|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| **N480** | 2021/12/22 |
| **Source** | Chuanmin Jia |
| **Title** | State of the art of end-to-end video coding |
| **Target** | MPAI Community |

# Introduction

This document is anticipated to provide a comprehensive overview for the state-of-the-art end-to-end video (EEV) coding solutions. The algorithmic details, technical analysis as well as the future research directions are elaborated.

# Report of selected papers

(up to ~1 page per paper)

*(a. List of papers considered, and more useful papers that might not be covered in discussion*

*b. Description of the algorithm, method*

*c. Comparison with other similar methods, improvements, why it improves, why does not improve…is not part but is used in chapter 3*

*d. …)*

Table 1 The presentations during MPAI-EEV meeting

|  |  |
| --- | --- |
| Presenter | Title |
| Chuanmin | DAST-NVC: Dual Attentional Spatial Transformer Inspired Neural Video Compression |
| Alessandra | End-to-End Learning for Video Frame Compression with Self-Attention |
| Antonio | ELF-VC: Efficient Learned Flexible-Rate Video Coding |
| Asfa | An End-to-End Learning Framework for Video Compression |
| Giovanni | Neural Video Compression Using Spatio-Temporal Priors |
| Giovanni | Learning for Video Compression With Recurrent Auto-Encoder and Recurrent Probability Model |
| Roberto | End-to-End Learning of Video Compression Using Spatio-Temporal Autoencoders |
| Roberto | Learning for Video Compression with Hierarchical Qualityand recurrent enhancement |
| Roberto | On the OpenDVC |

## DAST-NVC: Dual Attentional Spatial Transformer Inspired Neural Video Compression

### Description of the algorithm

This paper has two main contributions, which is shown in Fig. 1. First, the authors propose dual attention spatial transformer (DAST) for better motion compensation in EEV. Second, noticing the fact that in single-reference video coding schemes, the reconstruction error accumulates as the number of decoded frames gets larger, the authors incorporate the in-loop filter technique into the restoration of a decoded frame before it serves as the reference frame of the next frame to be encoded.

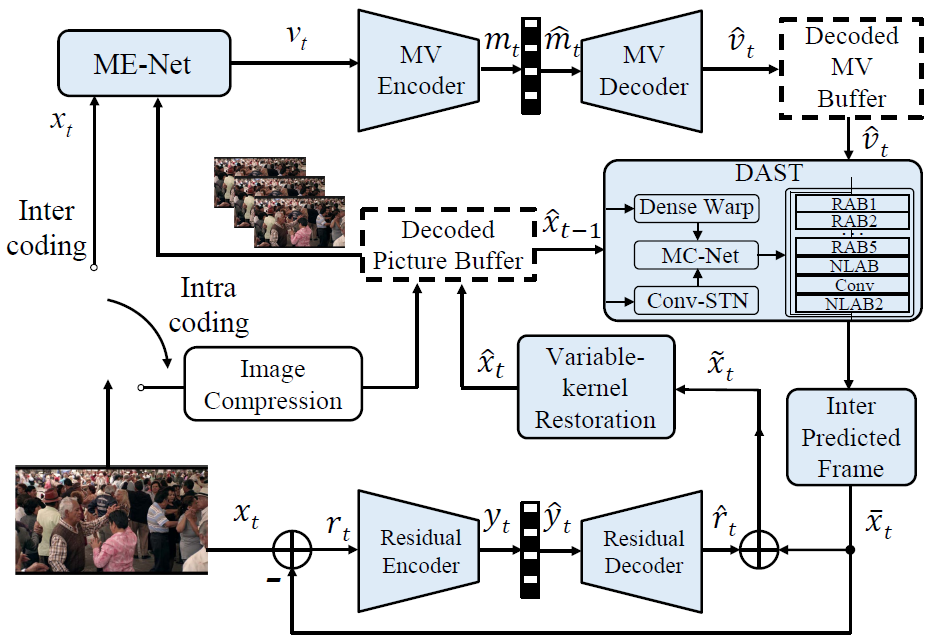


Fig. 1 The proposed DAST-NVC scheme, in which conv STN indicates spatial transformer network, ME/MC-Net is motion estimation/compensation network, RAB represents residual attention block and NLAB is the non-local attention block.

Depicted in Fig. 2, the DAST module is composed of three main parts. First, spatial transformer network (STN) provides a rough spatially-transformed frame. The second part called MC-net is a simple module consisting of several convolutional layers for motion compensation. After the motion compensation, a global identity connected module is designed to enhance the restoration quality of the predicted frame. As the DAST illustrates, the restoration quality enhancing module consists of two attention blocks which utilizes the capability of self-attention mechanism to better exploit both translational and non-translational movement information between adjacent frames, thus boosting the restoration performance. The in-loop filter module adopts variable kernel convolutions. The authors notice that current works of learning based image compression and quality restoration achieve outstanding performance. Therefore, they penetrate this success into the in-loop filter stage in EEV. Specifically, a well-designed, global identity connected, and residual attention block based deep network is proposed to enhance the reconstruction quality of the next reference frame.

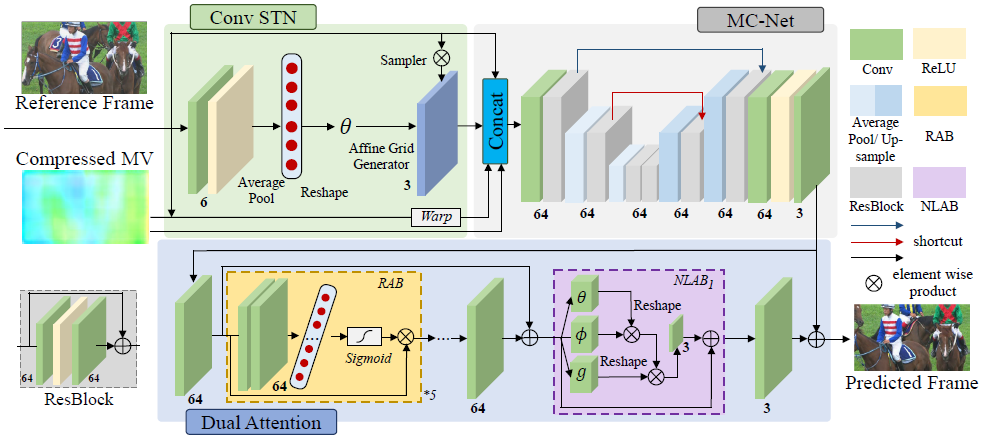


Fig. 2 The overall network architecture of the DAST based neural inter prediction module

This work surpasses the SOTA single-reference EEV scheme and achieves comparable coding performance with VVC. In detail, by proposing DAST, more than 48% and 42% bit-rate could be saved against x264 and x265 respectively under HEVC Class D dataset. Compared with other deep learning-based methods, the proposed method realizes 31.57% bit-rate reduction over the DVC method in Lu2019(CVPR) and obtain 21.08%, 14.96% and 6.50% coding gain over the schemes Hu2020, Lu2020 and Yang2020, respectively in HEVC Class D dataset.

