Deep Multiframe Enhancement for Motion Prediction in Video Compression

Deep Multiframe Enhancement for Motion Prediction in Video Compression

1st Nicola Prette
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
nicola.prette@polito.it

2nd Diego Valsesia
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
diego.valsesia@polito.it

3rd Tiziano Bianchi
Department of Electronics and Telecommunications
Politecnico di Torino
Turin, Italy
tiziano.bianchi@polito.it

***Abstract*—This work proposes a novel Deep Learning technique to increase the efficiency of currently available video compression techniques based on motion compensation. The goal is to improve the frame prediction task, whereby a more accurate prediction of the motion from the reference frames to the target frame allows to reduce the rate needed to encode the residual. This is achieved by means of a convolutional neural network (CNN) architecture that processes the basic block-based motion-compensated prediction of the current frame as well as predictions from past reference frames. This method allows to reduce typical artifacts such as blockiness, and achieves a more accurate prediction of motion thanks to the representation capabilities of CNNs, leading to smaller prediction residuals. Preliminary results show that the proposed approach is capable of providing BD-rate gains up to 6%.**

***Index Terms*—Frame Enhancement, Video compression, Deep learning**

I. Introduction

Advancing video compression algorithms towards higher efficiency is an important task, as the quantity and resolution of videos that are created, stored, and shared dramatically increases every year. The last decade has shown how techniques based on deep learning can successfully tackle challenging problems in image and video processing, so it only seems natural to apply them to the field of video compression.

In this paper we are concerned with a deep learning approach to enhance the rate-distortion performance of video compression schemes, such as current state-of-the-art HEVC (high efficiency video coding [1]), based on the motion compensation paradigm to exploit temporal redundancy. In this coding paradigm compression is achieved by generating a prediction of the current frame based on past decoded frames and only encoding the prediction error, i.e., the residual signal that the motion compensation algorithm has not been able to predict. The decoder is able to generate the predicted frame from the motion vectors written in the compressed file, which are used to signal block displacements with respect to the reference frames. However, these generated motion-compensated (MC) frames tend to be a rough approximation of the frame to be compressed because of the limited descriptive power of modeling motion by means of block-wise translation.

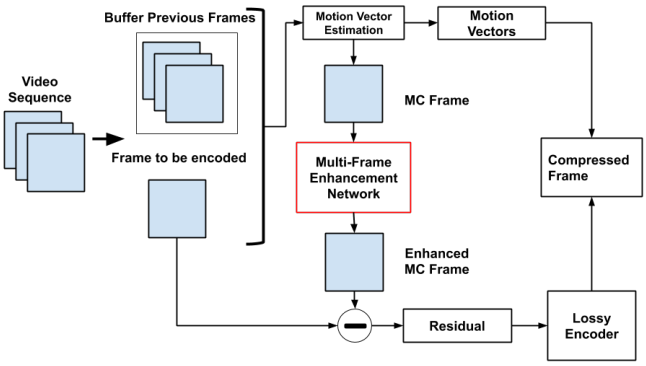

Fig. 1. Proposed compression scheme: a neural network uses the motion compensated frame and past decoded frames to improve the quality of the prediction and generate a residual which is easier to compress.

This work proposes a framework based on CNNs to improve the quality of frame prediction, overviewed in Fig. 1. The role of the CNN is to provide a causal model of motion, with the objective to improve motion prediction without using additional side information. In particular, the proposed scheme uses the traditional MC frame at time t and decoded frames at past times to generate a first set of predictions for time t, which are then merged and refined. The result is a better prediction for the frame at time t thanks to the improved motion model learned by the CNN, which allows us to reduce the entropy of the prediction residual, and in turn the rate of the compressed file at constant quality, at no additional cost in terms of side information. The frame is reconstructed at the decoder side

This work was supported by RAI – Radiotelevisione italiana and the SmartData@PoliTO center.

following the same steps: first the MC frame is generated using the motion vectors, then its quality is enhanced using the neural network. Finally the residual is added to the enhanced frame. We remark that the proposed scheme differs from other enhancement schemes present in the literature as it uses a multi-frame architecture which targets the enhancement of the predicted frame before encoding of the residual. It should be regarded as an improvement of the motion estimation stage instead of a neural network version of in-loop filtering or post-processing of the decoded video.

II. Background

In the last few years several works used deep neural networks to tackle different aspects of video compression. A comprehensive review of what has been done throughout the years can be found in [2]. The contributions present in the literature can be broadly categorized as i) end-to-end compression schemes and ii) neural network video compression tools.

The first approach tries to design an entire compression algorithm from scratch, using fully differentiable and optimizable pipelines. While being an interesting effort, it cannot yet compete with the highly optimized algorithms employed in current video coding standards. Some promising examples include [3] and [4] in which an end-to-end deep learning technique is described to reach performance comparable to H.264/AVC. [5] recreates the structure of the hybrid coding architecture using multiple neural networks and manages to reach the performances of HEVC in terms of MS-SSIM.

Most of the research in this field falls under the latter approach, i.e., it is concerned with the development of tools to improve the performance of tasks already performed in video coding algorithms. Works in this category have a greater impact as a narrower task can be solved more effectively. The works in this field can be categorized based on which part of the video compression pipeline is addressed. Several papers deal with improving the in-loop filter, such as [6] which uses a multi-frame approach. Other approaches focus on the Intra-prediction mode and capitalize on the advancements deep learning has done in image compression [7] [8]. Some other works tackle several aspects of motion compensation. Zhang et al. [9] propose a CNN to improve the performance of the fractional pixel interpolation process, while Zhao et al. [10] use a CNN for the purpose of improving the way multiple motion-compensated estimates are merged in the bidirectional prediction process. Two papers that are close to our approach are [11], in which Huo et al. propose a CNN to refine the quality of inter-predicted blocks using already reconstructed parts of the frame as an aid, and [12], in which Choi et al. introduce a new inter prediction mode which tries to predict a frame directly without the use of side information. The technique proposed in this paper can be regarded as a generalization of [12], where we are also able to exploit side information in the form of motion vectors of past MC frames. Indeed, we can tune the desired tradeoff between no side information as in [12] and highly accurate motion vectors; this latter case is studied in detail in this paper. Moreover, we extend the approach in [11] by exploiting the temporal correlation in multiple previously decoded frames.

