Deep motion-compensation enhancement in video compression

N. Prette,[1,✉] D. Valsesia,[1] T. Bianchi,[1] E. Magli,[1]
M. Naccari,[3] and A. Fiandrotti[2]
[1] *Department of Electronics and Telecommuncations, Politecnico di Torino, Italy*
[2] *Università di Torino, Italy*
[3] *IEEE*
✉Email: nicola.prette@polito.it

This work introduces the multiframe motion-compensation enhancement network (MMCE-Net), a deep-learning tool aimed at improving the performance of current video coding standards based on motion-compensation, such as H.265/HEVC. The proposed method improves the inter-prediction coding efficiency by enhancing the accuracy of the motion-compensated frame and thereby improving the rate-distortion performance. MMCE-Net is a neural network that jointly exploits the predicted coding unit and two co-located coding units from previous reference frames to improve the estimation of the temporal evolution of the scene. This letter describes the architecture of MMCE-Net, how it is integrated into H.265/HEVC and the corresponding performance.

Introduction: The creation of better video compression techniques is crucial in today's media-saturated environment. In this letter, we propose a deep learning approach to enhance the rate-distortion performance of video compression schemes based on motion-compensation, such as H.265/HEVC.

In such schemes, compression is achieved by generating an estimate of the current frame through a process called motion-compensation. What is encoded then is only the residual signal that the motion-compensation algorithm has not been able to predict. However, the motion-compensation algorithm tends to generate only rough approximations of the target frame because it cannot model complex forms of motion. For this purpose, we designed a neural network that leverages the information present in previously coded frames to enhance the accuracy of the motion compensated frame (MC-frame). This leads to a reduction of the entropy of the prediction residuals and thus to better compression performance. We denote this methodology as MMCE-Net, which stands for multiframe motion-compensation enhancement network.

Related works: In the past few years, several works proposed the use of deep neural networks for video compression (a review on past work can be found in ref. [1]). Most of the work concerns the creation of compression tools, whose purpose is to aid the inner workings of some already established codec. Since their purpose is to improve upon an already established compression algorithm, this kind of technique tends to tackle narrower sub-procedures of the task of video compression, which are easier to work on and allow to significantly gain over the performance of the established codecs. These techniques can be categorised depending on the particular step of the video compression algorithm being tackled.

A significant portion of the literature has addressed inter-prediction, which aims to exploit the temporal correlation with other frames in the sequence. Since temporally consecutive frames are very similar to each other, it is possible to reasonably estimate the frame to be encoded by taking neighbouring frames (called reference frames) as a basis. This is one of the greatest sources of bit-rate reduction in the final compressed file, and finding ways to improve the mechanism through which this is executed is fundamental in the quest for evolving modern video coding techniques. Zhao et al. in ref. [2] propose the construction of enhanced "virtual reference frames" to replace regular ones and thus improve the accuracy of the prediction. This particular technique was studied for bi-directional prediction and constructs the reference frame by interpolating between a future and a past frame using the FRUC (frame rate up-conversion) network deep voxel flow [3]. The obtained result is further enhanced using a residual CNN (convolutional neural network). Along the same lines, Lee et al. [4] also generate a virtual reference frame, but their method can work with past frames only, so the reference frame is not interpolated but predicted.

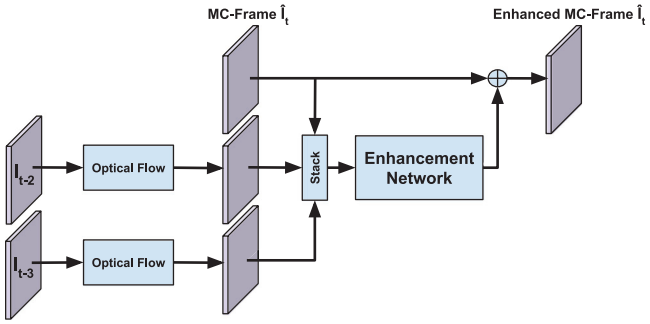

Fig. 1 *MMCE-Net structure: two preceding decoded frames are fed to an optical flow network so they can be warped to match the MC-frame. The warped frames are concatenated to the MC-frame and then merged. The final output is an enhanced MC-frame*