The reason why this method achieves such superior performance can be divided into two parts. First, the self-attention mechanism can comprehend more motion patterns (e.g. non-translational motion, non-uniform velocity motion) than pure CNN, which helps to better restore the predicted frame from the reference frame and motion vector. Second, the proposed global identity connected in-loop filter network inherits the success of deep neural networks in image compression and quality restoration, and achieves better restoration quality of decoded frames which will serve as reference frames in the latter coding procedure. For the fact that in single-reference video coding schemes, the restoration error of frames grows with its index in the group-of-picture (GOP), the proposed in-loop filter can serve as a correlation machine of all decoded frames before it becomes a reference frame thus boosting the coding performance.

### Technology analysis

*(thorough comparison of methods in sec. 2, first level of conclusions in the sense that a technology is retained/rejected)*

a) Compared with the existing methods, this paper proposes a highly efficient motion compensation module. It is an extension version of DVC model with enhanced methods. Specifically, DAST consists of three major subnets, convolutional spatial transformer network, motion compensation network and dual attention module. On the top of the straightforward warping operation, DAST module takes full advantage of the reference frame and compressed MV, which substantially improves the inter predictive coding performance. First, the reference frame is spatially transformed given the learned affine parameters. Subsequently, the reference frame, warped frame, spatially transformed frame and MV are merged together to derive the final inter prediction results. Dual attention including both channel attention and non-local attention mechanism acts the core role in the motion compensation module. Since inter predictive coding dominates the overall coding efficiency, it is of significance to refine the motion compensation module.

A novel in-loop restoration network is appended into the coding loop to reduce the compression noise and enhance the reconstructed quality. Considering the trade-off between performance and complexity, the network is fully convolutional and has variable kernel size. The reconstructed frame is fed into the loop filtering network before added into the DPB, which could provide higher quality temporal reference for subsequent inter coding. The in-loop filtering module is subjected to become the major component in the end-to-end learned video compression framework due to its efficiency.

### Promising ideas

*(choosing the really promising elements)*

1. Enhanced networks for motion-compensation.
2. Residual based in-loop filtering network.
3. Multiple reference frame based inter-predictive coding

## “End to End Learning for Video Frame Compression with Self Attention”

by *Nannan Zou, Honglei Zhang, Francesco Cricri, Hamed R. Tavakoli, Jani Lainema, Emre Aksu, Miska Hannuksela, Esa Rahtu.*

2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)

Results of Nokia proposal of CLIC 2020 (Challenge on Learned Image Compression - P frame compression track)

### Description of the algorithm

Diagram

Description automatically generated

Figure 1: Overview of the solution

In this paper Nokia proposed an end-to-end approach with the following main features: 1) extracting the deep embeddings of a frames and coding their difference in the latent space 2) using an attention mechanism .

**Encoder side.** At the encoding side, the embeddings of frames ft and ft-1 are extracted, and the difference is computed as Δe. This difference Δe is encoded by the embedding encoder NN.

An additional Neural Network (Importance Map Net in the Figure) analyzes the embedding-difference Δe and outputs a binary importance mask in the range of [0,1] denoted as *m*. This Importance Map zeros-out a varying number of channels and the output ym is represented by:

A picture containing text

Description automatically generated

ym is then quantized by 8 bit scalar quantization and then entropy coded by an arithmetic encoder.

**Decoder side.** At decoder side, bitstream is entropy decoded and dequantized; then the embedding decoder derives  and the reconstructed embedding for the current frame ft  is computed by adding the reconstructed embeddings of ft-1. The frame decoder neural network outputs a first version of the current frame by reprojecting the embeddings into the pixel space and an attention mechanism is applied to reconstruct Diagram

Description automatically generated.

**The attention mechanism.** The main idea is to combine the current frame and the previous frame taking into account the previous frame for all the detailed information already present in the reconstructed frame (pixel space) and the current frame only for the changed parts. The self-attention Neural Network analyses the two embeddings Diagram

Description automatically generatedand Diagram

Description automatically generated and outputs an Attention Map for Diagram

Description automatically generated and an Attention Map for ft-1 in the range [0,1]. These maps are then applied to the initial version of the current frame and the previous frame to reconstruct .

**Performances.** The proposed algorithm has been tested in the CLIC validation set and achieves MS-SSIM of 0*.*978, Peak Signal-to-Noise Ratio (PSNR) of 30*.*44*dB*, bits-per-pixel (BPP) of 0*.*0707.

## An End-to-End Learning Framework for Video Compression

1. **Introduction**

This document is aimed at providing a comprehensive overview of Paper “An End-to-End Learning Framework for Video Compression”. It gives an insight to the proposed framework, a brief insight to each module, technical advantages of proposed approach and any future prospects.

1. **Report of selected papers**

a. **Main Paper:**

Lu, G., Zhang, X., Ouyang, W., Chen, L., Gao, Z. and Xu, D., 2021. An End-to-End Learning Framework for Video Compression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10), pp.3292-3308.

b. **Algorithm Description and method**

This paper proposes an end-to-end video compression (DVC) model that replaces modules in traditional video compression pipeline with neural networks and jointly optimize them.

**The proposed video compression method comprises of following main features:**

1. Replacement of tradition motion estimation, motion compensation, residual compression, motion compression, and bit rate estimation modules with an end-to-end neural network.
2. Joint optimization of all modules based on rate-distortion tradeoff using a single loss function
3. Provision of flexible models DVC, DVC Lite, DVC Pro, keeping speed/ efficiency under consideration.
4. Reduction of parameters in variable bitrate coding using adaptive quantization technique.

**Overview of the Proposed End-to-end Deep Video Compression:**

Model follows the hybrid coding framework of motion-compensated prediction and residual transform coding, but all modules in this proposed approach are designed and implemented using convolution

networks.

Diagram

Description automatically generated

1. **Motion estimation Net:**

In order to estimate motion information, learning based optical flow method SpyNet is used which estimates optical flow between two neighboring frames using pyramid architecture.

Reference:

Ranjan and M. J. Black, “Optical flow estimation using a spatial pyramid network,” in CVPR, vol.2. IEEE, 2017, p. 2. 3, 4, 6, 7

1. **MV Encoder and Decoder Net**

An auto-encoder style network is used to compress pixel-level optical flow from motion estimation

Network.

Reference:

J. Ball´e, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in 5th International Conference on Learning Representations, ICLR, 2017. 1, 2, 5, 6, 15, 16

1. **Motion Compensation Net**

In motion compensation network, previous frame is warped around the motion vector. The obtained warped framed is concatenated with previous frame and sent to neural network. Network consists of Convolution layers followed by residual block and pooling layer.

1. **Residual Encoder and Decoder Net**

Here, an auto encoder network based on several convolution layers and GDN/IGDN layers is used

to compress residual information.

1. **Bit Rate Estimation Net**

An auto encoder network based on several convolution layers and GDN/IGDN layers is used

To estimated bit rates.

1. **Network Optimization:**

There are two main objectives:

1. Reduce number of bits
2. Reduce distortion between decoded and original image.

Loss function:



* λ determines the trade-off between the number of bits
* d() denotes the distortion and can be measured by mean square error (MSE) or multiscale structure similarity
* H(⋅) represents the number of bits used for encoding the representations

**c. Comparison with other similar methods, improvements**

In terms of PSNR score, the DVC (MSE) achieves 0.6dB gain at the same bpp level on the UVG dataset as compared to traditional codecs and recent deep learning framework.

