

# **AHG11: Complexity Reduction on Neural-Network Loop Filter**

**JVET-AB0136**

(related to JVET-AA0080)



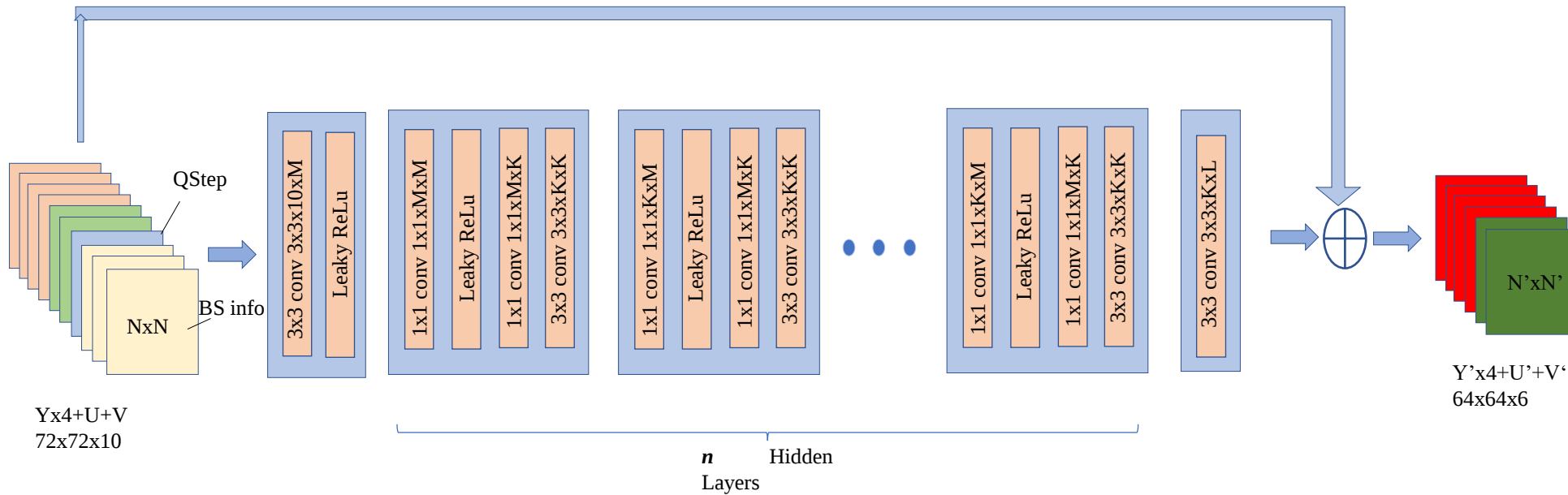
# Explore new aspects of reducing NNLF complexity

**Motivation:** Reduce Neural Network Loop Filter complexity

**This contribution studies following 3 methods to reduce NNLF complexity**

- **CP-decomposition: Canonical Polyadic Decomposition**
  - Also referred to as CANDECOMP/PARAFAC
  - Flexibility to select optimal rank for good trade-off (complexity vs coding gain)
- **Fuse adjacent 1x1 convolutions**
  - Applied after CP-decomposition to merge adjacent linear layers
- **New luma-chroma split NNLF architecture**
  - Provides comparable luma gains
  - Provides higher chroma gains
  - Reduces overall complexity
- All the above 3 aspects studied in previous meeting JVET-AA0080
  - AI results presented in previous meeting, RA results presented in this meeting
  - JVET-X0140 low complexity model used as baseline model