Methodology: Usually, in block-based motion-compensated predictive codecs, the MC-frame is generated by assembling sections of previously encoded frames (called reference frames). Besides a limited capability to model complex motion, this method has also the disadvantage of generating estimates which contain square-shaped artefacts, corresponding to the places where the blocks are stitched together.

MMCE-Net aims to generate a more realistic estimate of the MC-frame by exploiting the information contained in preceding frames (in display order). The reasoning is that by observing approximations of the original sequence at different time instants, MMCE-Net can learn complex motion patterns and use them to remove the artefacts in the MC-frame.

An overview of the proposed approach is shown in Figure 1. The core idea is to use a CNN that takes as inputs the MC-frame together with two temporally preceding decoded frames (I_{t-2}, I_{t-3}) to generate an enhanced MC-frame denoted as $\hat{I}_t$. I_{t-2}, I_{t-3} are not used directly: to help the work of the enhancement network, both are first warped according to the estimated optical flow between each of them and the MC-frame. This operation serves the purpose of roughly registering the information of the past frames to the content at time t so that the de-artefacting network does not also need to learn complex motion models. The previous frames are warped by re-sampling on the grid provided by the flow estimate with a suitable interpolation kernel (e.g. bicubic). We employ a neural network approach to optical-flow estimation, in the form of PWC-Net [5].

At this point, the MC-frame needs to be fused with the two warped frames. This is achieved with a simple concatenation over the colour channel axis. By stacking the images in this manner, the information is merged internally by the layers of the enhancement network, in a process which is known as direct fusion.

The input tensor is then fed into the main component of this architecture: the enhancement network. For the task of enhancement, neural network designs from the image restoration literature can be used. In particular, we employ a Dn-CNN architecture [6] as a baseline. This CNN adopts the residual learning strategy, by having a structure composed of repeated blocks of convolution, batch normalisation, and ReLU non-linearity with skip connections. The spatial resolution of the frames is never changed across the layers to avoid any loss of high-frequency information. Notice that the restoration network does not directly estimate the enhanced frame, but rather an additive correction applied to the MC-frame.

Intergration inside the H.265/HEVC reference model: MMCE-Net was implemented inside the currently ubiquitous H.265/HEVC standard, using its reference implementation HM (version 16.22). We chose this standard over the current state-of-the-art H.266/VVC because we have more experience working with the H.265/HEVC reference implementation. In any case, the method itself is applicable on every compression algorithm which employs motion-compensation.

During coding H.265/HEVC partitions the frame in square portions called coding tree units (CTU), which are further subdivided using a quad-tree structure in what are denominated coding units (CU). Since

H.265/HEVC processes the CUs sequentially (in the sense that for every CU the residual is evaluated before processing the following CU), also MMCE-Net was employed at the CU level. Furthermore, H.265/HEVC decides whether to use intra or inter prediction on every CU. We determined that most of the gain is experienced when the network is applied on the largest possible dimension for the CUs (64×64), so we avoided applying MMCE-Net on smaller ones. However, since the network is convolutional it can theoretically be applied to CUs of every dimension. Another choice was to use MMCE-Net only on the Y channel during processing since applying it on the chrominance channels also gave minor performance improvements. The enhancement network was integrated into HM as an optional mode that the compression algorithm may choose to use based on rate-distortion considerations. The chosen discriminant was the variance of the residual between the MC-frame and the original one before and after the enhancement. The integration of MMCE-Net as an additional mode requires the inclusion of a signalling procedure to enable the decoding of the compressed file. This involves adding a binary flag at the CU level to indicate whether the enhancement was used or not on the current CU. This flag is encoded using Context-adaptive binary arithmetic coding (CABAC) compression with a single context to model the probability distribution for the input data.