The proposed DVC(MS-SSIM) method outperforms the H.265/H.264 codecs by a large margin when measured by MS-SSIM. For example, the DVC(MS-SSIM) model has more than 0.005 gain in terms of MS-SSIM when compared with H.265 on the

HEVC Class C dataset.

1. **Technology analysis**

**PSNR evaluation:**

On most datasets, the proposed MSE based model outperforms the H.264 standard when measured by PSNR. It can also be observed that the performance of DVC Lite(MSE) is similar to DVC(MSE), while DVC Pro (MSE) outperforms DVC(MSE), especially for the HEVC Class C dataset. More importantly, the DVC Pro (MSE) model even achieves comparable compression performance with H.265 in terms of PSNR, which demonstrates the potential of the learning-based video compression approach.

Chart, line chart

Description automatically generated

MS-SSIM evaluation:

MSE based models achieve comparable or better compression performance than

H.265 in terms of MS-SSIM. It proves that the proposed framework can generate the reconstructed frames with better perceptual quality.

Chart

Description automatically generated

The DVC(MS-SSIM) model has more than 0.005 gain in terms of MS-SSIM when compared with H.265 on the HEVC Class C dataset.

Chart, line chart

Description automatically generated

**Bit rate:**

It is observed that the proposed DVC(MSE) can save more than 29% bit rate, while H.265 only saves 21.73%. Besides, the MS-SSIM based model DVC(MS-SSIM) saves up to 45.88%-bit rate when compared with H.264.

**Qualitative Comparison:**

Specifically, when compared with H.264/H.265, the proposed DVC model generates high-quality reconstructed frame at the same bpp level.

1. **Promising ideas**

* This is the first end to end deep learning model for video compression
* By jointly optimizing the loss function, the distortion between original and decoded frames is reduced significantly.
* The flexible model design help in tweaking the model based on requirements. For instance, If performance is ultimate goal, then DVC Pro can be used but if efficiency is required, then , it is better to DVC lite.
* Flexible model also allows to replace any module and thus, allow architecture to be modified easily.

1. **Potential reference model(s)**

This paper is an extension of DVC: An end-to-end deep video compression framework. It takes benefits from hybrid video compression framework and deep neural networks.

1. **Conclusions**

To conclude, this paper provides a deep learning approach to video compression problem. This approach takes full advantage of hybrid coding framework of video compression and replaces modules with neural networks. Also, the whole network is jointly optimized using a single loss function. This approach outperforms the traditional codecs and other recent state of art deep learning video compression frameworks.

## ELF-VC: Efficient Learned Flexible-Rate Video Coding

1. Introduction

The work of Rippel *et al.* is focused on achieving good bit-rate flexibility through a single set of learned parameters. In addition an efficient optimized backbone and a novel in-loop flow prediction scheme are proposed. The architecture shown in Figure 1 is composed by:

* the flow predictor block, that takes as inputs the previous reconstruction and the previous flow.
* the flow block, that refines the already defined flow obtained via the flow predictor block.
* the residue block, which reconstruct the output frame.

We can further notice the level maps present in the architecture, which are embedded in the learned operators. This one-hot vector allow all the networks present in this architecture to learn on multi-rate dataset. To that end, linear interpolation in the embedding space was used to target intermediate rates. Furthermore, this method does not require any additional training in order to interpolate the rate without sacrificing performance. This approach also helps to mitigate discontinuity regarding the rate curve. Old methods showed non-monotonic behavior a when varying discrete rate and quantization. In this approach the single rate parameter helps the quality increase monotonically as the bitrate increases.

Diagram

Description automatically generated

The backbone used for all the learnable operators is shown in Fig 2, it is based on DenseNet and provides for a lower number of concatenations, allowing faster models according to authors. The block referred also as Delayed Merge (DM) block, represents a balanced tradeoff between expressivity and computational performance. The original formulation, DenseNet, consists of many concatenation operations and convolutions with a small number of filters. This does not lend itself to efficient computation. With the Delayed Merge block the concatenation is performed only a single time hence the dimensionality is reduced.

Text

Description automatically generated

Results of the flow predictor block approach, can be seen in Figure 3. We can appreciate in Figure 3 that the (zero-bit) predicted flow (top left) is already very similar to the final flow (top right). So the flow block only needs to add some minors sparse touch ups, spending less bits.

Graphical user interface

Description automatically generated with medium confidence

2. Technology analysis

Benchmarks performed by the authors show improvements over both other ML-based codecs and traditional ones (44% BD-rate gain against H.264, 15% against AV1). Source code for this work is not available in open-source form, nonetheless some aspects and techniques are adoptable for the scope of MPAI.

A screenshot of a computer

Description automatically generated with low confidence

4. Promising ideas

1. Flow predictor block
2. Flexible rate single-parameter models

5. Potential reference model(s)

NA

6. Conclusions

This work leverages the flexibility of deep learning methods to achieve a flexible multi bit-rate encoding schema. It can be useful to highlight the use of the single set of learned parameters utilized to perform the various internal operations, against the multi-model approach utilized by previous methods.

## NEURAL VIDEO COMPRESSION USING SPATIO-TEMPORAL PRIORS

*Haojie Liu, Tong Chen, Ming Lu, Qiu Shen, and Zhan Ma, School of Electronic Science and Engineering, Nanjing University*, available here <https://arxiv.org/abs/1902.07383>

### Description of the algorithm

The basic idea described in the paper is to change “hand-crafted coding tools” in the classical hybrid codec architecture with AI-based coding tools. The typical block predictions (priors) created in the encoder are changed in AI-based, hopefully more efficient, priors. Spatial priors are created from downscaled low-resolution features while temporal priors, generated from previously co-decoded frames (references) are capture using a convolutional LSTM structure.

The coding tools are trained jointly (as opposed to single coding tool fine tuning characterizing the typical approach) to optimize rate-distortion performance.

