Learning for Video Compression with Hierarchical Quality and Recurrent Enhancement –Supplementary Material–

6. Details of our framework

6.1. The BDDC network

ME subnet. In our approach, we employ a 5-level pyramid network [26] for motion compensation, with the same structure and settings as [26]. However, [26] trains the network with the supervision of the ground-truth optical flow, but in our approach, we pre-train the ME subnet by minimizing the MSE between the warped frame and target frame. That is, we use the loss function of

$$L_{\text{ME}} = D(x_5, W_b(x_0^C, f_{5\to 0})) + D(x_5, W_b(x_{10}^C, f_{5\to 10})) \tag{14}$$

to initialize our ME subnet before jointly optimizing the whole BDDC network by (12) in our paper. Figure 10 shows the example frames warped by estimated motions, which are trained by ground-truth optical flow and the MSE loss of (14), and we also show the PSNR between the warped and target frames. It can be seen that the MSE optimized motion is able to reach higher PSNR for the warped frame, thus leading to better motion compensation.

x_0^C

x_5

Warped by ground-truth optical flow trained motion

PSNR = 24.29 dB

Warped by MSE trained motion $W_b(x_0^C, f_{5\to 0})$

PSNR = 29.73 dB

Figure 10. Example frames warped by estimated motions, which are trained by ground-truth optical flow and the MSE loss of (14).

MC and RC subnets. We follow [2, 3] to use the CNN-based auto-encoders in our MC and RC subnets, and they have the same structure in our approach. The detailed parameters are listed in Tables 2 and 3, in which GDN denotes the generalized divisive normalization [2] and IGDN is the inverse GDN [2].

MP subnet. We follow the motion compensation network in [22] to design our MP subnet, which is illustrated in Figure 11. In our MP subnet, all convolutional layers use a filter size of 3×3. The filter numbers of all layers excluding the output layer are 64, and the filter number of the output layer is set to 3. We use ReLU as the activation function for all layers.

Table 2. The encoder layers in the MC and RC subnets.

Layer	Conv 1	Conv 2	Conv 3	Conv 4
Filter number	128	128	128	128
Filter size	5×5	5×5	5×5	5×5
Activation	GDN	GDN	GDN	-
Down-sampling	2	2	2	2

Table 3. The decoder layers in the MC and RC subnets.

Layer	Conv 1	Conv 2	Conv 3	Conv 4
Filter number	128	128	128	3
Filter size	5×5	5×5	5×5	5×5
Activation	IGDN	IGDN	IGDN	-
Up-sampling	2	2	2	2

In our **SMDC network**, the subnets of ME, MC, MP and RC have the same architecture as introduced above.

6.2. The WRQE network

In the WG subnet of our WRQE network, we set the hidden unit number in the bi-directional LSTM as 256, and thus the layer d_1 has 512 nodes. We use 5×5 convolutional filters in all convolutional layers (the architecture is shown in Figure 5 of our paper). The filter numbers are all set to 24 before the output layer, and the filter number for the output layer is 3. We use ReLU as the activation function for all convolutional layers.

7. Additional experiments

Configurations of x264 and x265. In Figure 6 and Table 1 of our paper, we follow [22] to use the following settings for x264 and x265, respectively:

```
x264: ffmpeg -pix_fmt yuv420p -s WidthxHeight
-r Framerate -i Name.yuv -vframes Frame
-c:v libx264 -preset veryfast -tune
zerolatency -crf Quality -g 10 -bf 2
-b_strategy 0 -sc_threshold 0 Name.mkv
x265: ffmpeg -pix_fmt yuv420p -s WidthxHeight
-r Framerate -i Name.yuv -vframes Frame
-c:v libx265 -preset veryfast -tune
zerolatency -x265-params
"crf=Quality:keyint=10:verbose=1" Name.mkv
```

In the commands above, we use Quality = 15, 19, 23, 27 for the JCT-VC dataset, and Quality = 11, 15, 19, 23 for UVG videos.

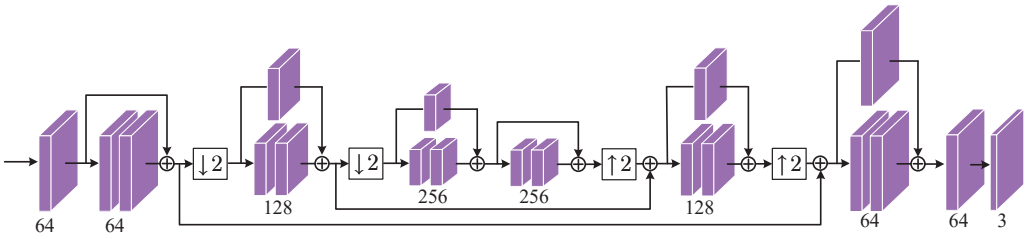


Figure 11. Architecture of our MP subnet.

