|  |
| --- |
| **INTERNATIONAL ORGANIZATION FOR STANDARDIZATION ORGANISATION INTERNATIONALE DE NORMALISATION ISO/IEC JTC 1/SC 29/WG 5 MPEG JOINT VIDEO EXPERTS TEAM WITH ITU-T SG 16** |
| **ISO/IEC JTC 1 / SC 29 / WG 5 N 264** |
| **Online – 17–26 January 2024** |
| |  |  | | --- | --- | | **Source:** | **Convenor (Jens-Rainer Ohm)** | | **Title:** | **Exploration experiment on neural network-based video coding (EE1)** | | **Type:** | **General** | | **Subtype:** | **Other** | | **Status:** | **Approved** | | **Date:** | **2024-02-26** | | **Expected Action:** | **Info** | | **Action due date:** | **N/A** | | **Pages:** | **13** (not including this cover page) | | **Email of convenor:** | **ohm @ ient . rwth-aachen . de** | | **Committee URL:** | **https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-5** | |

|  |  |
| --- | --- |
| **Joint Video Experts Team (JVET)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29**  33rd Meeting, by teleconference, 17–26 January 2024 | Document: JVET-AG2023-v2 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Exploration Experiments on Neural Network-based Video Coding (EE1)** | | |
| *Status:* | Output document to JVET | | |
| *Purpose:* | Report | | |
| *Author(s) or Contact(s):* | E. Alshina, R. Chang, F. Galpin,  Y. Li, D. Rusanovskyy, M. Santamaria, J. [Ström](mailto:jacob.strom@ericsson.com), and Z. Xie | Email: | [elena.alshina@huawei.com](about:blank)  [renjiechang@tencent.com](mailto:renjiechang@tencent.com)  [franck.galpin@interdigital.com](mailto:franck.galpin@interdigital.com)  [yue.li@bytedance.com](mailto:yue.li@bytedance.com)  dmytror@qti.qualcomm.com  [maria.santamaria\_gomez@nokia.com](mailto:maria.santamaria_gomez@nokia.com)  [jacob.strom@ericsson.com](mailto:jacob.strom@ericsson.com)  [xiezhihuang@oppo.com](mailto:xiezhihuang@oppo.com) |
| *Source:* | EE Coordinators | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This document summarizes Exploration Experiment 1 (EE1) tests to be performed between the JVET-AG and JVET-AH meetings to evaluate **Neural Network-based Video Coding (**NNVC) technologies, analyze their performances and complexity aspects.

# Introduction

Code base for the test should be NNVC-8.0, anchor is default configuration of NNVC-8.0 (NN-Intra and LOP-2 filter enabled). NNVC common test conditions, results and complexity reporting template must be used.

For tests competing with technologies in NNVC-8 it is recommended to configure the proposed solution to have a complexity close to related NNVC(LOP/HOP), but not exceeding:

* kMAC/pxl of EE1 test ≤ kMAC/pxl NNVC (*must*),
* Number of Parameters EE1 test ≤ Number of Parameters NNVC (*if possible*).

NN architecture provided in this test description should not be changed, beside minor adjustment for parameters (such as channels number) in order to meet recommendation above. Exact parameters settings can be specified by 2nd AhG11/14 teleconference Feb. 28.

# List of tests

This round of EE1 tests will include:

* EE1-1: LOP in-loop filter
  + EE1-1.0: LOP re-training training strategy [JVET-AG0156](https://jvet-experts.org/doc_end_user/current_document.php?id=13712)
  + EE1-1.1: training of EE1-0 with architecture change [JVET-AG0163](https://jvet-experts.org/doc_end_user/current_document.php?id=13719) (merge convolutions) & [JVET-AG0155](https://jvet-experts.org/doc_end_user/current_document.php?id=13711) (separable convolutions).
  + EE1-1.2: training of EE1-0 for [JVET-AG0163](https://jvet-experts.org/doc_end_user/current_document.php?id=13719) (merge convolutions) & [JVET-AG0155](https://jvet-experts.org/doc_end_user/current_document.php?id=13711) (separable convolutions) & [JVET-AG0069](https://jvet-experts.org/doc_end_user/current_document.php?id=13625) (transformed input).
  + EE1-1.3: training from the scratch for [JVET-AG0163](https://jvet-experts.org/doc_end_user/current_document.php?id=13719) (merge convolutions) & [JVET-AG0155](https://jvet-experts.org/doc_end_user/current_document.php?id=13711) (separable convolutions) & [JVET-AG0069](https://jvet-experts.org/doc_end_user/current_document.php?id=13625) (transformed input).
  + EE1-1.4: LOP2 architecture (training of EE1-0) with content adaptivity [JVET-AG0111](https://jvet-experts.org/doc_end_user/current_document.php?id=13667)
* EE1-2: HOP in-loop filter
  + EE1-2.1: Separate HOP model [JVET-AG0175](https://jvet-experts.org/doc_end_user/current_document.php?id=13731)
  + EE1-2.2: Wide activation HOP model [JVET-AG0179](https://jvet-experts.org/doc_end_user/current_document.php?id=13735)
  + EE1-2.3: HOP with transformers [JVET-AG0162](https://jvet-experts.org/doc_end_user/current_document.php?id=13718)

*Note*: training should be aligned with NNVC-8.0 HOP3.