# Overview of JVET-X0140 baseline architecture

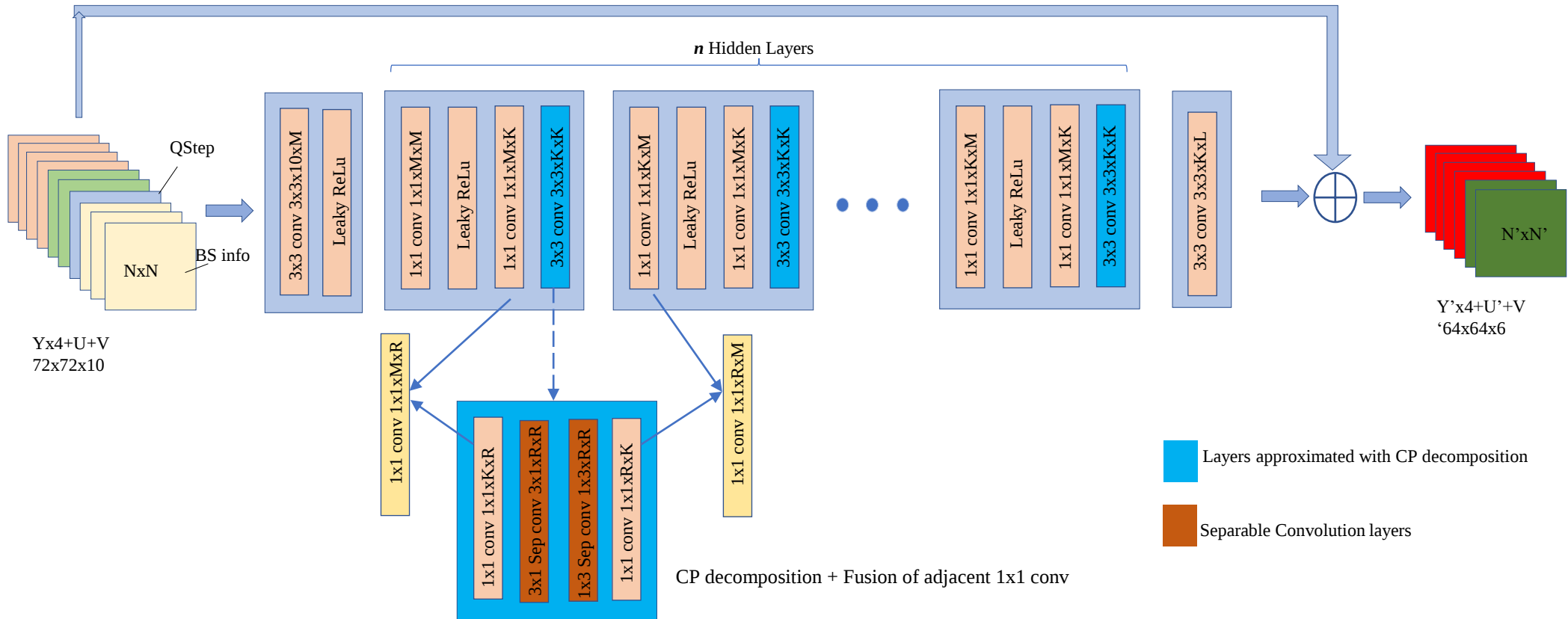


- Low complexity Joint Luma-Chroma Filter
- Placed before SAO, deblocking is disabled
- Inputs:  $72 \times 72 \times 10$  tensor
  - $144 \times 144$  YUV420 block converted to  $72 \times 72 \times 6$  tensor (4 luma, 2 chroma)
  - BS map and QP map
- Major complexity is due  $3 \times 3$  convolution
  - $3 \times 3$  convolution MACs =  $3 \times 3 \times K \times K$

## Model parameters and Complexity

- $M = 72$ ,  $K = 24$
- $n = 11$  (number of hidden layers)
- **Model complexity = 33.6 kMAC/pixel**
- 108k params per model
- 4 models (2 models for Intra slice)

# CP-decomposition and fusion architecture



3x3 convolutions of each hidden layer are decomposed into 4 layers with rank R:

- 1st layer:  $1 \times 1 \times K \times R$  pointwise convolution
- 2nd layer:  $3 \times 1 \times R \times R$  separable convolution
- 3rd layer:  $1 \times 3 \times R \times R$  separable convolution
- 4th layer:  $1 \times 1 \times R \times K$  pointwise convolution

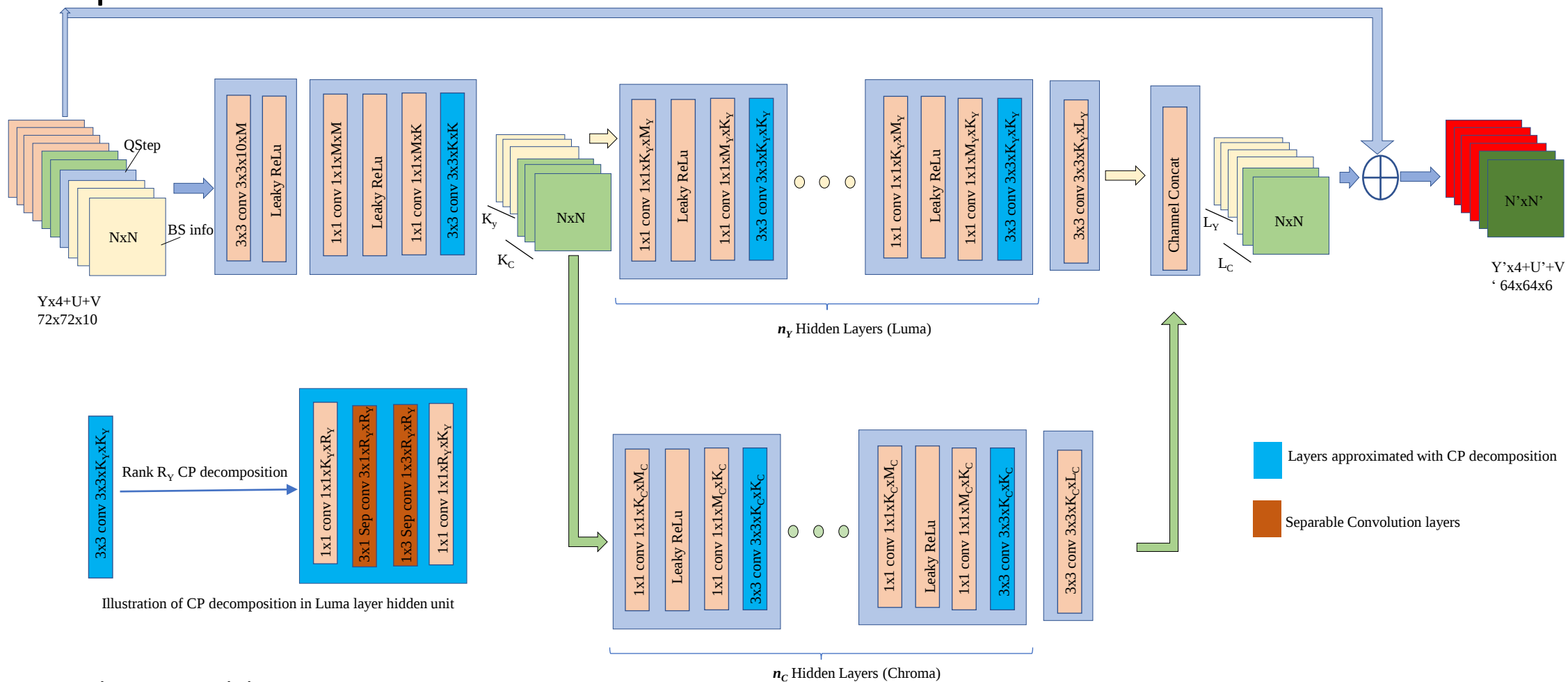
Fusion of adjacent  $1 \times 1$  conv layers