Training and dataset: MMCE-Net was trained using the VIMEO 90K dataset [7]. It consists of 90,000 video sequences extracted from the Vimeo video streaming website comprising seven frames with a resolution of 448×256. This dataset was chosen for the great variety in the scenes depicted and the moderate expensiveness in terms of memory. For each sequence, the images were converted to the YCbCr colour space using the ITU-R BT.709 colour primaries and concatenated to create a raw video sequence. The generated video sequences were then compressed using an optimised implementation of the H.265/HEVC standard called x265, setting a constant Quantisation Parameter (QP) level and the use of only unidirectional prediction (i.e. P-frames only). All the other parameters were set to the default values of $\times 265$. Two datasets were created: one using $QP = 22$, the second obtained randomly applying one of the three standard values $\{27; 32; 37\}$. Then the reconstructed and motion-compensated frames were extracted in the form of raw images again in the YCbCr colour space. The reconstructed file was extracted using the FFmpeg software while the MC-frame was extracted using a video coding analysis software. After completing the procedure, for every scene of the dataset we were left with the original frame I_t (used as ground truth), the corresponding MC-frame, and the two temporally preceding reconstructed frames ($\hat{I}_{t-2}$ and $\hat{I}_{t-3}$) to complete the network's input. The network was first trained on the bigger dataset obtained using only $QP = 22$ and was then later fine-tuned using the dataset obtained using multiple QPs. In this way, the final network is capable of generalising for all the levels of distortion, which is much more convenient than creating different networks specialised for particular ranges of QP.

The enhancement network was trained by minimising an L2 loss between the enhanced MC-frame and the label using Adam optimisation algorithm with a learning rate that was set to 10^{-4}. The hardware setup consisted of an NVIDIA Tesla V100 SXM2 GPU. The training lasted 10^6 iterations using batches of 16 sequences cropped to the dimension of 256×256 as the implementation of PWC-Net needs inputs with side dimensions which are multiple of 64, and since the network is intended to be applied on square CU extracted during H.265/HEVC coding, we preferred using square images as input.

Performance analysis: The tests were carried out on common testing sequences provided by the joint video experts team (JVET). Since the network was trained to work only in the causal scenario (no information from temporally future frames is employed for the enhancement) in this work HM is configured according to the low-delay P configuration, which only employs P-frames.

In Table 1, we compare the gains of MMCE-Net in terms of rate-distortion performance of the enhanced scheme compared with the HM version 16.22 implementation of H.265/HEVC using the Bjøntegaard Delta metric on the rate (BD-Rate).

As can be observed, there is high variability in the amount of gain provided by MMCE-Net depending on the content of the sequence. Most of the gain is seen in the B and E Class of the JVET test sequences,

Table 1. BD-rate relative to HM-16.2 using low-delay P configuration

Class	Sequence	fps	BD-rate (%)
A - 2560 × 1600	PeopleOnStreet	30	−0.37
	Traffic		−1.75
B - 1920 × 1080	BQTerrace	60	−7.49
	BasketballDrive	50	−1.47
	Cactus		−1.22
	Kimono	24	−2.75
	ParkScene		−0.67
C - 832 × 480	BQMall	60	−1.28
	BasketballDrill	50	−0.44
	PartyScene		−0.71
	RaceHorsesC	30	−0.60
D - 416 × 240	BQSquare	60	−1.08
	BasketballPass	50	−0.12
	BlowingBubbles		−0.38
	RaceHorses	30	−0.12
E - 1280 × 720	FourPeople	60	−1.04
	Johnny		−6.08
	KristenAndSara		−2.77
Total average			−1.69