* EE1-3: NN-inter prediction
  + EE1-3.1: Deep Reference Frame Generation for Inter Prediction Enhancement [JVET-AG0122](https://jvet-experts.org/doc_end_user/current_document.php?id=13678)
* EE1-4: NN-super-resolution
  + EE1-4.0: adaptive resolution (s=2) coding for NNVC-8 (NNIntra and LOP2 filter) with NNVC SR filter
  + EE1-4.1: adaptive resolution coding for NNVC-8 (NNIntra and LOP2 filter) with RPR
    - EE1-4.1.1: scaling factor s=2
    - EE1-4.1.2: scaling factor s=1.5
  + EE1-4.2: adaptive resolution coding for NNVC-8 (NNIntra and LOP2 filter) with [[JVET-AG0129](https://jvet-experts.org/doc_end_user/current_document.php?id=13685)](https://jvet-experts.org/doc_end_user/current_document.php?id=13686)
    - EE1-4.2.1: scaling factor s=2
    - EE1-4.2.2: scaling factor s=1.5
  + EE1-4.3: adaptive resolution (s= 2) coding for NNVC-8 (NNIntra and LOP2 filter) with [JVET-[AG0114](https://jvet-experts.org/doc_end_user/current_document.php?id=13670)](https://jvet-experts.org/doc_end_user/current_document.php?id=13686) (LOP version)

*Note*: strategy for scaling factor, QP selection to be same for all tests, choose best strategy for RPR (in EE1-4.1), announced during 1st teleconference and shared in all tests of this category.

* EE1-5. Very Low Operational Point
  + EE1-5.1 Study of the NN architecture at Very Low Operational Point (JVET-AG0155, JVET-AG0057, JVET-AF0206)

# Tests description

## EE1-1: LOP in-loop filter

### EE1-1.0: LOP re-training training strategy [JVET-AG0156](https://jvet-experts.org/doc_end_user/current_document.php?id=13712)

In this test, original LOP2 architecture is used, see Figure 1. Participants will jointly train LOP2 Stage 3 with batch size 64 and learning rate 0.0008, as proposed in JVET-AG0156 and JVET-AG0163 (EE1-2.3). The goal of the test is to verify stability, replicability of the proposed training strategy. No changes to LOP2 architecture and training dataset are introduced. Changes to number of training epochs, L1/L2 and learning rate points are anticipated.

Training: Modified LOP Stage 3 training strategy.

Inference: the NNVC 8.0 software is used for inference, following LOP anchor settings.

Crosscheck: Dolby, Ittiam and Qualcomm will independently train/crosscheck each other for the training and inference part.

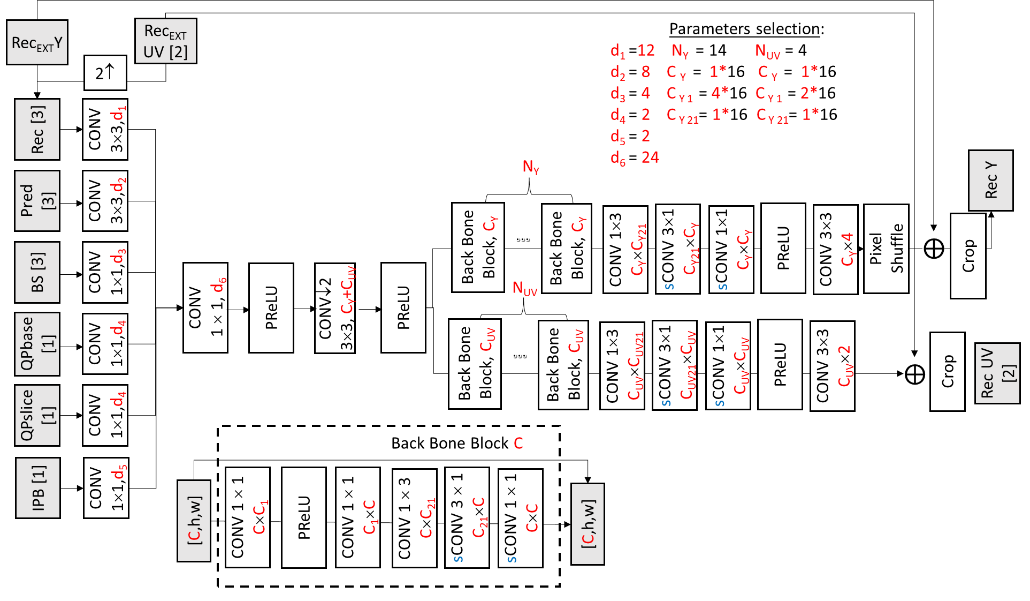


Figure 1.

### EE1-1.1: LOP candidate with Low Complexity TB (JVET-AG0155) and new BB structure (JVET-AG0163)

In this test, a combination of separable convolutions in fusion/transition block (JVET-AG0155) and modified BB block (JVET-AG0163) is jointly trained by proponents. A candidate architecture is shown in Figure 2, its parameters may be changed to meet target complexity of 17.05kMAC/pixel and balance luma/chroma performance.

LOP Stage 3 training will be conducted, following the strategy selected from EE1-1.0. Sub-test with modified training may be additionally presented, supplied with results of the same training applied to LOP2.

Training: LOP Stage 3 training strategy from EE1-1.0.

Inference: the NNVC 8.0 software is used for inference, following LOP anchor settings.

Crosscheck: Dolby, Ittiam and Qualcomm will independently train/crosscheck each other for the training and inference part.