A high level illustration of the architecture is given in the following figure

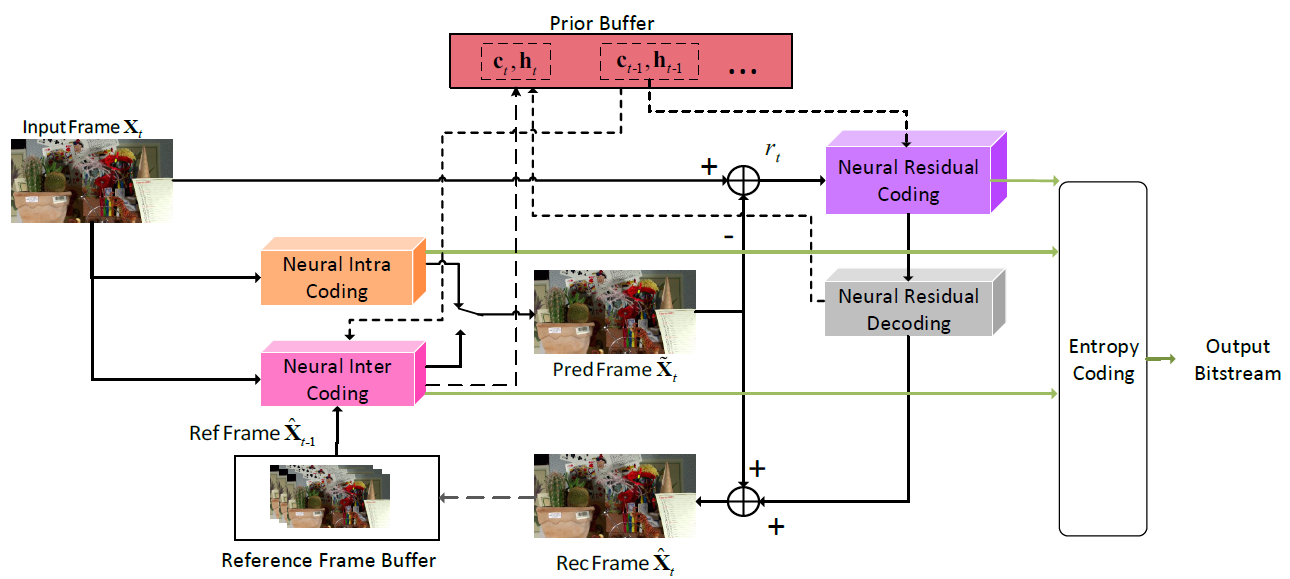


Figure 2 - High level architecture

The intra-coding part (illustrated below together with quantization and probability models adaptation) is based on an autoencoder with embedded hyperpriors which is followed by a quantization stage (uniform noised used in the training phase with rounding at the quantization stage). The distribution of the quantized features and the hyperpriors are used to optimize the probability distributions used in the entropy coding phase (Gaussian distributions are assumed, thus the µ and σ parameters are determined). The ICN network is used to fuse and concatenate decoded hyperpriors and residuals for the reconstruction.

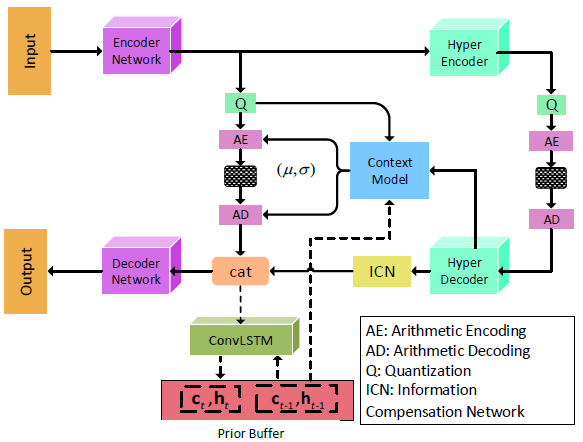


Figure 3 - Intra coding, quantization and probability adaptation

Inter-coding (illustrated in the figure below), is based on optical flow calculation, warping of the decoded temporally preceding frame followed by super-resolution filtering to eliminate blurring and a parallel temporal recurrent network (ConvLSTM) to capture high frequency details that would otherwise be lost.

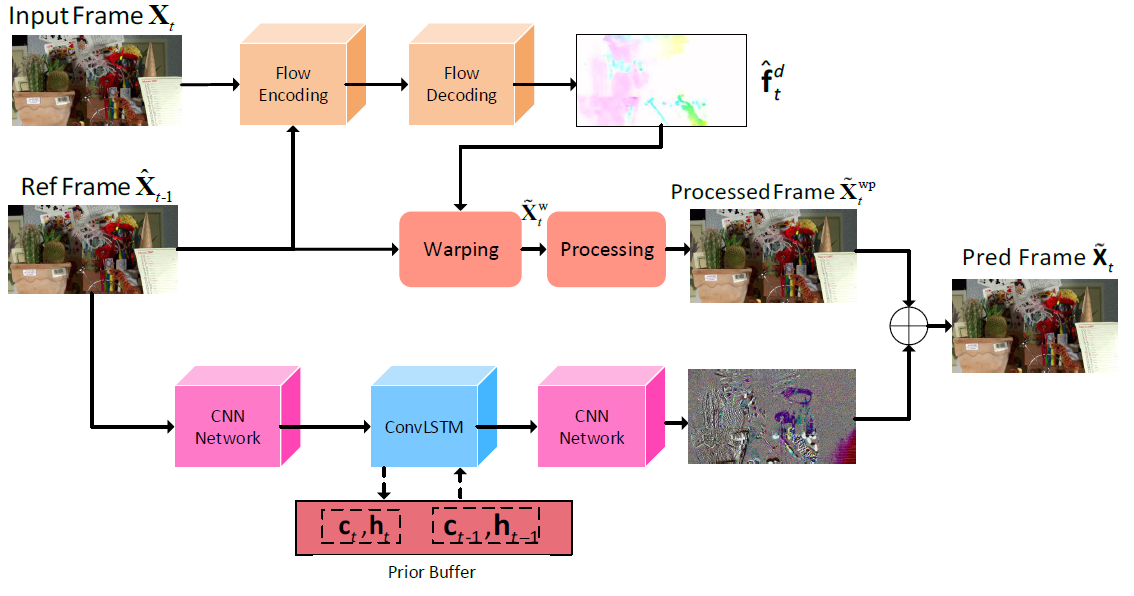
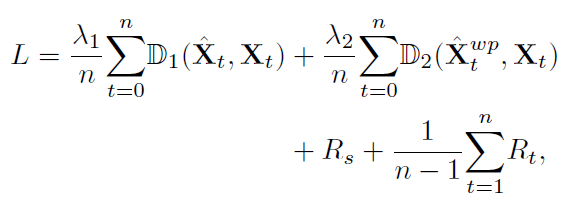


Figure 4 - Inter coding

Residuals are coded using an architecture similar to that used for intra coding, augmented with a ConvLSTM to keep track of temporal dipendencies.

Every single network is independently trained at first, then a joint training phase follows to optimize for rate-distortion using the cost function below



D1 representing MS-SSIM and D2 the warp loss.

### Results

The authors claim an average 38% improvement over HEVC but:

* Test have been made against x265 which is known to be on average 14% less efficient than HM reference software (see, e.g., <https://ieeexplore.ieee.org/document/7906321>)
* The GOP configuration chosen to create the references is IPPPPPPP, with encoder configured in low-delay mode, hardly providing a realistic evaluation of HEVC performance (see, e.g., <https://ieeexplore.ieee.org/document/4037003>)

Claimed figures should be thoroughly evaluated but are likely to be largely inferior to effective HEVC performance.

## Learning for Video Compression With Recurrent Auto-Encoder and Recurrent Probability Model

*Ren Yang , Fabian Mentzer, Luc Van Gool , Member, IEEE, and Radu Timofte , Member, IEEE,* available here <https://ieeexplore.ieee.org/abstract/document/9288876>

### Description of the algorithm

The paper proposes a recurrent approach to challenge the “hand-crafted” tool selection approach used to design well-known hybrid codecs such as AVC and HEVC. The basic idea is to define a framework that can be optimized and-to-end instead of relying on single tool customization and tuning.