Figure 12. Example frames after motion compensation in our SMDC network at $\lambda = 1024$.

Motion estimation in our SMDC network. Recall that, in our SMDC network, $\tilde{x}_2$ is generated by motion compensation with the compressed motion $\hat{f}_{2\to0}$. Then, we estimate the motions of $\hat{f}_{1\to0}$ and $\hat{f}_{1\to2}$ from $\hat{f}_{2\to0}$ using the inverse operation (refer to (8) and (9) in Section 3.3) to generate $\tilde{x}_1$.

However, as shown in Figure 12, $\tilde{x}_1$ has even higher PSNR than $\tilde{x}_2$. Moreover, Table 4 shows the averaged PSNR of $\tilde{x}_1$ and $\tilde{x}_2$ among all videos in the JCT-VC dataset. The results in Table 4 also prove that our SMDC network generates $\tilde{x}_1$ with higher PSNR than $\tilde{x}_2$. It is probably because of the bi-directional motion used for $\tilde{x}_1$, and the shorter distance between $\tilde{x}_1$ and the reference frame x_0^C. These results validate that our SMDC network accurately estimates multi-frame motions from a single motion map.

Table 4. Average PSNR (dB) of $\tilde{x}_1$ and $\tilde{x}_2$ in our SMDC network.

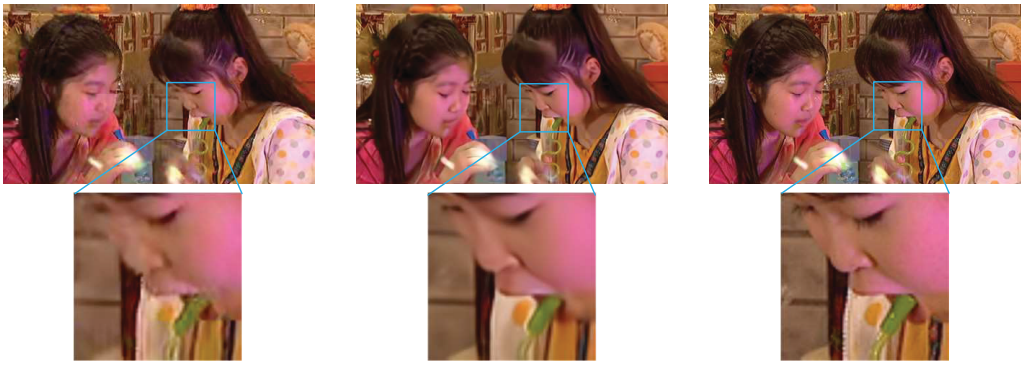
	$\lambda = 256$	$\lambda = 512$	$\lambda = 1024$	$\lambda = 2048$
PSNR of $\tilde{x}_1$	27.67	28.65	29.43	29.92
PSNR of $\tilde{x}_2$	26.42	27.44	28.22	28.58

In conclusion, the benefits of our SMDC network can be summarized as:

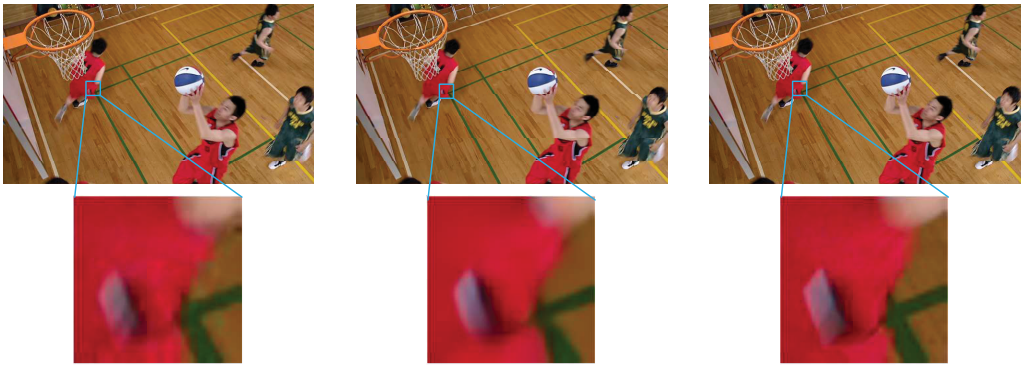
(1) As discussed in Section 4.3 of our paper, our SMDC network reduces the bit-rate for motion information, due to compressing a single motion map in SMDC.

(2) Our SMDC network generates $\tilde{x}_1$ with even higher quality than $\tilde{x}_2$, and thus leads to fewer residual between $\tilde{x}_1$ and x_1 to encode. This facilitates the residual compression subnet to achieve better compression performance.

