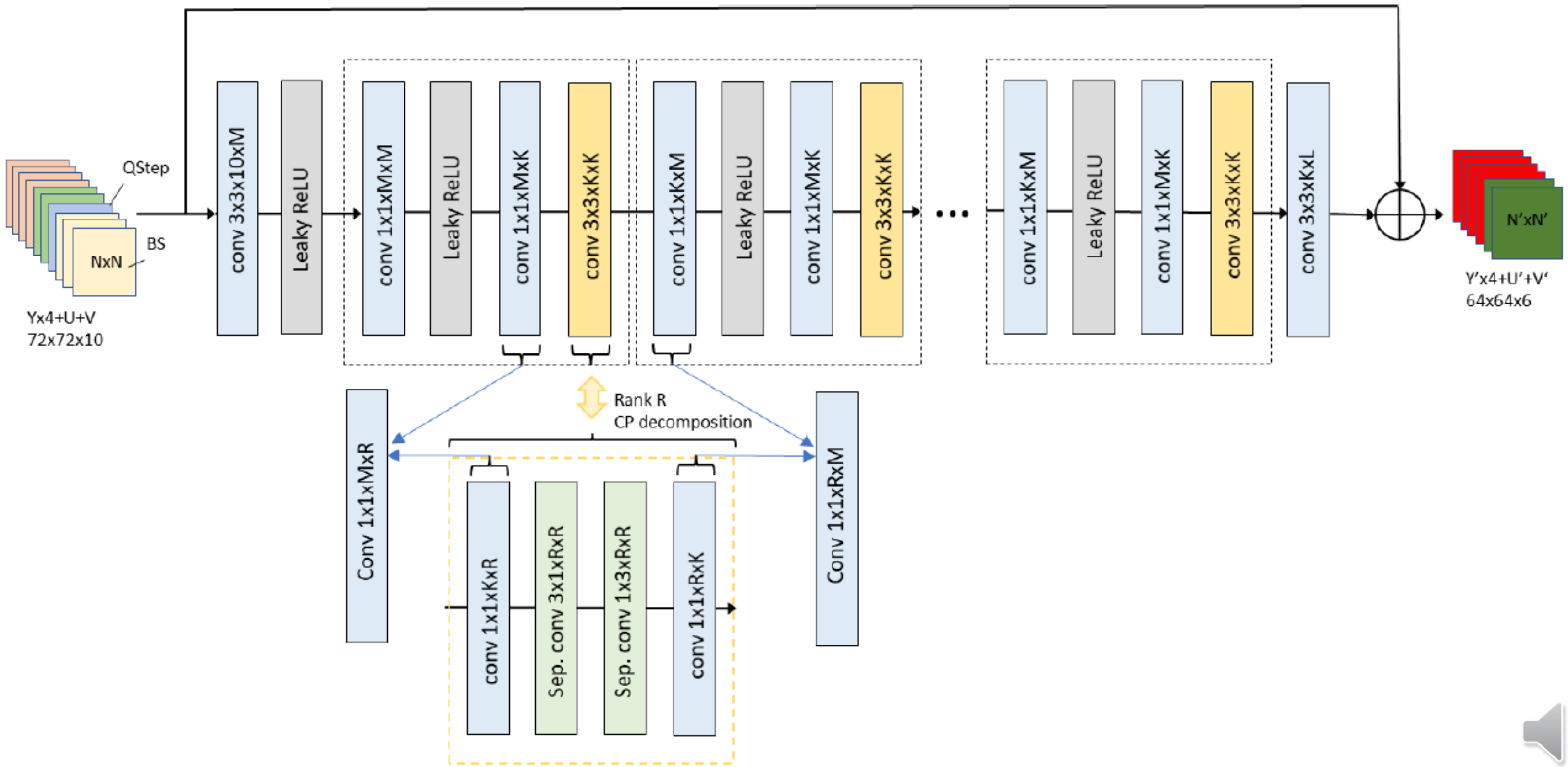
Proposed model: baseline model

- Baseline model: JVET-X0140 CNN based model, which has around 5% gain with 33.6 KMAC/Pixel
- Inputs: 4 luma tensors, 2 chroma, 1 Quantization Step, 3 Boundary Strength