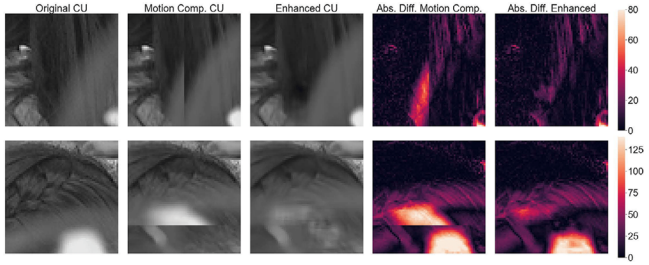

Fig. 2 *Examples of the results of the enhancement on* 64×64 *CU performed by MMCE-Net. The network is capable of reducing the blockiness present in motion-compensated frames, which is one of its greatest sources of approximation error*

where there are peaks above the absolute value of 2.5% (e.g. BQTerrace, −7.49%, Johnny, −6.08%). On all other classes the gains are smaller but even in the most difficult clips (e.g. PeopleOnStreet, BasketballPass) the technique always outperforms the baseline HM. It is interesting to note that, on average, most of the reduction in terms of rate happens when the level of distortion is low (QP Levels = 22, 27), while the gap decreases in situations of lower rate and higher distortion. For Class B, on average, we see a 4% reduction in terms of rate for $QP = 22$, while for $QP = 37$ it reduces to 2%. A more extreme version of this phenomenon happens for Class E, where we see a reduction from 7% for $QP = 22$ to less than 1% for $QP = 37$.

We also show visual results of the enhanced frames yielded by MMCE-Net. As can be observed in Figure 2, the greatest source of performance improvement are cases in which the CU presents artefacts due to the composition of different blocks. The network is capable of localising the regions interested by such artefacts and minimising the magnitude of the residual by either blurring the area or, when possible, by providing a more likely estimate using the spatial and temporal context of the sequence.

It is interesting to observe that the sequences with the highest gains seem to have high resolution, while the worst performance is seen on the fast-moving sequences at low resolution. This behaviour can be justified by the fact that at higher resolution CU of size of 64×64 are chosen much more often, so there are more occasions in which the

Table 2. BD-rate (%) comparison for ablation study

Class	Sequence	MMCE-Net	Single frame	Single warped frame	Multiframe no warping	Enhancement always on
B - 1920 × 1080	ParkScene	**−0.67**	0.00	−0.04	−0.07	−0.36
	Kimono	**−2.75**	−2.00	−1.41	−1.65	−2.19
C - 832 × 480	BQMall	**−1.28**	−0.16	−1.13	−0.27	−1.22
	PartyScene	−0.71	−0.43	−0.74	**−0.77**	−0.61
D - 416 × 240	BQSquare	−1.08	−0.28	−1.30	**−1.40**	−1.04
	BlowingBubbles	−0.38	−0.10	−0.37	**−0.69**	−0.34
E - 1280 × 720	Johnny	**−6.08**	−1.57	−2.73	−3.02	−5.60
Complexity (number of parameters)		14.79×10^6	0.71×10^6	14.78×10^6	0.71×10^6	14.79×10^6

network can introduce improvements. This also explains the lower gains for the classes C and D. Similar considerations can be done for rapidly changing sequences, since again on such sequences 64 × 64 CU are chosen more sparingly. It has also to be considered that when drastic changes between frames occur, the preceding frames are less helpful at providing support for the enhancement since they are less correlated with the CU that is being enhanced.

Ablation tests: We performed some ablation tests to justify the various design choices of the network. These tests were done only on a subset of the JVET sequences for reasons of brevity. The results of the ablation tests are reported in Table 2. "single frame" refers to the performances achieved by not employing previous frames and feeding only the MC-Frame to the enhancement network. "single warped frame" refers to the use of only one reference frame instead of two (in particular we used I_{t-2}), while in "multiframe no warping" we reinserted the second reference frame but removed the registration mechanism with the optical flow. The last column, "enhancement always on" refers to the performances achieved when MMCE-Net is applied on all CUs. Using this direct implementation also makes it unnecessary to signal whether the enhancement is applied or not to each block, at the cost of applying MMCE-Net also when it is not beneficial.