- No ReLu layers before and after decomposed layer
- 1<sup>st</sup> decomposed layer merged with previous  $1 \times 1$  conv
- 4<sup>th</sup> decomposed layer merged with next  $1 \times 1$  conv
- Only separable convolution layers left after fusion

Model parameters and Complexity

- $M = 72$ ,  $K = 24$ ,  $R = 24$
- $n = 11$  (number of hidden layers)
- Model complexity = 16.2 kMAC/pixel
- 52.5k params per model

# Split Luma-Chroma Architecture



## Split Luma-Chroma Model Motivation

- Initial layers process luma and chroma jointly: Extract cross dependencies
- Split luma and chroma filtering subsequently: Independent filtering uses reduced number of features
- Different combinations of  $K_Y$ ,  $K_C$ ,  $M_Y$ ,  $M_C$ ,  $n_Y$ ,  $n_C$  provide good trade-offs
- Apply CP-decomposition and 1x1 fusion for all hidden layers of luma and chroma

| Split luma-chroma model parameters<br>$M_Y=3*K_Y$ $M_C=3*K_C$ $n_Y=n_C=10$ |            |
|----------------------------------------------------------------------------|------------|
| $(K_Y, K_C)$                                                               | KMAC/Pixel |
| (24L, 8C)                                                                  | 17.7       |
| (20L, 8C)                                                                  | 14.2       |
| (16L, 8C)                                                                  | 11.3       |

# JVET-X0141-1 Baseline results

- Summary of JVET-X0140 EE1-1.4.1 RA results

| Random Access Main10        |        |        |        |      |       |
|-----------------------------|--------|--------|--------|------|-------|
| BD-rate Over VTM-11.0+V0056 |        |        |        |      |       |
|                             | Y-PSNR | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                    | -5.95% | -4.46% | -3.98% | 131% | 4371% |
| Class A2                    | -5.96% | -5.91% | -3.61% | 120% | 3776% |
| Class B                     | -5.51% | -7.19% | -5.43% | 123% | 4311% |
| Class C                     | -5.88% | -8.72% | -6.38% | 118% | 3886% |
| Class E                     |        |        |        |      |       |
| <b>Overall</b>              | -5.78% | -6.80% | -5.03% | 123% | 4095% |
| Class D                     | -7.21% | -8.37% | -6.50% | 115% | 4539% |
| Class F                     | -2.80% | -5.78% | -4.42% | 127% | 5012% |
| Class H                     | -5.56% | -5.35% | -5.47% | 129% | 5100% |

| Random access Main10                |        |        |        |      |             |
|-------------------------------------|--------|--------|--------|------|-------------|
| BD-rate Over VTM11_nnnvc-2.0 anchor |        |        |        |      |             |
|                                     | Y-PSNR | U-PSNR | V-PSNR | EncT | DecT (Ours) |
| Class A1                            | -5.01% | -2.96% | -2.67% |      | 4321%       |
| Class A2                            | -5.08% | -4.27% | -2.28% |      | 4485%       |
| Class B                             | -4.66% | -5.29% | -3.55% |      | 4570%       |
| Class C                             | -5.29% | -7.00% | -4.44% |      | 4267%       |
| Class E                             |        |        |        |      |             |
| <b>Overall</b>                      | -4.98% | -5.08% | -3.36% |      | 4401%       |
| Class D                             | -6.74% | -6.73% | -5.25% |      | 4647%       |
| Class F                             | -2.32% | -4.66% | -3.06% |      | 5012%       |
| Class H                             | -4.49% | -3.46% | -3.65% |      | 5100%       |