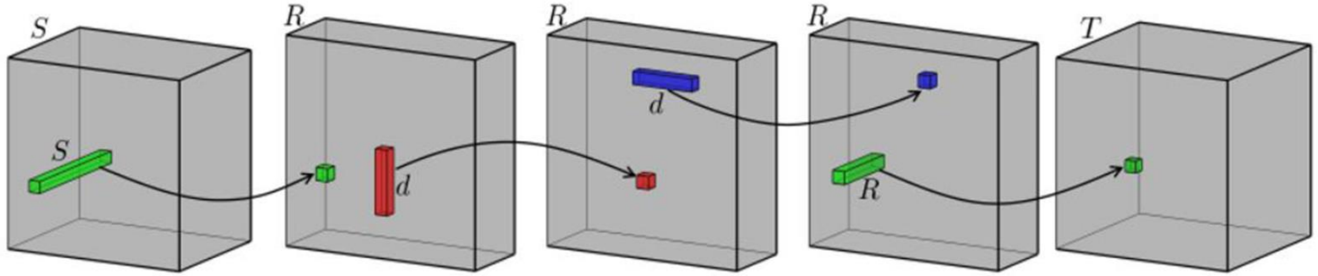


Proposed model: Convolutional Neural Network with Fused CP Decomposition (CP Fused)

- The 3x3 convolutions of each hidden layer are decomposed into 4 layers with rank R, i.e., CP decomposition:
  - 1st layer: 1x1xKxR pointwise convolution
  - 2nd layer: 3x1xRxR separable convolution
  - 3rd layer: 1x3xRxR separable convolution
  - 4th layer: 1x1xRxK pointwise convolution
- K=24, M=72, R=24, L=6, n = 11



Proposed model: CP Decomposition



- Regular convolution for output channel t can be written as:

$$V(x,y,t) = \sum_{i=x-\delta}^{x+\delta} \sum_{j=y-\delta}^{y+\delta} \sum_{s=1}^{S} K(i-x+\delta, j-y+\delta, s, t)\, U(i,j,s)$$

- where, U is input tensor with S channels, K is kernel of size δxδxS per output channel, and V is output tensor.

The CP rank R approximation for the above convolution for channel t can be written as:

$$V(x,y,t) = \sum_{r=1}^{R} K^t(t,r) \left(\sum_{i=x-\delta}^{x+\delta} K^x(i-x+\ \delta, r) \left(\sum_{j=y-\delta}^{y+\delta} K^y(j-y+\delta, r) \left(\sum_{s=1}^{S} K^s(s,r)\, U(i,j,s) \right) \right) \right)$$