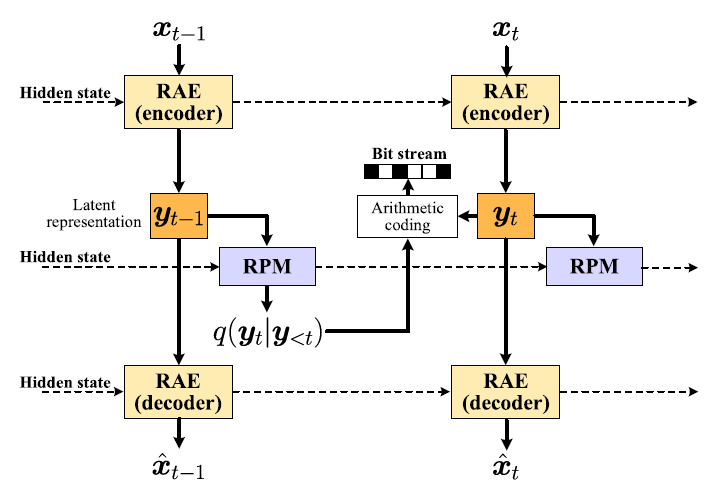


Figure 5 - High level architecture

Recurrent autoencoders (RAEs) are used to create a representation in the latent space of frames at a given instant in time (*x*ts) that are quantized and entropy coded. A recurrent probability model provides updates of the probability distributions needed for entropy coding.

A more detailed view of the coding structure is given in the following figure

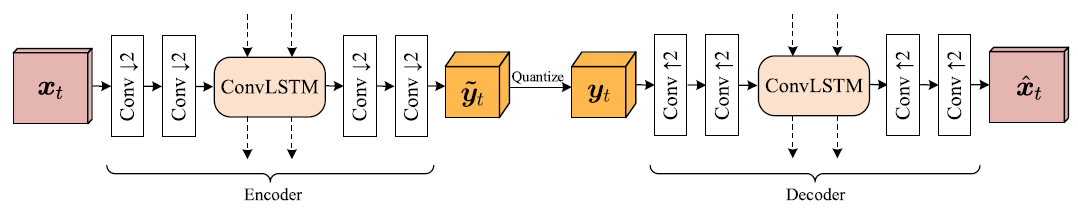


Figure 6 - Hidden state propagation

As indicated by the dashed lines, hidden states are propagated to take into account long term dependencies between representations of frames in the latent space.

Motion compensation and estimation are implemented using optical flow detection of motion vector fields.

The structure of the recurrent autoencoders is illustrated in the following figure



Recurrent cells in the encoder and the decoder is needed to account for long term dependencies from remote frames that would otherwise be lost.

The RPM network shown below is used to entropy code the sequence of latent representations



The training follows a stepped approach. At first, the motion estimation network is trained using the loss function



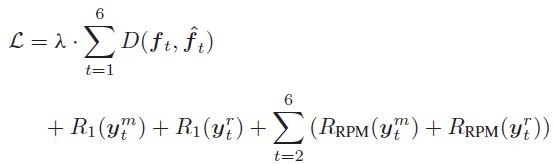
x being the input frames and W the warping operation. After convergence, the part that account for motion compression and motion compensation is added



and a final step of joint training is added at last



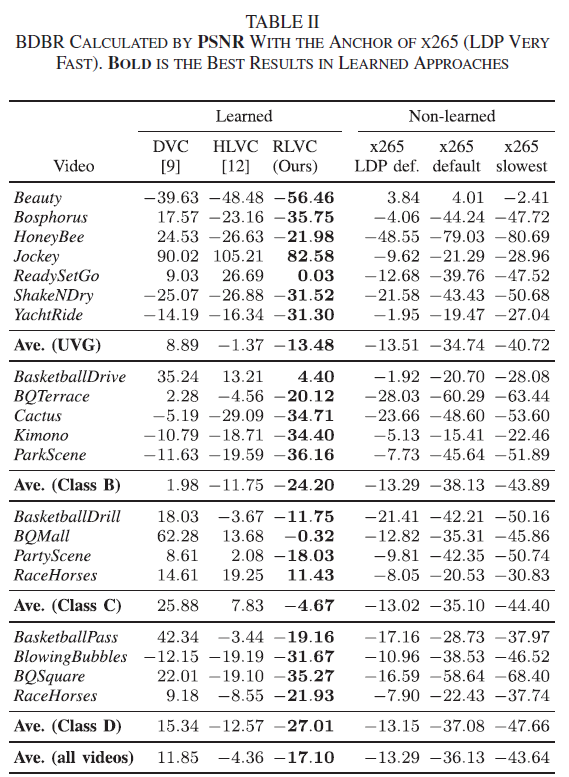
The resulting end-to-end training function is



### Results

In this case also, results have been evaluated using x265 as the reference encoder in low delay mode with IPPPPP GOP structure (see comments already made in paragraph 2.5.2 above.

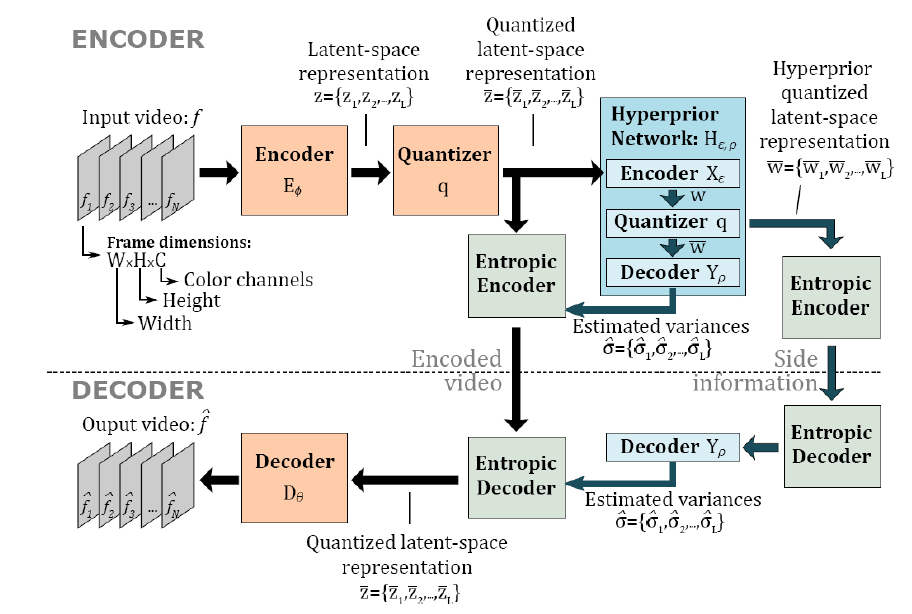
The table below summarizes results of the current RLVC model against other models known in the literature and against x265.



It has to be noted that RLVC model seems to be outperformed by x265 when non in “Very fast” (the most unfavorable configuration used for very high speed encoding) mode.

## End-to-End Learning of Video Compression Using Spatio -Temporal Autoencoders

Yang, R., Mentzer, F., Van Gool, L., & Timofte, R. (2020). Learning for video compression with hierarchical quality and recurrent enhancement. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6627–6636. <https://doi.org/10.1109/CVPR42600.2020.00666>



The authors propose a rate-distortion optimization approach for video compression that, in addition to learning a 3D latent-space representation of a video, enforces temporal consistency between frames without introducing undesirable artifacts, e.g. flickering.

This architecture is quite similar to the one from Ballè et al., ‘Variational image compression with a scale hyperprior’, where there are two transforms: a projection transform to represent the original video in a latent space (in Figure x called Encoder), and a synthesis transform to reconstruct the video into the original space (in Figure x called Decoder).