| Network Information in Inference Stage |                                                                              |                  |                 |
|----------------------------------------|------------------------------------------------------------------------------|------------------|-----------------|
|                                        |                                                                              | EE1-1.4.1        | EE1-1.4.2       |
| Mandatory                              | HW environment:                                                              |                  |                 |
|                                        | GPU Type                                                                     | CPU only         |                 |
|                                        | Framework:                                                                   | Tensorflow 2.2.0 |                 |
|                                        | Number of GPUs per Task                                                      | 0                |                 |
|                                        |                                                                              |                  |                 |
|                                        | Number of Parameters (Each Model)                                            | 107,670          | 990,796/989,498 |
|                                        | Total Parameter Number                                                       | 430,680          | 7,921,176       |
|                                        | Parameter Precision (Bits)                                                   | 32               |                 |
|                                        | Memory Parameter (MB)                                                        | 1.64             | 30.22           |
|                                        | MAC/Pixel (Giga)                                                             | 3.36e-5          | 6.24e-4         |
| Optional                               |                                                                              |                  |                 |
|                                        | Total Conv. Layers                                                           | 35               | 38              |
|                                        | Total FC Layers                                                              | 0                |                 |
|                                        | Total Memory (MB)                                                            |                  |                 |
|                                        | Batch size:                                                                  |                  |                 |
|                                        | Patch size                                                                   |                  |                 |
|                                        | Changes to network configuration or weights required to generate rate points |                  |                 |
|                                        | Peak Memory Usage (Total)                                                    | 1.43MB           | 8.54MB          |
|                                        | Peak Memory Usage (per Model)                                                | 1.43MB           | 4.27MB          |
|                                        | Other information:                                                           |                  |                 |

Note: DecT(Ours) is the DecT numbers for JVET-X0140 EE-1.4.1 as measured by us

# RA Test Results

Table 1 Simulation results of CP decomposition + fusion of baseline architecture

| CPD + fusion<br>(baseline) | Random access Main10       |        |        |      |       |
|----------------------------|----------------------------|--------|--------|------|-------|
|                            | BD-rate Over VTM-11_nnv2.0 |        |        |      |       |
|                            | Y-PSNR                     | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                   | -4.88%                     | -3.60% | -4.36% |      | 3567% |
| Class A2                   | -4.57%                     | -4.95% | -3.70% |      | 3397% |
| Class B                    | -4.07%                     | -6.08% | -5.93% |      | 3770% |
| Class C                    | -4.52%                     | -7.27% | -6.01% |      | 3538% |
| Class E                    |                            |        |        |      |       |
| <b>Overall</b>             | -4.45%                     | -5.68% | -5.19% |      | 3590% |
| Class D                    | -5.90%                     | -6.23% | -6.70% |      | 3837% |

Table 2 Simulation results of Split Luma-Chroma Architecture (24,8)

| Split Luma-<br>Chroma (24,8) | Random access Main10       |         |         |      |       |
|------------------------------|----------------------------|---------|---------|------|-------|
|                              | BD-rate Over VTM-11_nnv2.0 |         |         |      |       |
|                              | Y-PSNR                     | U-PSNR  | V-PSNR  | EncT | DecT  |
| Class A1                     | -4.63%                     | -5.84%  | -6.30%  |      | 4736% |
| Class A2                     | -4.88%                     | -9.21%  | -5.34%  |      | 4518% |
| Class B                      | -4.33%                     | -8.91%  | -8.88%  |      | 4993% |
| Class C                      | -4.93%                     | -10.51% | -9.45%  |      | 4655% |
| Class E                      |                            |         |         |      |       |
| <b>Overall</b>               | -4.66%                     | -8.78%  | -7.81%  |      | 4753% |
| Class D                      | -6.37%                     | -9.99%  | -10.66% |      | 5097% |