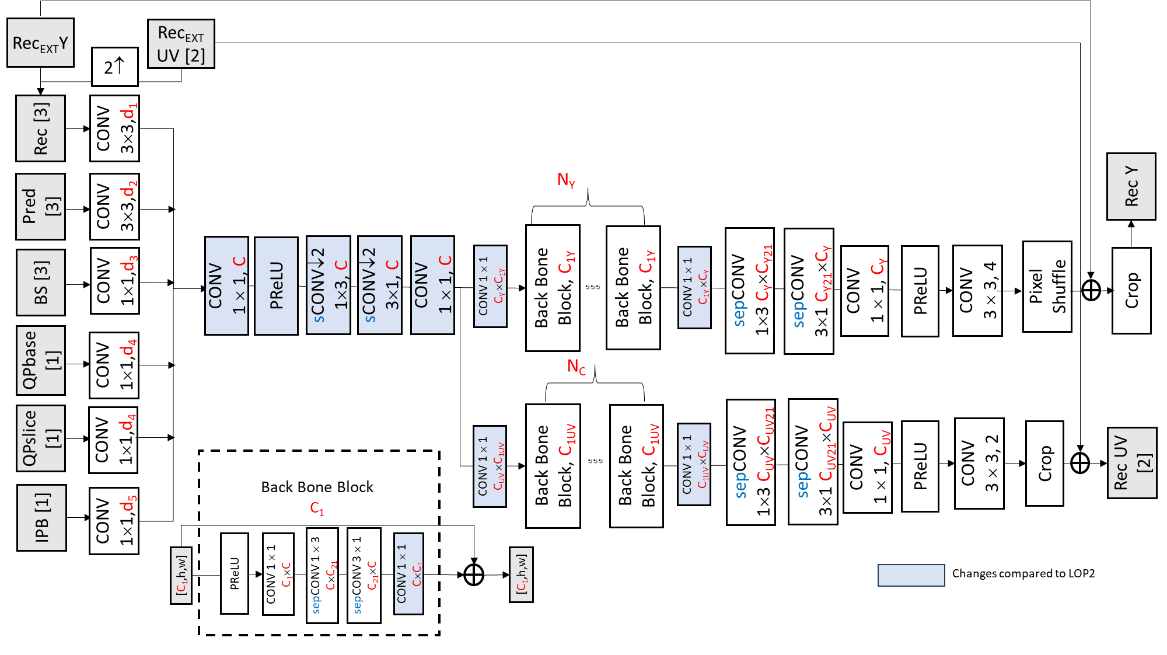


Figure 2.

### EE1-1.2: training of EE1-0 for [JVET-AG0163](https://jvet-experts.org/doc_end_user/current_document.php?id=13719) (merge convolutions) & [JVET-AG0155](https://jvet-experts.org/doc_end_user/current_document.php?id=13711) (separable convolutions) & [JVET-AG0069](https://jvet-experts.org/doc_end_user/current_document.php?id=13625) (transformed input)

In this test, the transformed inputs from JVET-AG0069 are applied to the combined model from JVET-AG0163 and JVET-AG0155. A candidate architecture in shown in Figure 3. The number of channels and backbones blocks may be modified to meet the target complexity of 17 kMAC/pixel and balance luma/chroma performance.

The model will be trained at stage 3 following training strategy from EE1-0. Subtests with modified parameters may be additionally presented.

Training: Training strategy from EE1-1.0 at stage 3.

Inference: The NNVC-8.0 software is used for inference, following LOP anchor settings.

Proponent: Ericsson will run training and inference.