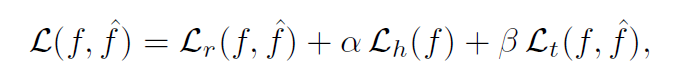
Usually video is thought as a sequence of 2D frames that are provided to the network. Their approach is based on considering a video as a 3D object handled by 3D convolutions to achieve a better performance, as it processes spatial and temporal information simultaneously.

The Hyperprior network computes the distribution of the quantized latent space representation and feed these statistics into the entropy coder to optimize the probability distributions used in the entropy coding phase.

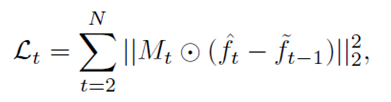
In the training phase to overcome the problem that quantizer would block the backpropagation

of the gradients through the bottleneck, quantization is replaced by additive uniform noise.

The network is end-to-end optimized for both the distortion (Lr) and the rate (entropy, Lh). The interesting part of this paper is the third loss term that take into account the temporal consistency between consecutive frames (Lt):



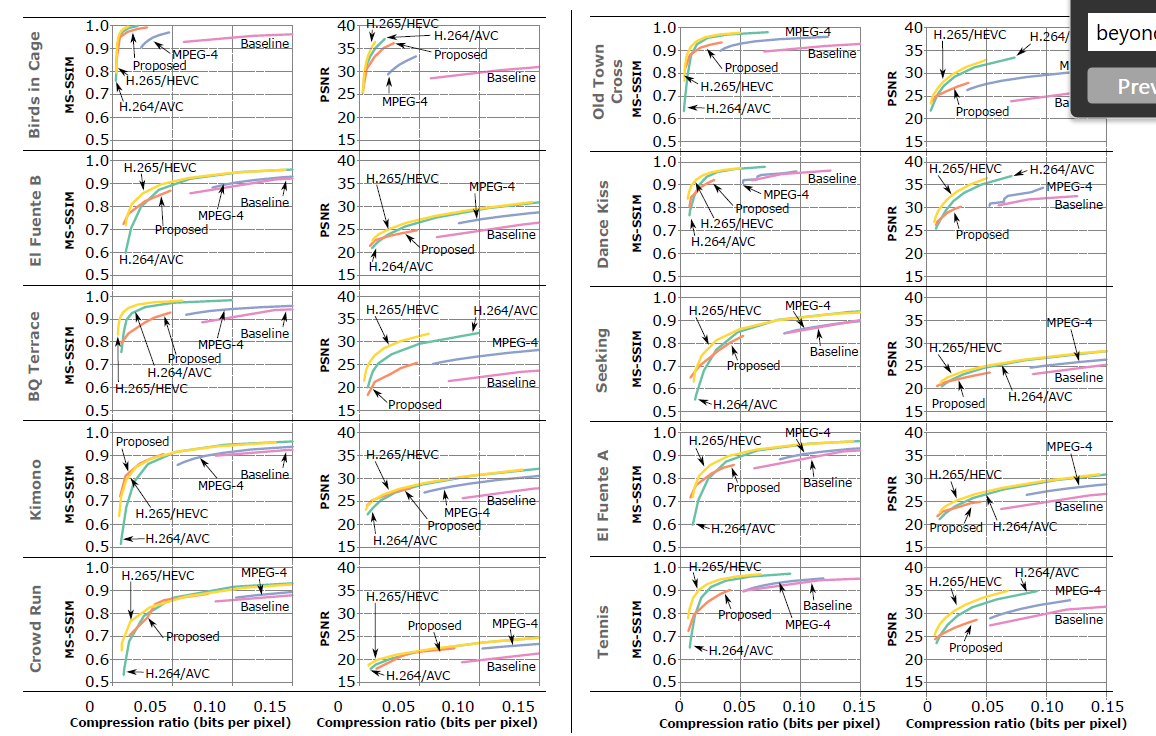
This temporal loss is defined as:



where ⊙ denotes pixel-wise product, is the result of warping the frame to time t - 1 by using the estimated backwards optical flow from to and Mt is a binary occlusion mask excluding pixels that are not present in both and

The meaning of this term is that a pixel in the previous frame can move anywhere in the current frame but it must be that pixel without any changes in luma, because it will lead to the flickering effect.

### Results



Individual rate-distortion curves for the compressed videos (RGB color space)

The authors claim that their approach has competitive performance to both H.265/HEVC and H.264/AVC at the lower bitrates.

### Promising ideas

The paper contains interesting ideas:

1. video considered as a 3D object
2. loss function that take into account temporal consistency

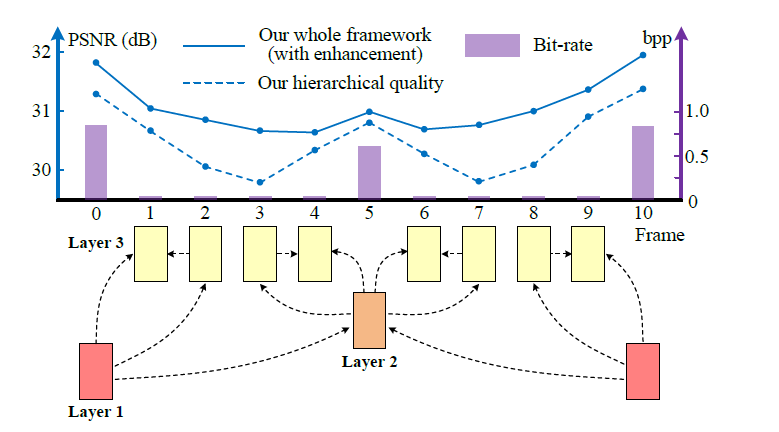
### Pain points

From the Figure x is clear that their method very few times is over HEVC in the low bit rate but most of the time is under HEVC. They don’t provide an overall BD-rate.

Moreover, the authors don’t share the HEVC parameters that have used so it is difficult to judge their results.

## Learning for Video Compression with Hierarchical Quality and Recurrent Enhancement

Yang, R., Mentzer, F., Van Gool, L., & Timofte, R. (2020). Learning for video compression with hierarchical quality and recurrent enhancement. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6627–6636. https://doi.org/10.1109/CVPR42600.2020.00666



The hierarchical layers and the rate-distortion performance

This paper proposes a learned video compression approach with hierarchical quality and recurrent enhancement. In the Figure x the frames in layers 1, 2 and 3 are compressed