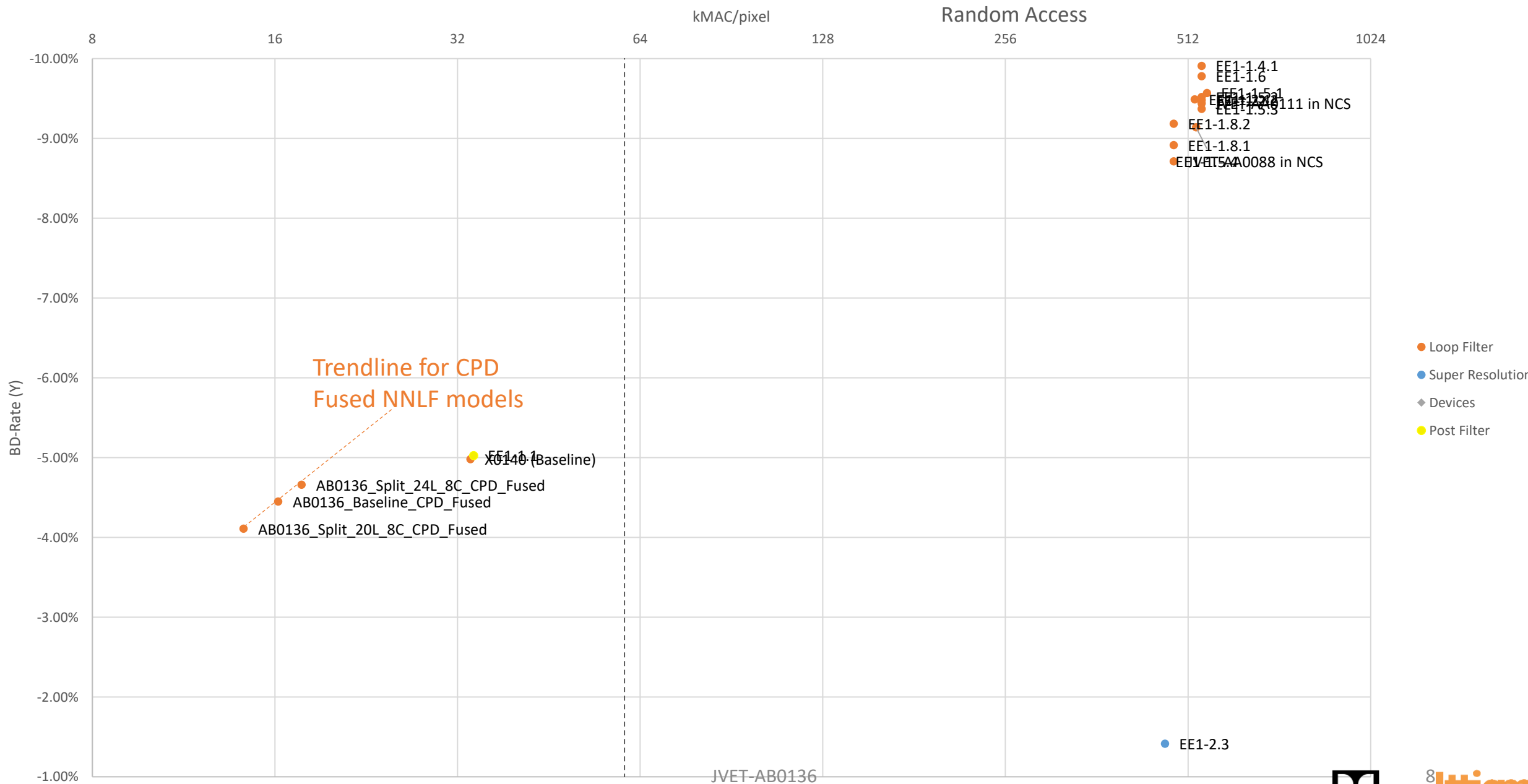
Table 3 Simulation results of Split Luma-Chroma Architecture (20,8)

| Split Luma-<br>Chroma (20,8) | Random access Main10       |         |        |      |       |
|------------------------------|----------------------------|---------|--------|------|-------|
|                              | BD-rate Over VTM-11_nnv2.0 |         |        |      |       |
|                              | Y-PSNR                     | U-PSNR  | V-PSNR | EncT | DecT  |
| Class A1                     | -4.38%                     | -5.24%  | -4.03% |      | 4756% |
| Class A2                     | -4.34%                     | -8.03%  | -3.97% |      | 4519% |
| Class B                      | -3.75%                     | -8.65%  | -6.43% |      | 5034% |
| Class C                      | -4.18%                     | -9.78%  | -7.03% |      | 4709% |
| Class E                      |                            |         |        |      |       |
| <b>Overall</b>               | -4.11%                     | -8.15%  | -5.62% |      | 4785% |
| Class D                      | -5.52%                     | -10.26% | -7.94% |      | 5121% |

Table 4 Summary of block level complexity (KMAC/Pixel) and # of parameters per model

| Architecture                          | KMAC/Pixel | Number of Parameters<br>(each model) |
|---------------------------------------|------------|--------------------------------------|
| JVET-X0140-1                          | 33.6       | 107,670                              |
| Baseline CPD + Fusion                 | 16.2       | 52,566                               |
| Luma Chroma Split (24,8) CPD + Fusion | 17.7       | 57,014                               |
| Luma Chroma Split (20,8) CPD + Fusion | 14.2       | 46,066                               |
| Luma Chroma Split (16,8) CPD + Fusion | 11.3       | 36,974                               |

# Loop filter complexity vs gain trade-off (RA)



# Summary

## **Proposed new methods to reduce NNLF complexity**

- CP-decomposition
- Fuse adjacent 1x1 convolutions
- New luma-chroma split NNLF architecture

**The complexity vs coding gains trade-offs are promising for AI and RA**

**It is proposed to further study the reduced complexity NNLF models in EE1**

# BACKUP SLIDES

# Network training and inference information (RA)

| Network Information in Inference Stage |                                   |                                     |       |                    |                    |
|----------------------------------------|-----------------------------------|-------------------------------------|-------|--------------------|--------------------|
|                                        |                                   | Baseline CP Fused                   |       | Split LC (20L, 8C) | Split LC (24L, 8C) |
| Mandatory                              | HW environment:                   |                                     |       |                    |                    |
|                                        | GPU Type                          | CPU only                            |       |                    |                    |
|                                        | Framework:                        | Tensorflow 2.8.0                    |       |                    |                    |
|                                        | Number of GPUs per Task           | 0                                   |       |                    |                    |
|                                        |                                   |                                     |       |                    |                    |
|                                        | Number of Parameters (Each Model) | 52566                               | 46066 | 57014              |                    |
|                                        | Total Parameter Number            | 105132                              | 92132 | 114028             |                    |
|                                        | Parameter Precision (Bits)        | Floating (32bits)                   |       |                    |                    |
|                                        | Memory Parameter (MB)             | 0.411                               | 0.368 | 0.435              |                    |
|                                        | MAC/Pixel (Kilo)                  | 16.2                                | 14.2  | 17.7               |                    |
| Optional                               |                                   |                                     |       |                    |                    |
|                                        | Total Conv. Layers                | 46                                  | 88    | 88                 |                    |
|                                        | Total FC Layers                   | 0                                   |       |                    |                    |
|                                        | Total Memory (MB)                 |                                     |       |                    |                    |
|                                        | Batch size:                       |                                     |       |                    |                    |
|                                        | Patch size                        | 72x72 (4 luma and 2 chroma planes ) |       |                    |                    |