Crosschecker: Tencent will run crosscheck for training and inference

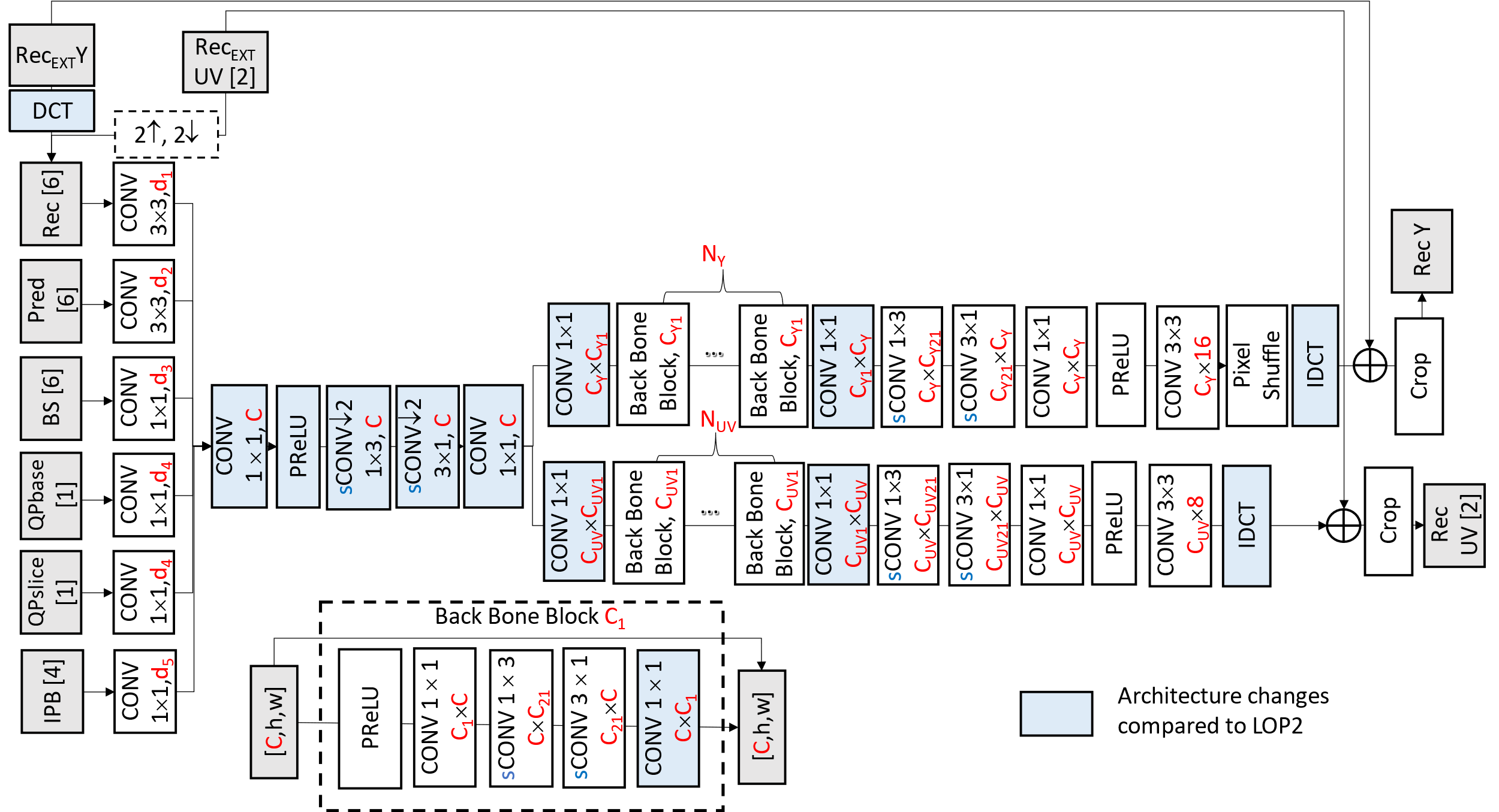


Figure 3.

### EE1-1.3: training from the scratch for [JVET-AG0163](https://jvet-experts.org/doc_end_user/current_document.php?id=13719) (merge convolutions) & [JVET-AG0155](https://jvet-experts.org/doc_end_user/current_document.php?id=13711) (separable convolutions) & [JVET-AG0069](https://jvet-experts.org/doc_end_user/current_document.php?id=13625) (transformed input)

In this test, the model from EE1-1.2 will be trained from scratch. A candidate architecture in shown in Figure 3. The model will be trained from scratch, and training strategy may be changed for the tested model. Changes in training dataset and luma/chroma weighting are anticipated.

Training: The model will be trained from scratch.

Inference: The NNVC-8.0 software is used for inference, following LOP anchor settings.

Proponent: Ericsson will run training and inference.

Crosschecker: Nokia will run crosscheck for training and inference EE1-1.4: LOP2 architecture (training of EE1-0) with content adaptivity [JVET-AG0111](https://jvet-experts.org/doc_end_user/current_document.php?id=13667)

This test will study the application of content adaptability to LOP2 architecture.

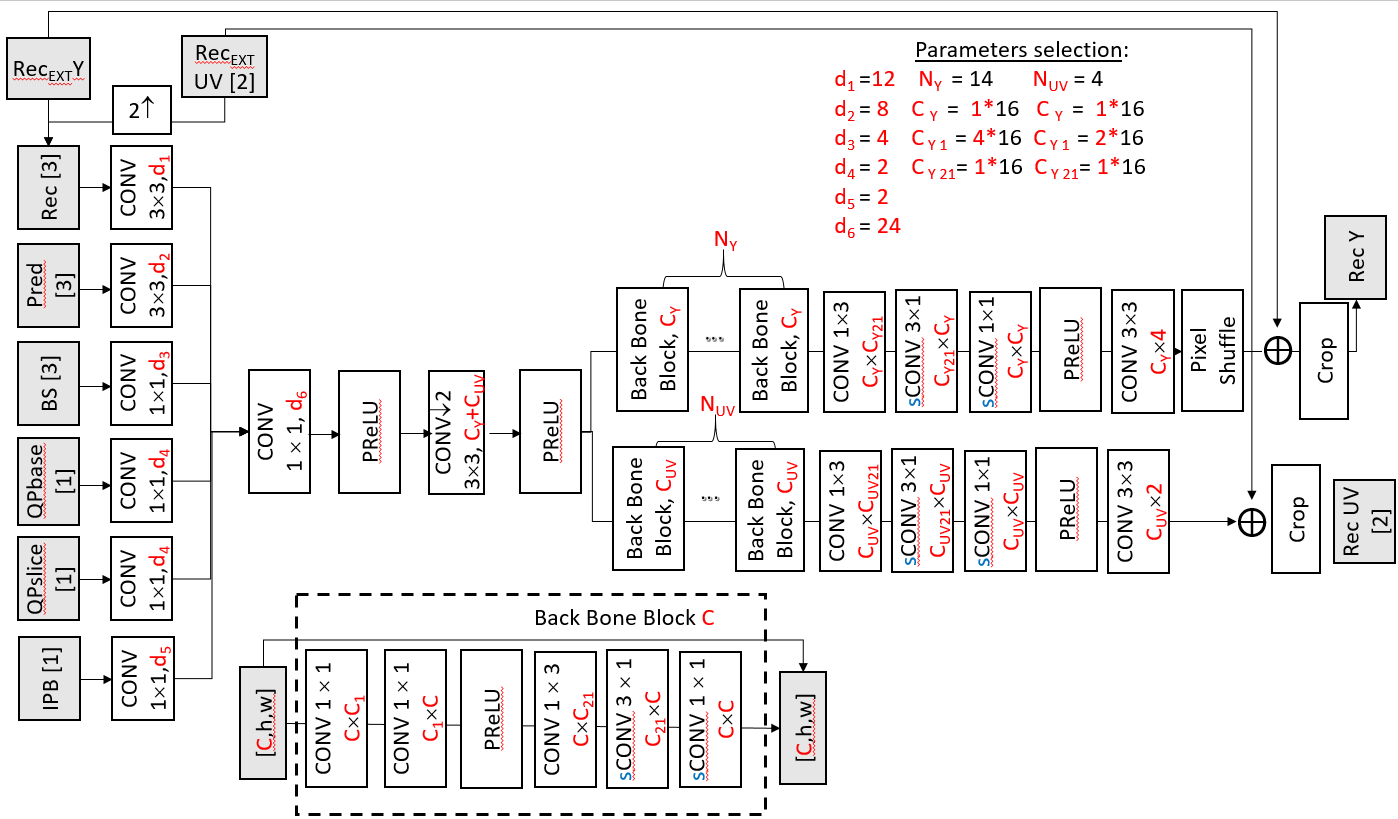


Figure 4.