Visual results. Then, we demonstrate more visual quality results of our PSNR and MS-SSIM models and the latest

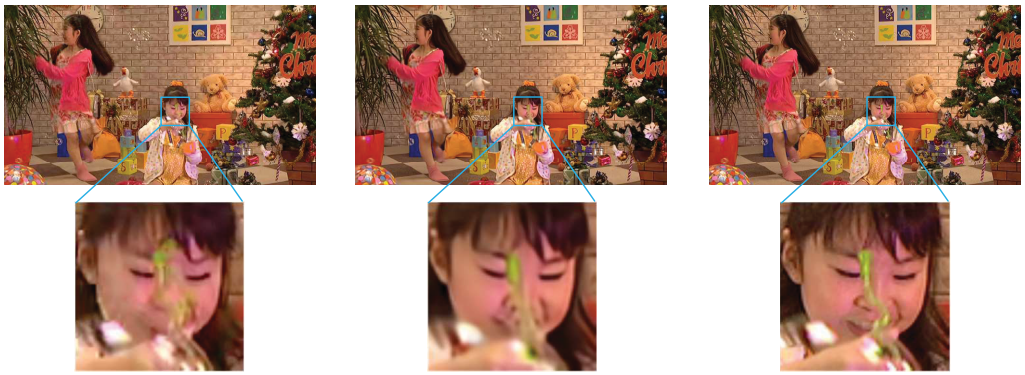


H.265, bpp = 0.1349 | Our PSNR model, bpp = 0.1134 | Raw



H.265, bpp = 0.2029 | Our PSNR model, bpp = 0.2023 | Raw

H.265, bpp = 0.2115 | Our MS-SSIM model, bpp = 0.2114 | Raw



H.265, bpp = 0.1558 | Our MS-SSIM model, bpp = 0.1401 | Raw

Figure 13. Visual results of our PSNR and MS-SSIM models in comparison with H.265.

video coding standard H.265 in Figure 13. The bit-rates in Figure 13 are the average values among all frames in each video, and the frames in Figure 13 are selected from layer 3, *i.e*., the lowest quality layer in our approach. It can be seen from Figure 13 that both our PSNR and MS-SSIM models have less compression artifacts than H.265, in case that our models consume lower bit-rate. That is, the frames in the lowest quality layer of our approach still achieve better visual quality, when the average bit-rate of the whole video is lower than H.265.

Different GOP sizes. Our method is able to adapt to different GOP sizes, since our BDDC and SMDC networks can be flexibly combined. In Figure 2, one more SMDC module can be inserted between the two SMDC modules, enlarging the GOP size to 12. More SMDC modules can be inserted to further enlarge the GOP size. In DVC [22], GOP = 12 is applied on the UVG dataset. For fairer comparison, we also test our HLVC approach on UVG with GOP = 12. Our PSNR performance on UVG drops to BDBR = 1.53% with the same anchor in Table 1, but we still outperform Wu *et al*. [38] and DVC [22].

Comparison with different configurations of x265. In our paper, we compare with the "LDP very fast" mode of x265. However, since our HLVC model has a "hierarchical B" structure, we further compare our approach with x265 configured with "b-adapt=0:bframes=9:b-pyramid=1" instead of "zerolatency". The detailed configuration is as follows.

```
ffmpeg -pix_fmt yuv420p -s WidthxHeight
-r Framerate -i Name.yuv -vframes
Frame -c:v libx265 -preset
veryfast/medium -x265-params
"b-adapt=0:bframes=9:b-pyramid=1:
crf=Quality:keyint=10:verbose=1" Name.mkv
```

With the anchor of x265 "hierarchical B", we achieve BDBR = -9.85% and -10.55% for the "medium" and "very fast" modes on MS-SSIM. For PSNR, we do not outperform x265 "Hierarchical B" (BDBR = 20.87%). Besides, we also compare our approach with the "LDP medium" mode of x265, where we obtain BDBR = -34.45% on MS-SSIM. In terms of PSNR, we are comparable with x265 "LDP medium" (BDBR = -1.08%).

Comparing ablation results with DVC [22]. DVC [22] has the IPPP prediction structure, while our approach employs the bi-directional hierarchical structure. To directly compare the performance of these two kinds of frame structure, we calculated the BDBR values of our ablation models with the anchor of DVC [22] on JCT-VC. The results of our "baseline+HQ" and "baseline+HQ+SM" models are -5.50% and -7.37% vs. DVC [22], respectively. It can be seen that our hierarchical model (+HQ) outperforms the IPPP structure of DVC (BDBR$<$0). This validates the effectiveness of our hierarchical layers.