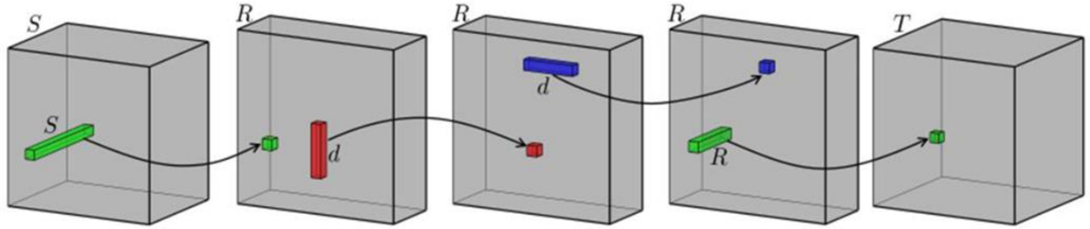
Where kernel K is approximated as

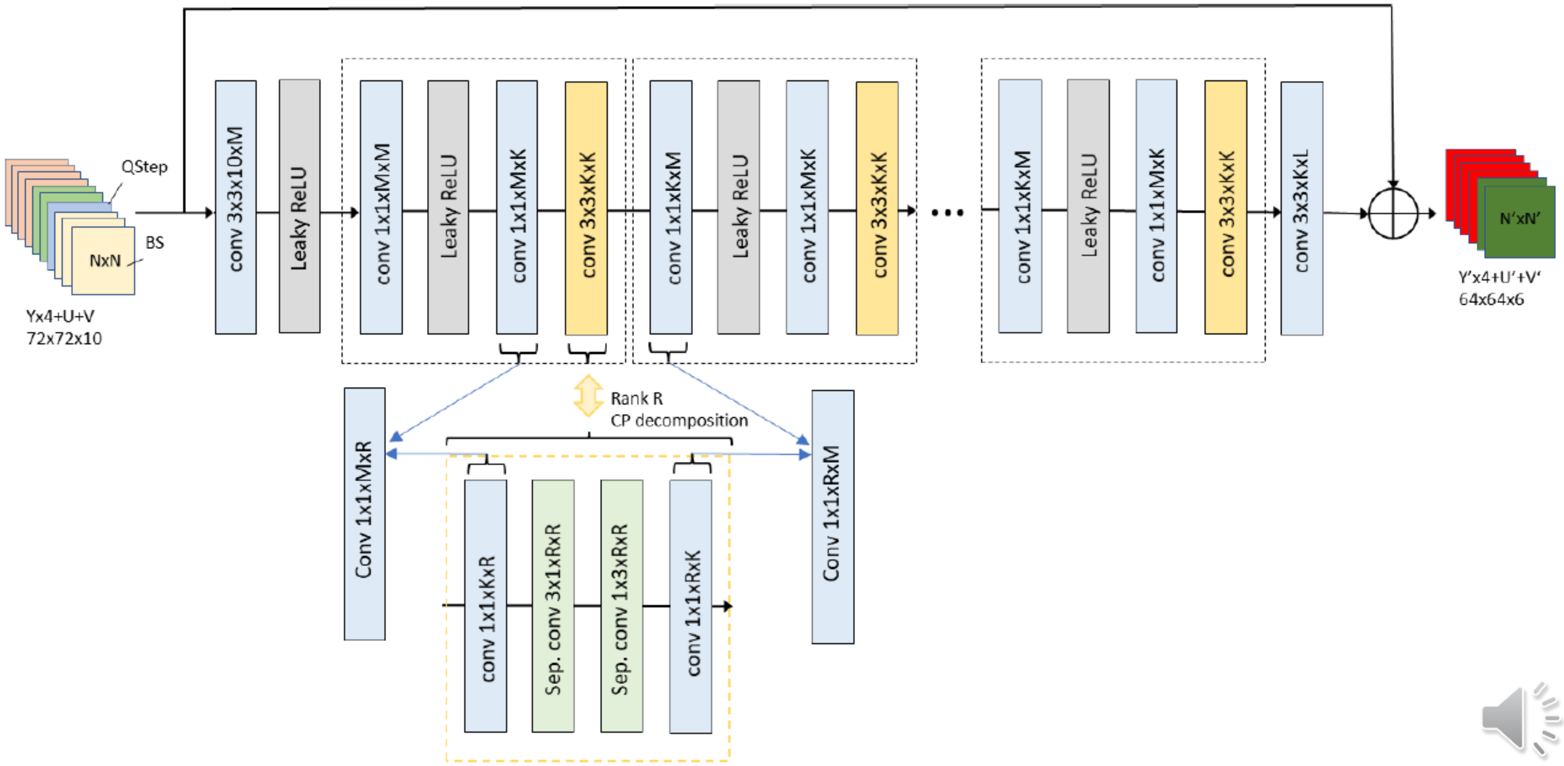
$K(i,j,s,t) = \sum_{r=1}^{R} K^x(i-x+\delta,r)\, K^y(j-y+\delta,,r)\ \ K^s(s,r)\ \ K^t(t,r)$and K^x, K^y, K^s, K^t are δxR, δxR, SxR, TxR tensors along different dimensions.