Training: overfitting for each sequence (segment-wise) and encoding configuration

Inference: The NNVC-8.0 software is used as base (plus NNFU APS and other changes), following LOP anchor settings

Proponent: Nokia

Cross-checker: Ericsson will run crosscheck for training and inference.

## EE1-2: HOP in-loop filter

### EE1-2.1: Separate HOP model [JVET-AG0175](https://jvet-experts.org/doc_end_user/current_document.php?id=13731)

In this test, separate filter architectures for luma and chroma will be studied. Candidate architectures are shown in the figures below, their parameters may be changed to meet target complexity of HOP3.



Figure 4. Luma filter architecture in EE1-2.1



Figure 5. Chroma filter architecture in EE1-2.1

HOP Stage 3 training will be conducted. Sub-tests with different network hyper-parameters may be presented.

Training: HOP Stage 3 training strategy.

Inference: the NNVC 8.0 software is used for inference, following HOP anchor settings.

Cross-checker: Bytedance, InterDigital, Qualcomm, Tencent will independently train/crosscheck each other for the training and inference part.

Back Bone Block, C

CONV

3 × 3

PReLU

CONV

3 × 3, 4

Pixel Shuffle

Rec Y

Crop

CONV 1 × 1 (C1+C22)×C

gCONV 3 × 1 C×C31

CONV 1 × 3 C31×C

[C,h,w]

Parameters selection:

d1 =12\*16 C = 4\*16 N = TBD

d2 = 2\*16 C1 = TBD d6 = 3\*16

d3 = 1\*16 C21= 2\*16

d4 = 1\*16 C22= TBD

d5 = 1\*16 C31= 3\*16

### EE1-2.2: Wide activation HOP model [JVET-AG0179](https://jvet-experts.org/doc_end_user/current_document.php?id=13735)

In this test, a filter architecture for HOP is developed by rebalancing the network depth and width. In particular, the number of feature maps in the backbone block is increased while the number of backbone blocks is reduced. A candidate architecture is shown in the figure below, its parameters may be changed to meet target complexity of HOP3.



Figure 6. Filter architecture in EE1-2.2

HOP Stage 3 training will be conducted. Sub-tests with different network hyper-parameters may be presented.

Training: HOP Stage 3 training strategy.

Inference: the NNVC 8.0 software is used for inference, following HOP anchor settings.

Proponent: Bytedance

Cross-checker: xx

### EE1-2.3: HOP with transformers [JVET-AG0162](https://jvet-experts.org/doc_end_user/current_document.php?id=13718)

This test is based on JVET-AF0158, where transformer was introduced to the backbone block of HOP architecture and JVET-AG0162, where a low complexity attention mechanism, comprising a sub-set of transformer, is introduced. Two sub-tests are planned, respectively.

A diagram of a computer

Description automatically generated

Filter architecture: HOP3-based architecture, implementation of attention mechanism and transformer block.

Target complexity: HOP.

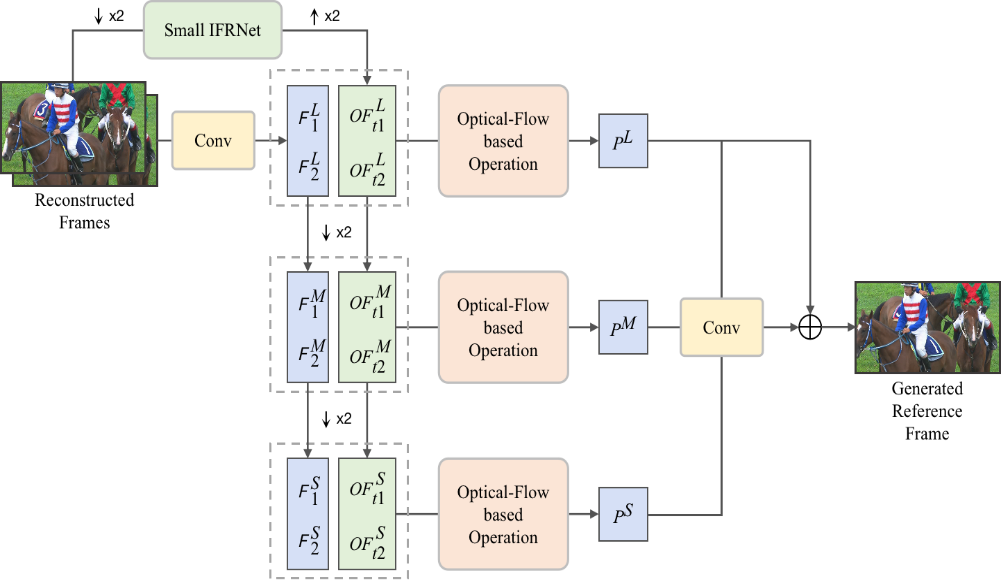
Training strategy: HOP-based.

Proponent: Qualcomm.