| Network Information in Training Stage |                                                                                            |                                                                                                                                                                                                                                               |
|---------------------------------------|--------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mandatory                             | GPU Type                                                                                   | Nvidia GeForce RTX3090                                                                                                                                                                                                                        |
|                                       | Framework:                                                                                 | Tensorflow 2.8.0                                                                                                                                                                                                                              |
|                                       | Number of GPUs per Task                                                                    | 1                                                                                                                                                                                                                                             |
|                                       |                                                                                            |                                                                                                                                                                                                                                               |
|                                       | Epoch:                                                                                     | (30-50 per each RA stage)<br><ul style="list-style-type: none"> <li>Baseline CP Fused - Stage1:50 Stage2:50</li> <li>Split LC CP Fused (24,8) - Stage1:50 Stage2:30</li> <li>Split LC CP Fused (20,8) - Stage1:30 (Stage2 pending)</li> </ul> |
|                                       | Batch size:                                                                                | 16                                                                                                                                                                                                                                            |
|                                       | Training time:                                                                             | 200hours per each RA stage                                                                                                                                                                                                                    |
|                                       | Training data information:                                                                 | BVI-DVC                                                                                                                                                                                                                                       |
|                                       | Training configurations for generating compressed training data (is different to VTM CTC): | QP=20,25,30,35,40,45                                                                                                                                                                                                                          |
|                                       |                                                                                            |                                                                                                                                                                                                                                               |
| Optional                              | Number of iterations                                                                       |                                                                                                                                                                                                                                               |
|                                       | Patch size                                                                                 | 72x72 (4 luma and 2 chroma planes)                                                                                                                                                                                                            |
|                                       | Learning rate:                                                                             | <ul style="list-style-type: none"> <li>Epoch 0-20 :1E-4</li> <li>Epoch 20-30 : 1E-5</li> <li>Epoch &gt;30 : 1E-6</li> </ul>                                                                                                                   |
|                                       | Optimizer:                                                                                 | ADAM                                                                                                                                                                                                                                          |
|                                       | Loss function:                                                                             | 12-1-1 Weighted L2                                                                                                                                                                                                                            |
|                                       | Preprocessing:                                                                             | Convert 144x144 YUV420 signal to 6 72x72 blocks. Normalize to 0~1                                                                                                                                                                             |
|                                       | Other information:                                                                         |                                                                                                                                                                                                                                               |
|                                       |                                                                                            |                                                                                                                                                                                                                                               |

# AI Test Results (JVET-AA0080)

Table 1 Simulation results of CP decomposition + fusion of baseline architecture

| CPD + fusion<br>(baseline) | All Intra Main10            |        |        |      |       |
|----------------------------|-----------------------------|--------|--------|------|-------|
|                            | BD-rate Over VTM-11.0+V0056 |        |        |      |       |
|                            | Y-PSNR                      | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                   | -4.56%                      | -4.48% | -4.23% |      | 2343% |
| Class A2                   | -4.24%                      | -5.95% | -4.69% |      | 1858% |
| Class B                    | -4.23%                      | -5.90% | -4.97% |      | 1924% |
| Class C                    | -4.52%                      | -6.76% | -5.07% |      | 1498% |
| Class E                    | -6.21%                      | -5.02% | -4.88% |      | 2288% |
| <b>Overall</b>             | -4.68%                      | -5.72% | -4.81% |      | 1925% |
| Class D                    | -4.94%                      | -5.07% | -4.86% |      | 1550% |

Table 2 Simulation results of Split Luma-Chroma Architecture (24,8)

| Split Luma-<br>Chroma (24,8) | All Intra Main10            |        |        |      |       |
|------------------------------|-----------------------------|--------|--------|------|-------|
|                              | BD-rate Over VTM-11.0+V0056 |        |        |      |       |
|                              | Y-PSNR                      | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                     | -4.51%                      | -7.07% | -6.72% |      | 3133% |
| Class A2                     | -4.58%                      | -9.10% | -6.39% |      | 2463% |
| Class B                      | -4.50%                      | -7.85% | -8.00% |      | 2558% |
| Class C                      | -5.05%                      | -8.52% | -8.68% |      | 2377% |
| Class E                      | -6.59%                      | -6.92% | -7.53% |      | 3558% |
| <b>Overall</b>               | -4.99%                      | -7.92% | -7.59% |      | 2733% |
| Class D                      | -5.45%                      | -7.50% | -9.52% |      | 2643% |