The complexity of CP decomposition in terms of MAC/pixel is $R(S + 2\delta + T)$ as compared to $ST\delta^2$ for the regular convolution.

Proposed model: Convolutional Neural Network with Fused CP Decomposition (CP Fused)

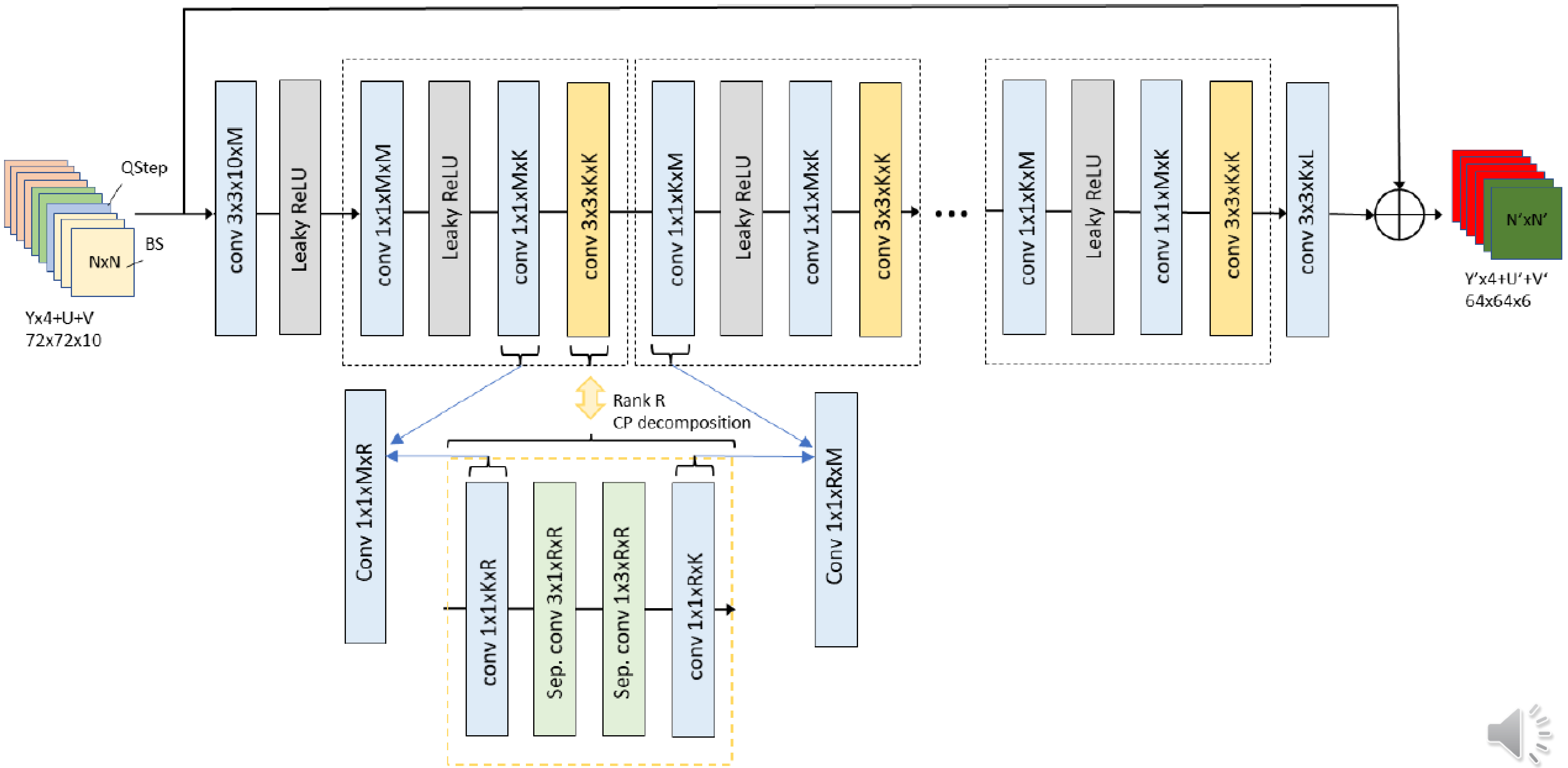
- CP decomposition:
  - 1st layer: 1x1xKxR pointwise convolution
  - 2nd layer: 3x1xRxR separable convolution
  - 3rd layer: 1x3xRxR separable convolution
  - 4th layer: 1x1xRxK pointwise convolution
- MACs/Pixel = 20.093 KMAC/Pixel (33.6 before)