MMCE-Net outperforms the ablation almost always, with considerable gains at high-resolution. The gap get smaller for the classes C and D, which contain low-resolution videos, and for the sequences PartyScene, BQSquare and BlowingBubbles the "multiframe no warping" configuration slightly outperforms MMCE-Net. In general, the system introduces less improvement on small-resolution images because on these resolutions 64×64 CUs are used seldom. Furthermore, at small resolutions moving objects have dimensions comparable to the size of the CU. This leads to blocks extracted from previous frames that are very different from the MC-frame and thus the registration tends to generate warped frames which are very distorted and artificial-looking. These slight problems could be solved by creating resolution-specific networks.

Another notable fact is that using the "enhancement always on" does not massively alter the performance, with the biggest gap seen for the sequence Kimono, indicating that MMCE-Net introduces an improvement on almost every CU on which is employed.

In the final row, we show the complexity of MMCE-Net and its ablations measuring the number of parameters. The complexity of our method is comparable to most deep learning techniques in the field of video compression. It can be observed that most of the parameters originate from the use of PWC-Net, while the enhancement network is relatively small.

This is further illustrated by the fact that, for QP=22, on average the network improves the accuracy of the motion-compensated CU on 80% of the blocks where it is applied (here are some examples of success rates on the tested sequences: ParkScene 83.26%, Kimono 77.39%, BQMall 85.74%, PartyScene 84.18%, BQSquare 66.62%, BlowingBubbles 89.53%, Johnny 73.33%).

Conclusion: We introduced MMCE-Net, a deep neural network methodology to improve the performance of block-based motion-compensation algorithms commonly used in video compression. We showed how the use of a multiframe approach complemented with an optical flow for the registration of previous frames is capable of greatly boosting the enhancing capability of the Dn-CNN network employed for this purpose. While it was tested for H.265/HEVC, the proposed technique is not standard-dependent and could be easily applied to standards such as H.264/AVC and H.266/VVC.

This technique is capable of improving the performance across all the sequences where it was tested, achieving the greatest gains in the B and E classes of the JVET test sequences. These results are very encouraging and future work will focus on investigating systems in which the motion-compensation algorithm is integrated into the enhancement architecture. In this way it would be possible, for example, to reduce the costs of side information by using a simpler version of the compensation algorithm, while keeping constant the entropy of the residual using the enhancement network.

Acknowledgement: This research was supported by RAI—Radiotelevisione Italiana and the SmartData@PoliTO center. The authors also want to thank the developers from Elecard for letting us use the video coding analysis suite StreamEye for this work.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: Author elects to not share data: Research data are not shared.

Received: *24 December 2021* Accepted: *2 March 2022*
doi: 10.1049/ell2.12475

References

1 Ma, S., Zhang, X., Jia, C., Zhao, Z., Wang, S., Wang, S.: Image and video compression with neural networks: a review. *IEEE Trans. Circuits Syst. Video Technol.* **30**(6), 1683–1698 (2020)
2 Zhao, L., Wang, S., Zhang, X., Wang, S., Ma, S., Gao, W.: Enhanced motion-compensated video coding with deep virtual reference frame generation. *IEEE Trans. Image Process.* **28**(10), 4832–4844 (2019)
3 Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4473–4481. IEEE, Piscataway, NJ (2017)
4 Lee, J.K., Kim, N., Cho, S., Kang, J.W.: Deep video prediction network-based inter-frame coding in HEVC. *IEEE Access* **8**, 95906–95917 (2020)
5 Sun, D., Yang, X., Liu, M., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943. IEEE, Piscataway, NJ (2018)
6 Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **26**(7), 3142–3155 (2017)
7 Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. *Int. J. Comput. Vision* **127**(8), 1106–1125 (2019)