Table 3 Simulation results of Split Luma-Chroma Architecture (20,8)

| Split Luma-<br>Chroma (20,8) | All Intra Main10            |        |        |      |       |
|------------------------------|-----------------------------|--------|--------|------|-------|
|                              | BD-rate Over VTM-11.0+V0056 |        |        |      |       |
|                              | Y-PSNR                      | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                     | -4.37%                      | -7.15% | -6.18% |      | 2867% |
| Class A2                     | -4.24%                      | -8.29% | -6.13% |      | 2259% |
| Class B                      | -4.15%                      | -7.98% | -7.62% |      | 2345% |
| Class C                      | -4.45%                      | -8.37% | -7.53% |      | 1846% |
| Class E                      | -6.10%                      | -7.71% | -8.03% |      | 2808% |
| <b>Overall</b>               | -4.59%                      | -7.93% | -7.18% |      | 2355% |
| Class D                      | -4.89%                      | -7.35% | -8.43% |      | 1950% |

Table 4 Simulation results of Split Luma-Chroma Architecture (16,8)

| Split Luma-<br>Chroma (16,8) | All Intra Main10            |        |        |      |       |
|------------------------------|-----------------------------|--------|--------|------|-------|
|                              | BD-rate Over VTM-11.0+V0056 |        |        |      |       |
|                              | Y-PSNR                      | U-PSNR | V-PSNR | EncT | DecT  |
| Class A1                     | -4.09%                      | -7.22% | -6.90% |      | 2514% |
| Class A2                     | -3.99%                      | -8.46% | -6.50% |      | 1995% |
| Class B                      | -3.92%                      | -8.48% | -8.56% |      | 2078% |
| Class C                      | -4.23%                      | -8.97% | -8.23% |      | 1621% |
| Class E                      | -5.69%                      | -7.88% | -7.33% |      | 2463% |
| <b>Overall</b>               | -4.32%                      | -8.28% | -7.66% |      | 2074% |
| Class D                      | -4.67%                      | -8.29% | -9.37% |      | 1712% |

Table 5 Summary of block level complexity (KMAC/Pixel) and # of parameters (each model)

| Architecture                          | KMAC/Pixel | Number of Parameters<br>(each model) |
|---------------------------------------|------------|--------------------------------------|
| JVET-X0140-1                          | 33.6       | 107,670                              |
| Baseline CPD + Fusion                 | 16.2       | 52,566                               |
| Luma Chroma Split (24,8) CPD + Fusion | 17.7       | 57,014                               |
| Luma Chroma Split (20,8) CPD + Fusion | 14.2       | 46,066                               |
| Luma Chroma Split (16,8) CPD + Fusion | 11.3       | 36,974                               |

# Loop filter complexity vs gain trade-off (AI)



# Network training and inference information CP-fused

| Network Information in Inference Stage |                                                                              |                                    |                            |
|----------------------------------------|------------------------------------------------------------------------------|------------------------------------|----------------------------|
|                                        |                                                                              | CP-decomposed                      | CP-decomposed + 1x1 fusion |
| Mandatory                              | HW environment:                                                              |                                    |                            |
|                                        | GPU Type                                                                     | CPU only                           |                            |
|                                        | Framework:                                                                   | Tensorflow 2.8.0                   |                            |
|                                        | Number of GPUs per Task                                                      | 0                                  |                            |
|                                        |                                                                              |                                    |                            |
|                                        | Number of Parameters (Each Model)                                            | 64902                              | 52566                      |
|                                        | Total Parameter Number                                                       | 129804                             | 105132                     |
|                                        | Parameter Precision (Bits)                                                   | Floating (32bits)                  |                            |
|                                        | Memory Parameter (MB)                                                        | 0.507                              | 0.411                      |
|                                        | MAC/Pixel (Kilo)                                                             | 20.1                               | 16.2                       |
| Optional                               |                                                                              |                                    |                            |
|                                        | Total Conv. Layers                                                           | 68                                 | 46                         |
|                                        | Total FC Layers                                                              | 0                                  |                            |
|                                        | Total Memory (MB)                                                            |                                    |                            |
|                                        | Batch size:                                                                  |                                    |                            |
|                                        | Patch size                                                                   | 72x72 (4 luma and 2 chroma planes) |                            |
|                                        | Changes to network configuration or weights required to generate rate points |                                    |                            |
|                                        | Peak Memory Usage (Total)                                                    |                                    |                            |
|                                        | Peak Memory Usage (per Model)                                                |                                    |                            |
|                                        | Other information:                                                           |                                    |                            |

| Network Information in Training Stage |                                                                                            |                                                                   |
|---------------------------------------|--------------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| Mandatory                             | GPU Type                                                                                   | Nvidia GeForce RTX3090                                            |
|                                       | Framework:                                                                                 | Tensorflow 2.8.0                                                  |
|                                       | Number of GPUs per Task                                                                    | 1                                                                 |
|                                       |                                                                                            |                                                                   |
|                                       | Epoch:                                                                                     | ~374                                                              |
|                                       | Batch size:                                                                                | 16                                                                |
|                                       | Training time:                                                                             | 124h                                                              |
|                                       | Training data information:                                                                 | DIV2K                                                             |
|                                       | Training configurations for generating compressed training data (is different to VTM CTC): | QP=20,25,30,35,40,45                                              |
|                                       |                                                                                            |                                                                   |
| Optional                              | Number of iterations                                                                       |                                                                   |
|                                       | Patch size                                                                                 | 72x72 (4 luma and 2 chroma planes)                                |
|                                       | Learning rate:                                                                             | Epoch : LR, 0-56: 2E-5, 56-374: 2E-6                              |
|                                       | Optimizer:                                                                                 | ADAM                                                              |
|                                       | Loss function:                                                                             | 12-1-1 Weighted L2                                                |
|                                       | Preprocessing:                                                                             | Convert 144x144 YUV420 signal to 6 72x72 blocks. Normalize to 0~1 |
|                                       | Other information:                                                                         |                                                                   |
|                                       |                                                                                            |                                                                   |
|                                       |                                                                                            |                                                                   |
|                                       |                                                                                            |                                                                   |

