# Efficient Neural Video Coding for CLIC 2022

Runsen Feng, Zongyu Guo, Yixin Gao, Xiaohan Pan, Zhibo Chen
University of Science and Technology of China
fengruns@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

## Abstract

*This one-page fact sheet describes our proposed method for the video track of Challenge on Learned Image Compression (CLIC) 2022. Our scheme follows the typical hybrid coding framework with some new techniques. Firstly, we introduce a novel cross-scale inter-frame prediction module that can support fine-grained adaptation to diverse video content especially diverse motion cases. Secondly, we employ a spatially-variable-rate intra-frame codec and perform online encoder optimization to improve rate-distortion performance of each specific sample. Our team name is IM-CLVC.*

## 1. Introduction

In the video track of CLIC2022, participants are required to submit a codec to compress a set of 10-second sequences with diverse contents and frame rates. The winners will be chosen based on human perceptual visual quality. To meet the requirements, we propose an efficient neural video codec optimized for MS-SSIM. In this fact sheet, we briefly describe the design and training details.

## 2. Proposed Methods

The proposed end-to-end optimized video codec follows the typical hybrid coding framework with some new techniques.

Firstly, we apply a cross-scale inter-frame prediction module that achieves better motion compensation. Previous hybrid coding approaches rely on optical-flow [3] or Gaussian-scale flow [1] for motion compensation, which cannot support fine-grained adaptation to diverse video content especially diverse motion cases. Our proposed cross-scale prediction module is able to adapt to translational or complication motion. Specifically, on the one hand, we produce a reference feature pyramid as prediction sources, then transmit cross-scale flows that leverage the feature scale to control the prediction precision. On the other hand, we introduce the mechanism of weighted prediction into single-reference motion compensation, where cross-scale weight maps are also transmitted to indicate the importance of different feature scales. The detailed description of the cross-scale prediction module can be found in [2].

Secondly, to improve rate-distortion performance of a specific sample, we employ a spatially-variable-rate (SVBR) intra-frame codec and perform online encoder optimization. In this way, both the network parameters and the spatial bit allocation can be more adapted to each specific sample.

### 2.1. Network Structure

For inter-frame codec, the main network structures are almost the same as [2]. The only difference is that the quantized motion variables is at scales $\frac{H}{32} \times \frac{W}{32}$ instead of $\frac{H}{16} \times \frac{W}{16}$. For SVBR intra-frame codec, we use the network structure described in [2] and simply add SFT module [4] after each Resblocks of the transform encoder/decoder.

### 2.2. Training Details

First, We separately optimize the (SVBR) intra-frame codec and the single-rate video codec with MS-SSIM. The detailed training process of the two models can be found in [2, 4]. Then, to perform online encoder optimization, we fixed the network parameters of the decoder and jointly optimize the whole codec with Eq. 1 given a specific sequence.

$$\mathcal{L} = \mathcal{R}_I + \lambda \cdot \mathcal{D}_I + \sum_{t=1}^{T-1} (\mathcal{R}_t + \lambda \cdot \mathcal{D}_t). \tag{1}$$

Here, $R_I$-$D_I$ and $R_t$-$D_t$ represent the rate-distortion of the I-frame and the $t$-th P-frame.T is the training GoP size and $\lambda$ controls the trade-off of rate-distortion optimization.

## References

[1] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2020. 1

[2] Zongyu Guo, Runsen Feng, Zhizheng Zhang, Xin Jin, and Zhibo Chen. Learning cross-scale prediction for efficient neural video compression. *arXiv preprint arXiv:2112.13309*, 2021. 1

[3] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1

[4] Myungseo Song, Jinyoung Choi, and Bohyung Han. Variable-rate deep image compression through spatially-adaptive feature transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2380–2389, 2021. 1