Cross-checker: Tencent (training and inference)

## EE1-3: NN-inter prediction

### EE1-3.1: Deep Reference Frame Generation for Inter Prediction Enhancement [JVET-AG0122](https://jvet-experts.org/doc_end_user/current_document.php?id=13678)



This test is based on JVET-AG0122, aiming at network optimization and training dataset extension. When Vimeo is utilized as the training dataset, this test will optimize the network architecture of NN-Inter for better performance. Additionally, this test will continue to extend the existing training dataset under NNVC standardization.

Training: Training strategy from JVET-AG0122.

Inference: The NNVC 8.0 software is used for inference under NNVC RA configuration.

Proponent: Wuhan University and Tencent.

Cross-check: Inference and training cross-checks are planned.

Cross-checker: OPPO.

## EE1-4: NN-super-resolution

### EE1-4.0: adaptive resolution (s=2) coding for NNVC-8 (NNIntra and LOP2 filter) with NNVC SR filter

Performance provided by AhG14 as reference point. The architecture is shown in Figure EE1-4.0.



Figure EE1-4.0

### EE1-4.1: adaptive resolution coding for NNVC-8 (NNIntra and LOP2 filter) with RPR

EE1-4.1.1 scaling factor s=2

EE1-4.1.2 scaling factor s=1.5

In this test, a study on encoder strategy for RPR with NN LF and NN-intra tools enabled is conducted. No training is expected because RPR filter is used.

Inference: the NNVC 8.0 software is used for inference, following LOP anchor settings.

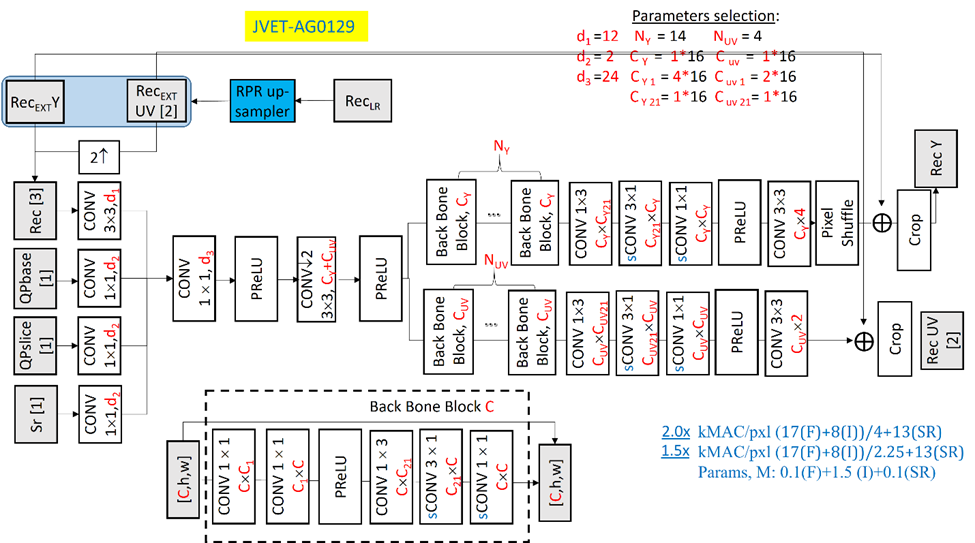
Provider: Performance will be provided by Bytedance (s=2) and vivo (s=1.5).

Crosschecker: TBA

### EE1-4.2: adaptive resolution coding for NNVC-8 (NNIntra and LOP2 filter) with [[JVET-AG0129](https://jvet-experts.org/doc_end_user/current_document.php?id=13685)](https://jvet-experts.org/doc_end_user/current_document.php?id=13686)

EE1-4.2.1 scaling factor s=2

EE1-4.2.2 scaling factor s=1.5

This test aims to investigate the use of a unified CNN-based super-resolution method to handle both scaling ratios of 2x and 1.5x. The network structure shown above will be studied with minor architectural changes, dataset generation, and training strategies.

Training: xx

Inference: The NNVC 8.0 software is used for inference.

Proponent: vivo.

Cross-check: Inference and training cross-checks are planned.

Cross-checker: TBA.

### EE1-4.3: adaptive resolution (s= 2) coding for NNVC-8 (NNIntra and LOP2 filter) with [[JVET-AG0114](https://jvet-experts.org/doc_end_user/current_document.php?id=13670)](https://jvet-experts.org/doc_end_user/current_document.php?id=13686) (LOP version)

HOP version network architecture.