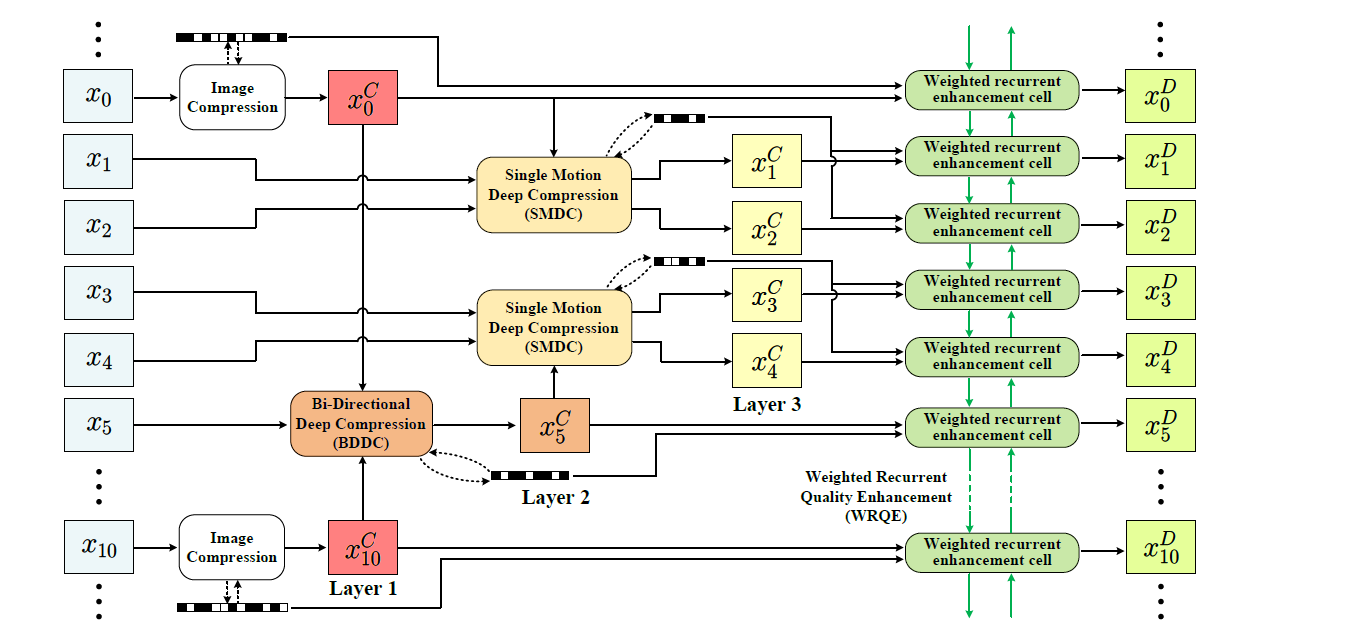
with the highest, medium and the lowest quality, respectively.

Two main ideas justify their approach, one encoder side and the other decoder side:

* + 1. the high quality frames are able to improve the compression performance of other frames at the encoder side because they provide high quality references

* + 1. due to the high correlation among neighbouring frames, the low quality frames can be enhanced by making use of the redundant information in high quality frames, at the decoder side. The enhancement is done without bit-rate overhead.

For example, the frames 3 and 7 in Figure x, which belong to layer 3, are compressed with low quality (PSNR under 30) and the bit-rate is very low (the histogram is negligible). After their approach the PSNR is almost 31 for both frames, without increasing in bit-rate.



The overall framework of our HLVC approach

This architecture is quite complex: each block is another architecture of neural networks rising concerns about complexity.

On the left of Figure x there is Group of Picture (GOP), in this case 10.

The first layer is encoded by image compression method BPG, Better Portable Graphics, and

denotes the compressed frames. This is similar to the “I-frames” in traditional codecs.

Pictures in layer 2 are processed by the Bi-directional Deep Compression (BDDC) network that takes the previous, e.g. and the upcoming, e.g. compressed frames from layer 1 as bi-directional references. To take into account that the interval between the frames in layers 1 and 2 is long (e.g., 5 frames in Figure x), the authors apply a pyramid network to handle large motions.

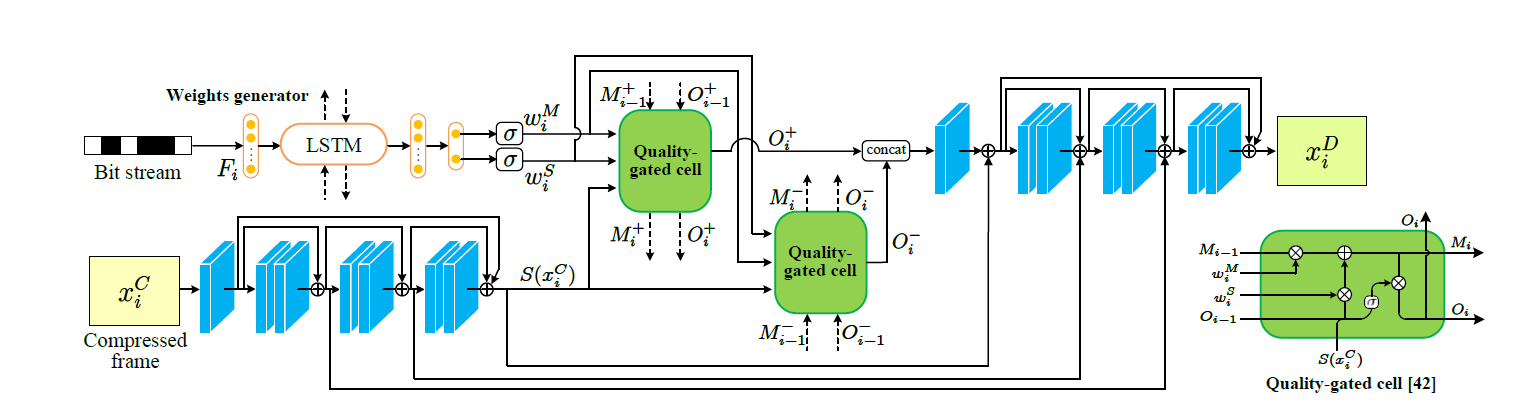
Another interesting idea it to use backward warping i.e. they estimate backward motions to handle occlusions and dis-occlusions.

The Single Motion Deep Compression (SMDC) network, applies a single motion map to describe the motions between multiple frames, and therefore the bit-rate can be reduced.

As illustrated in Figure x, the frames x1 and x2 are compressed using a single motion map with

the reference of , while and x3 and x4 use as reference.

BDDC and SMDC for motion compression and residual compression are based on the neural network architecture from Ballè et al., ‘Variational image compression with a scale hyperprior’ (see 2.7).



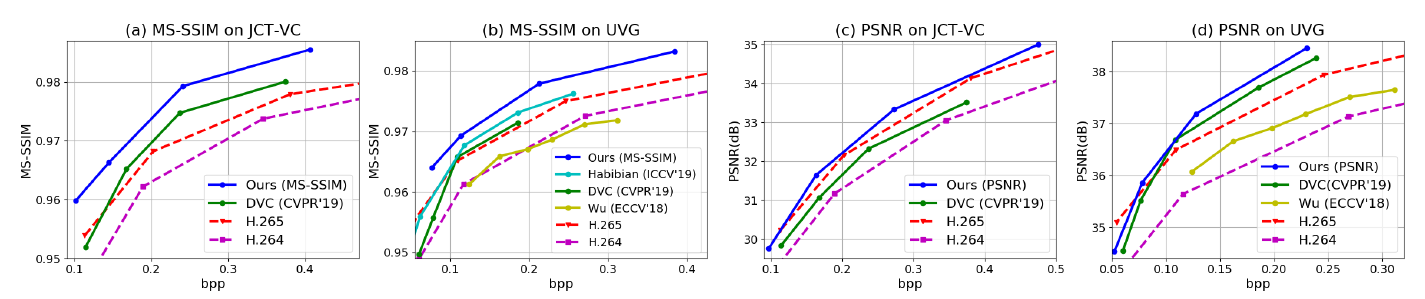
The architecture of our WRQE network

The last architecture is the Weighted Recurrent Quality Enhancement (WRQE) network, in Figure x, based on recurrent cells (ConvLSTM). These are network with memory and they have good performance with sequential data. Frames that belong to different layers of quality are fed to the network in order to exploit the high correlation among frames. It has to be noted that the compression quality is encoded in the bit stream and this info is fed to the recurrent cell to tka into account that the three layers are very different in quality and this info help the network to differentiate the picture that belongs to different layers.