Proposed model: Convolutional Neural Network with Fused CP Decomposition (CP Fused)

- Fusion of adjacent convolutional layers
- MACs/Pixel = 16.265 KMAC/Pixel



Experimental results

- Model trained using DIV2K dataset for AI and BVI-DVC dataset for RA, Tensorflow
- VVC reference software VTM-11.0-nnvc (NNVC-1.0)
- RD performance and CPU decoding time comparisons
- CTC test sequences, All Intra (AI) and Random Access (RA)

Experimental results

- Compared to NNVC-2.0 anchor
- 4.45%, 5.68%, 5.19% (Y, U, V, respectively) for RA
- 4.68%, 5.72%, 4.81% (Y, U, V, respectively) for AI

Table 1: BD-Rate (%) of the proposed fused CP decomposition model compared to VTM NNVC-2.0 anchor, under RA and AI configurations. Negative value means coding gain.

Class	Random Access			All Intra		
	Y	U	V	Y	U	V
A1	-4.88%	-3.60%	-4.36%	-4.56%	-4.48%	-4.23%
A2	-4.57%	-4.95%	-3.70%	-4.24%	-5.95%	-4.69%
B	-4.07%	-6.08%	-5.93%	-4.23%	-5.90%	-4.98%
C	-4.52%	-7.27%	-6.01%	-4.52%	-6.76%	-5.08%
D	-5.90%	-6.23%	-6.70%	-4.94%	-5.07%	-4.86%
E	-	-	-	-6.21%	-5.02%	-4.88%
Overall	-4.45%	-5.68%	-5.19%	-4.68%	-5.72%	-4.81%

Experimental results

- Compared to baseline model JVET-X0140
- 0.56% luma loss, while decoding time is reduced by 19%

Table 2: BD-Rate (%) and CPU decoding time increase (%) of the proposed fused CP decomposition model compared to JVET-X0140 baseline model, under RA and AI configurations. Negative value means coding gain.

Class	Random Access				All Intra			
	Y	U	V	ΔDecT	Y	U	V	ΔDecT
A1	0.13%	-0.67%	-1.74%	-19%	0.17%	0.16%	0.16%	-24%
A2	0.54%	-0.72%	-1.45%	-19%	0.57%	-0.27%	-0.41%	-24%
B	0.62%	-0.84%	-2.46%	-19%	0.54%	0.36%	0.35%	-24%
C	0.82%	-0.26%	-1.62%	-18%	0.52%	-0.04%	0.33%	-24%
D	0.88%	0.55%	-1.52%	-19%	0.45%	0.18%	-0.05%	-26%
E	-	-	-		0.69%	0.82%	1.58%	-24%
Overall	0.56%	-0.63%	-1.89%	-19%	0.51%	0.21%	0.39%	-24%

Experimental results

- ❑ Better trade-off
- ❑ Compared with the ones with best RD performance, the proposed model can provide about half of the coding gain with 3% of the complexity (16.265 KMAC/Pixel).