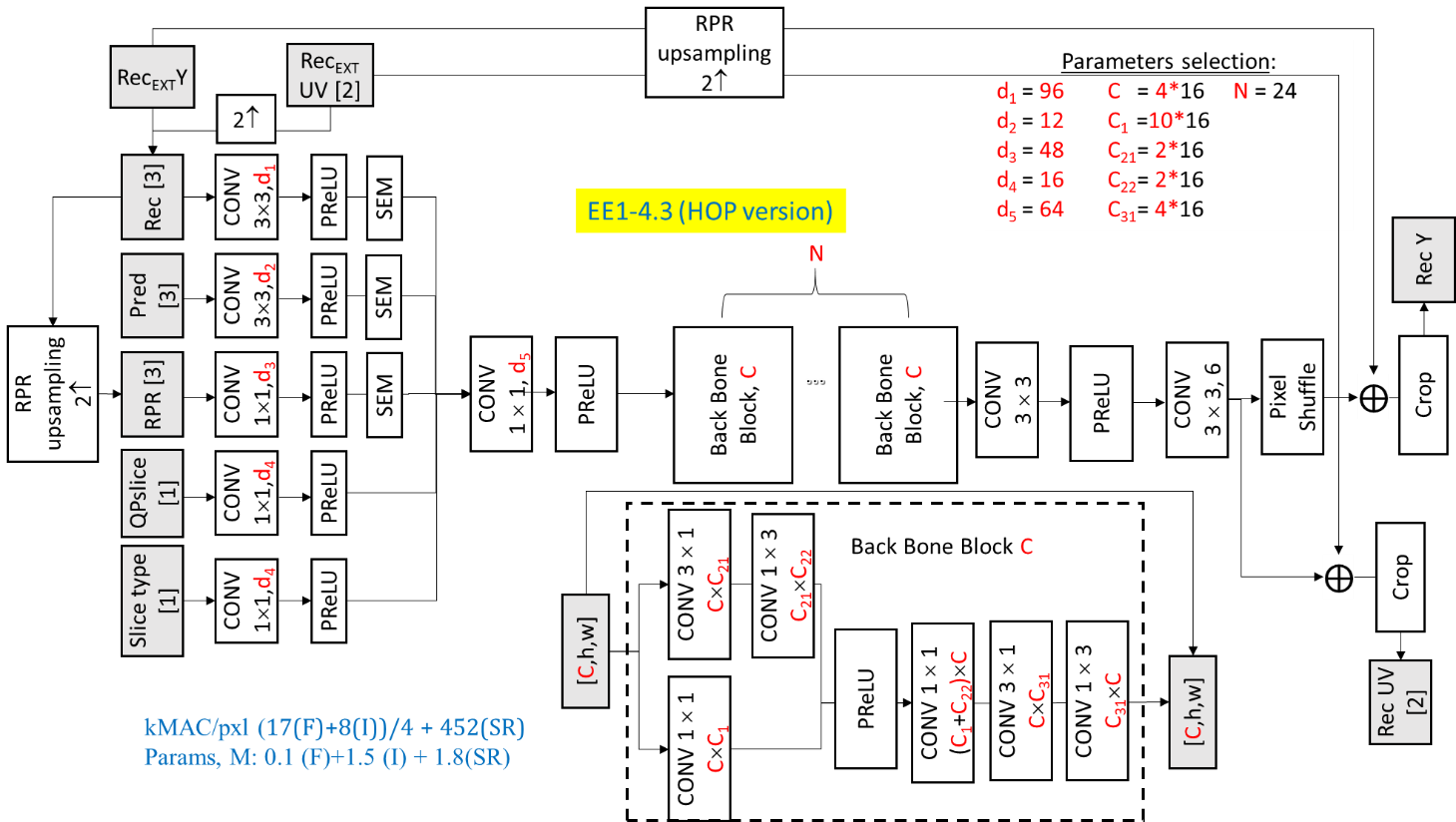


Figure EE1-4.3 HOP version

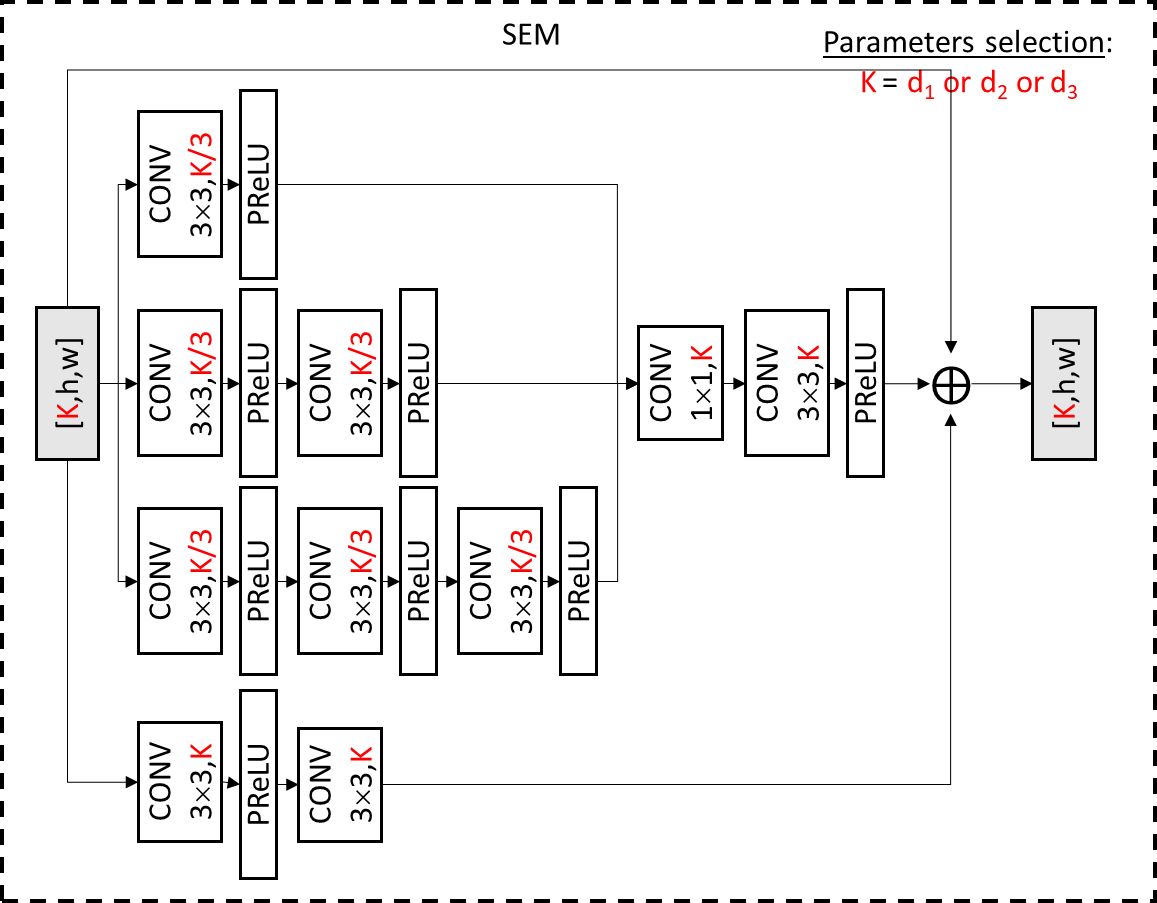


Figure shallow feature enhancement module (SEM)

This test aims to introduce RPR inputs and SEM preprocessing of shallow features into the SR network for better high resolution reconstruction. On the one hand, HOP version as proposed will be studied with minor architectural changes, dataset generation, and training strategies. On the other hand, Reduce the complexity of the HOP version and design the LOP version of proposed network structure.