# Network inference information luma-chroma split + CP + fused

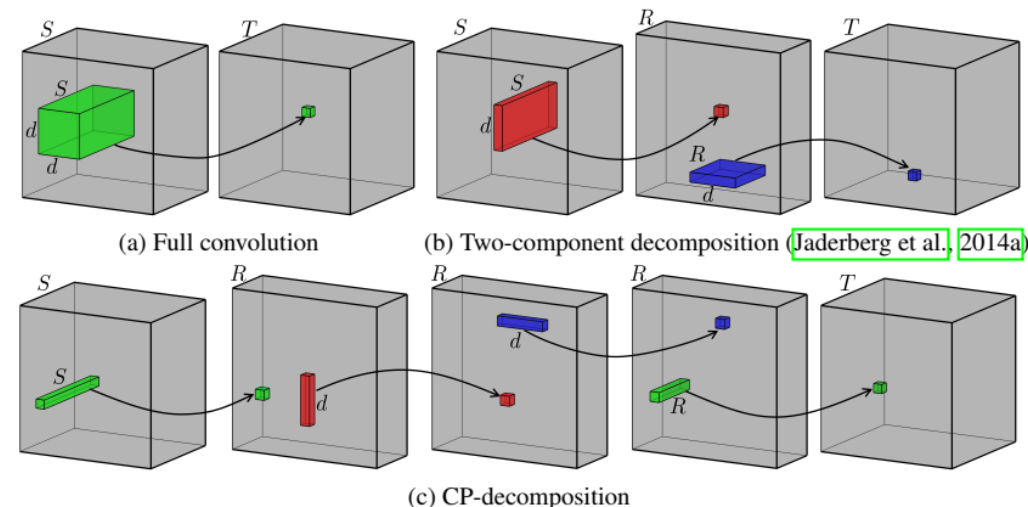
| Network Information in Inference Stage |                                                                              |                                     |                    |                    |
|----------------------------------------|------------------------------------------------------------------------------|-------------------------------------|--------------------|--------------------|
|                                        |                                                                              | Split LC (16L, 8C)                  | Split LC (20L, 8C) | Split LC (24L, 8C) |
| Mandatory                              | HW environment:                                                              |                                     |                    |                    |
|                                        | GPU Type                                                                     | CPU only                            |                    |                    |
|                                        | Framework:                                                                   | Tensorflow 2.8.0                    |                    |                    |
|                                        | Number of GPUs per Task                                                      | 0                                   |                    |                    |
|                                        |                                                                              |                                     |                    |                    |
|                                        | Number of Parameters (Each Model)                                            | 36974                               | 46066              | 57014              |
|                                        | Total Parameter Number                                                       | 73948                               | 92132              | 114028             |
|                                        | Parameter Precision (Bits)                                                   | Floating (32bits)                   |                    |                    |
|                                        | Memory Parameter (MB)                                                        | 0.282                               | 0.368              | 0.435              |
|                                        | MAC/Pixel (Kilo)                                                             | 11.3                                | 14.2               | 17.7               |
| Optional                               |                                                                              |                                     |                    |                    |
|                                        | Total Conv. Layers                                                           | 88                                  |                    |                    |
|                                        | Total FC Layers                                                              | 0                                   |                    |                    |
|                                        | Total Memory (MB)                                                            |                                     |                    |                    |
|                                        | Batch size:                                                                  |                                     |                    |                    |
|                                        | Patch size                                                                   | 72x72 (4 luma and 2 chroma planes ) |                    |                    |
|                                        | Changes to network configuration or weights required to generate rate points |                                     |                    |                    |
|                                        | Peak Memory Usage (Total)                                                    |                                     |                    |                    |
|                                        | Peak Memory Usage (per Model)                                                |                                     |                    |                    |
|                                        | Other information:                                                           |                                     |                    |                    |

# CP-Decomposition

## • Regular Convolution

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

- $U(:, :, :)$  input tensor of size  $X \times Y \times S$
- $K(:, :, :, :)$  4-D Kernel of  $d \times d \times S \times T$
- $V(:, :, :)$  output tensor
- $S, T$  Number of input and output feature maps
- Complexity MAC Per pixel  **$STd^2$**



## • CP Decomposition

- Rank  $R$  approximation of  $K$
- $K^x, K^y, K^s, K^t$  are  $\delta \times R, \delta \times R, S \times R, T \times R$  2D Tensors
- Complexity MAC/pixel  $R(S+2d+T)$  as compared to  **$STd^2$**  regular convolution
- $R \sim \frac{ST}{S+T}$ , then improvement factor is  $d^2$

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

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