### Results

In Figure x the rate-distortion curves on the JCT-VC and UVG datasets. The quality is evaluated in terms of MS-SSIM and PSNR, and the bit-rate is calculated by bits per pixel (bpp). As

shown in Figure x (a) and (b), the MS-SSIM model outperforms all learned approaches, and it has better performance than AVC and HEVC. The PSNR curves are illustrated in Figure 6 (c) and (d). It can be seen that the PSNR model outperforms HEVC on the JCT-VC dataset while on UVG, it reaches better performance than HEVC at high bit-rate.



### Promising ideas

There are interesting ideas in this paper:

1. backward warping to manage occlusion
2. pyramid network to handle large motions
3. full advantage of the high quality frames (layer 1) both at encoder and decode side
4. Multiple reference frames

### Pain points

The results of the paper should be thoroughly evaluated because of some concerns:

1. the trade-off between performance and complexity because of the many neural network architectures
2. In the Figure x the authors compare with 5 other methods and sometime with 4 or just with 3 methods without any explanation
3. The comparison is made with x265 and not with HEVC Test Model (HM reference software)
4. The comparison is made with x265 Low Delay P but they are using B-picture (in the Bi-directional Deep Compression)
5. The x265 setting is ‘very fas’t mode that gives low quality
6. The authors compared their method with the worse possible configuration for x265

## OpenDVC: An Open Source Implementation of the DVC Video Compression Method

Yang, R., Van Gool, L., & Timofte, R. (2020). *OpenDVC: An Open Source Implementation of the DVC Video Compression Method*. http://arxiv.org/abs/2006.15862

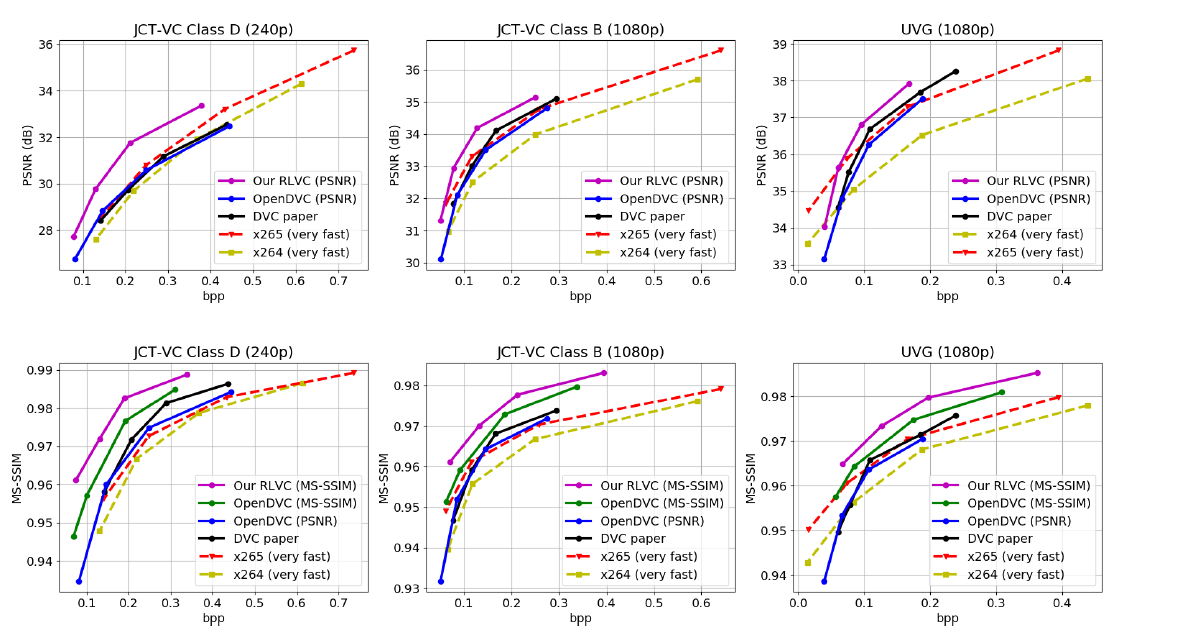
This paper proposes an open source implementation of DVC: An end-to-end deep video

compression framework (see Section 2.3). The main idea is to replace traditional motion estimation, motion compensation, residual and motion compression, and bit rate estimation modules, used in traditional hybrid block based video codec, with neural networks and jointly optimize them using a single loss function.

Different from the original DVC they implemented a PSNR-optimized model, denoted

by OpenDVC (PSNR), and they introduced a MS-SSIM-optimized model OpenDVC (MS-SSIM).

### Results



The performance of DVC and OpenDVC

The rate-distortion performance of OpenDVC is demonstrated in Figure x. It can be seen that the OpenDVC (PSNR) model achieves comparable performance with DVC and x265, and openDVC (MS-SSIM) outperforms DVC and x265 in terms of MS-SSIM. It has to be noted that the authors other approach, RLVC, see Section 2.6, outperforms both OpenDVC and x265.

### Promising ideas

There are interesting ideas in this paper:

1. pyramid network to estimate the motion between the current frame and the previous compressed frame
2. MS-SSIM-optimized model
3. Joint optimization of all modules using a single loss function to improve rate-distortion performance

### Pain points

The results of the paper should be thoroughly evaluated because of some concerns:

1. The comparison is made with x265 and not with HEVC Test Model
2. The authors used the x265 ‘very fast’ mode setting that gives low quality
3. The authors compared their algorithm with the worse possible configuration for x265 (LDP very fast)

# Potential reference model(s)

*(integration of elements in chapter 4 into reference model(s) may be part of another document)*

1. It is suggested to adopt the DVC/OpenDVC model as the reference model.
2. **Candidate technologies**
   1. **Technology 1: enhancement of compressed images in DVC**

Pros:

1) well-studied in image/video processing, quality enhancement has been hot topics for plenty of years. Plenty of prior existing knowledge could be borrowed;

2) the networks of post-processing have relatively smaller impact on the encoding of motion vectors and residuals than other tools in the coding loop.

3) directly reduce the coding error, which obvious brings coding gain

Cons:

1) no parameters, no gains. Often result in large quantities number of neurons.

2) complexity

Discussion:

* 1. **Technology 2: motion vector prediction**

Pros:

1) reduce the overhead of motion vector field (optical flow) coding, higher coding performance

Cons:

1) larger encoder/decoder buffer

2) design efficient prediction scheme is time-consuming, i.e., select reference frames, signalling

3) should it contain trainable parameters?

Discussion:

* 1. **Technology 3: multiple reference frame based motion compensation prediction (MCP)**

Pros:

1) rich temporal information for MCP, less prediction residual

2) better consistency between adjacent frames, higher visual quality

3)

Cons:

1) larger encoder/decoder buffer and higher complexity

2) signalling the index of ref frame, how many, which frames to be used

1. **First technology selection**

What is needed to develop the software required to implement the selected technologies in the software. (Should be further discussed)