Inference: The NNVC 8.0 software is used for inference.

Proponent: HUST, vivo.

Cross-checker: TBA.

## Very Low Operational Point

### EE1-5: Study of the NN architecture at Very Low Operational Point

In this test, a study on NN architecture for Very Low Operational Point, targeting complexity of 5kMAC/pixel is conducted. A combination of separable convolutions in fusion/transition block (JVET-AG0155) and filter configurations of JVET-AG0057 and JVET-AF0206 is jointly trained by proponents. A candidate architecture is shown in Figure EE1-5, its parameters may be changed to meet target complexity of 5kMAC/pixel and balance luma/chroma performance. Training is expected to follow LOP training strategy.

Training: Following LOP training strategy.

Inference: the NNVC 8.0 software is used for inference, following LOP anchor settings.

Crosscheck: Qualcomm, Bytedance and OPPO will independently train/crosscheck each other for the training and inference part.

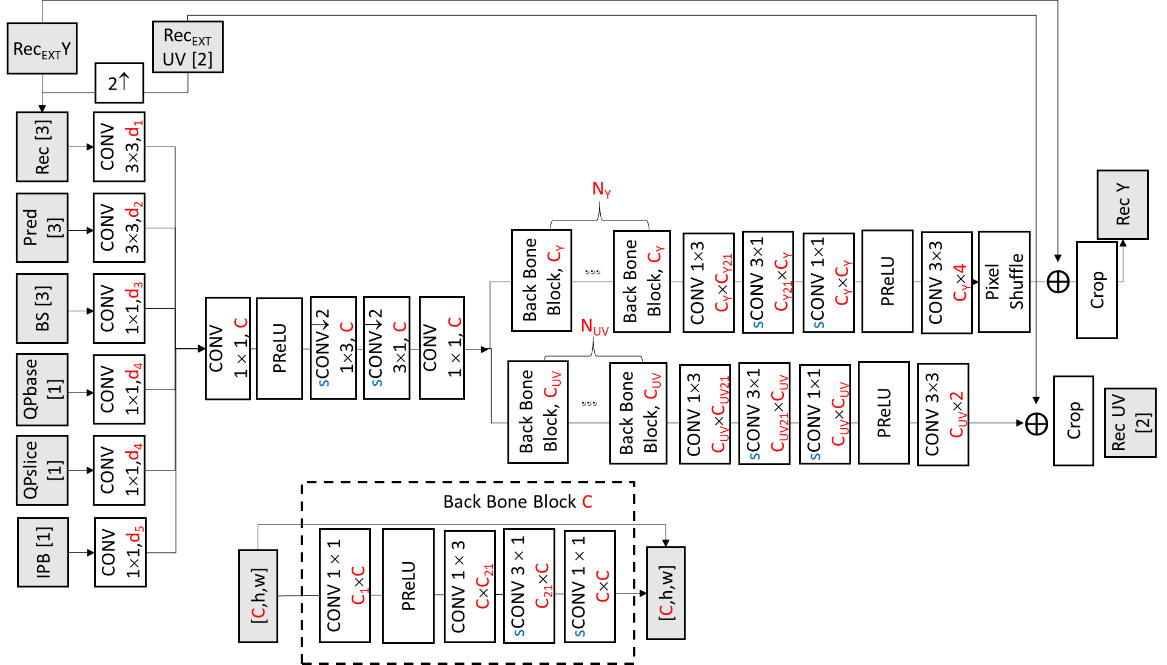


Figure EE1-5.

# Timeline

**T1 - 2 weeks after JVET-AG meeting (09-Feb-2024):** EE description (JVET-AG2023) finalized and uploaded.NNVC-8.0 software is available, including anchor performance.

**T2 – 2.5 weeks after JVET-AG meeting (13-Feb-2024) 1st teleconference.** EE1-1.0 LOP training is over. Results discussed during teleconference. Training of EE1-1.1,2,3 started. EE1-4.1 adaptive resolution and QP selection strategy for RPR ready.

**T2 – 5 weeks after JVET-AG meeting (28-Feb-2024) 2nd teleconference.** Final setting for parameters to be announced, partial results discussed, combinational tests (if any) decided.

**T3 – 3 weeks before T5 (27-Mar-2023)**: EE1 software is frozen (write access closed on 29-Mar-2024).

**T4 – 3 days before T5 (13-Apr-2024):** Cross-checkers report status to EE1 coordinators (sending e-mail).

**T5 –16-Apr-2024:** EE1 summary is uploaded as input contribution.

# SW location

https://vcgit.hhi.fraunhofer.de/jvet-ag-ee1/VVCSoftware\_VTM/-/tree/EE1-X.X

# References

[1] E. Alshina, R.-L. Liao, S. Liu, A. Segall Common test conditions and evaluation procedures for neural network-based video coding technology, [JVET-AF2016](https://jvet-experts.org/doc_end_user/current_document.php?id=13272).