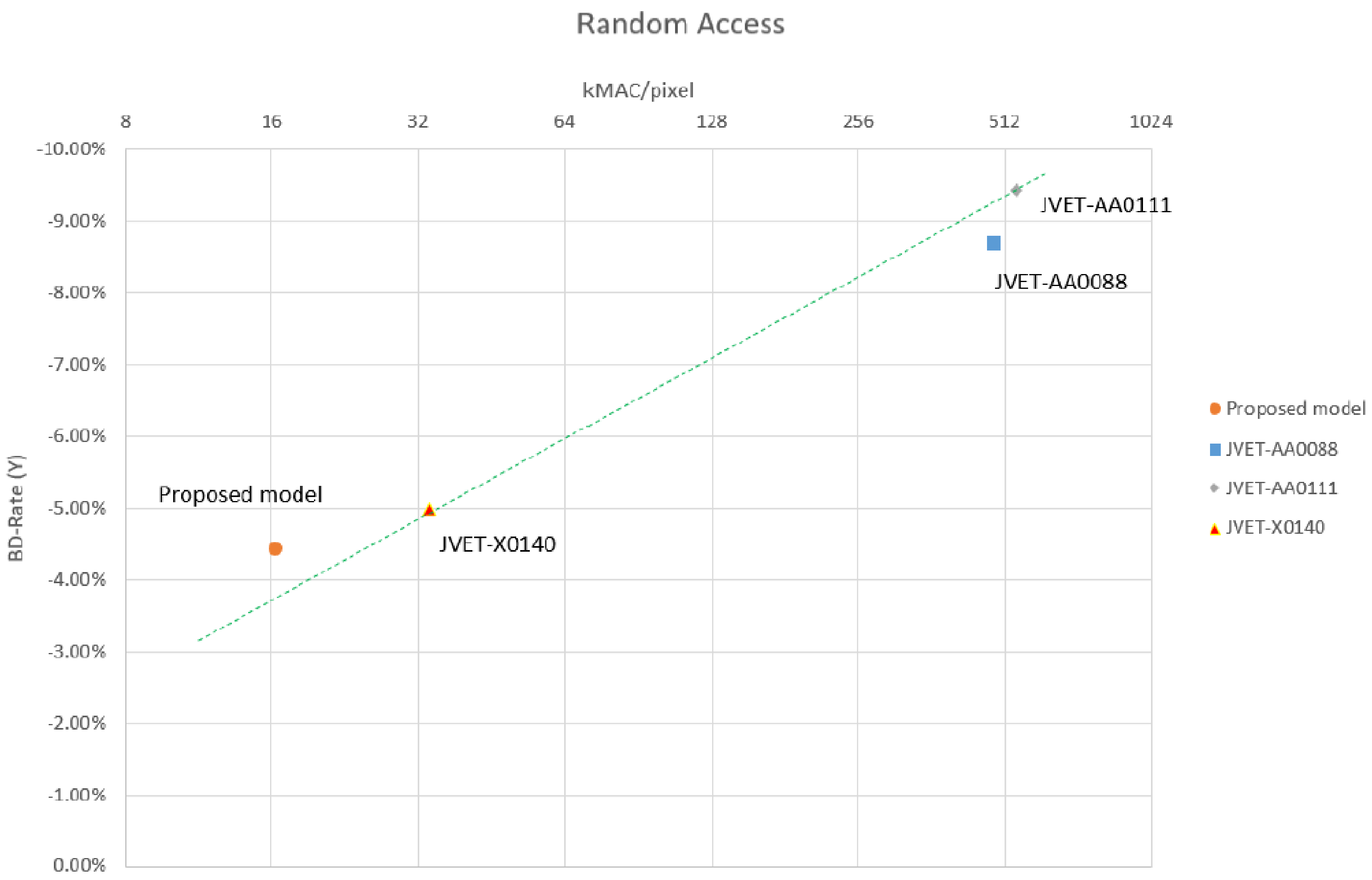


Figure 3: Complexity vs. gain trade-off comparisons of the state-of-the-art NNLF models under RA.

Conclusions

- Our proposed CP fused NNLF model provides 4.45% Luma gain with 16.265 KMAC/Pixel under NNVC-2.0 anchor.

- Compared to JVET-X0140, it shows 0.56% Luma loss, while decoding time is reduced by 19%.

- Compared to the 2 best performance filters in JVET NNVC-3.0, our model have only 3% of the complexity (KMACs) while maintain half of their coding gains.

- Our proposed model has a better BD-Rate vs. complexity trade-off according to the plot.

Q&A

Thanks!