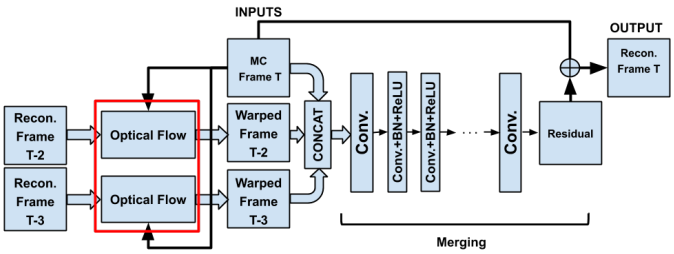

Fig. 2. Multiframe Enchancement Network structure: two preceding decoded frames are fed to an optical flow network so they can be warped to match the MC frame. The warped frames are concatenated to the MC frame and then merged. The final output is an enhanced MC frame.

III. Proposed method

This work proposes an approach to improve the quality of motion estimation with the goal of reducing the entropy of the prediction residual. The basic architecture was inspired by some recent developments in video quality enhancement using multi-image deep neural network schemes (such as [13], [14] and [15]). The critical difference is that our approach is applied as a frame prediction method, thus increasing compression efficiency rather than enhancing the quality of decoded frames. Using this method it would be possible to set the precision of the motion vectors as a tunable parameter and allow in some cases to greatly reduce the amount of side information without increasing the rate of the residual.

The main idea is to use a CNN to model more accurately the motion between past frames and the frame to be predicted. While a fully causal prediction, such as the one in [12], would not require any side information in the form of motion vectors, we argue that a more general approach is to combine side information and motion modeling by means of CNNs. In particular, as shown in Fig. 2, we rely on the MC frame P_t produced using motion vectors from decoded frame $\hat{I}_{t-1}$ as a rough estimate of the current frame I_t. We then use a CNN with the task of estimating the optical flow between past decoded frames $\hat{I}_{t-2}$, $\hat{I}_{t-3}$ and the MC frame P_t and warping $\hat{I}_{t-2}$, $\hat{I}_{t-3}$ accordingly using a spatial transform layer [16]. This operation produces multiple estimates of the frame at time t and registers the information at past times to the content at time t. The final part of the proposed architecture merges these multiple estimates to produce the final predicted frame. It does so in a residual fashion, i.e., only estimating the correction to the MC frame P_t.

Notice that we do not propose to warp decoded frame I_{t-1} by means of the optical flow network. The reason is that this operation would not be informative as the MC frame P_t is already obtained by block translation from I_{t-1}. We also remark that, while the architecture presented in Fig. 2 only uses past frames for ease of explanation, the method can be

easily extended to use bidirectional prediction from past and future frames.

The optical flow network is based on PWC-Net [17], a state-of-the-art CNN for this task based on a coarse to fine approach. The warping layer spatially transforms the two auxiliary frames by re-sampling these images using the optical flow as sampling grid. In each point indicated a sampling kernel is applied (in our case a cubic interpolator) and the output is mapped to the pixel corresponding the sampling point.

The network producing the final estimate is based on the DnCNN denoising and restoration architecture [18]. Its input is a 9-channel image obtained by stacking the MC frames and the warped frames. Notice that by stacking the input, the first convolutional layers of the network will already merge the information from all of them. In the field of convolutional spatio-temporal networks this process is referred as direct fusion. The DnCNN architecture is composed of a number of blocks of 2D convolution, batch normalization and ReLU non-linearity. The output is a residual correction to the MC frame.

IV. Experimental results

A. Experimental setup

The training set was built based on the VIMEO 90K dataset, a large raw video dataset assembled in [15] to be used for several video processing tasks such as denoising and super-resolution. The dataset is composed by 90,000 7-frame sequences with resolution 448×256. These sequences were first compressed with HEVC using x265 default settings, a constant quantization parameter (QP) and forcing no B-frames, to match the setup introduced in Sec. III.

To mimic the scenario in which the motion compensation is created by HEVC, the dataset is made of triplets composed of the MC Frame P_t, and the decompressed frames at times $\hat{I}_{t-2}$, $\hat{I}_{t-3}$.

A publicly available pretrained implementation [19] of PWC-Net was used. In particular, the PWCNET-LG-6-2-MULTISTEPS-CHAIRSTHINGSMIX model was chosen as it has better performance, albeit requiring more memory. Its weights were kept frozen to the pretrained values for the following experiments.

The DnCNN was trained from random initialization. Different setups is terms of depth of the network were tested obtaining the best results using 20 blocks of 2D convolution.

The network was trained on an NVIDIA Tesla V100 SXM2 GPUs using Adam optimization [20] to minimize the L2 loss between the enhanced MC Frame and the original frame with a learning rate of 10^{-3} for 10^6 iterations. Batches of 16 sequences, cropped to 256×256 pixels were used. The network was first trained using QP value equal to 22 and then fine-tuned using the same sequences quantized using QP $= \{27, 32, 37\}$ so that the network generalizes to different levels of quantization noise instead of being specialized on a particular one.

B. Performance evaluation

We present a set of preliminary experiments aimed at characterizing the performance of the proposed approach, i.e., whether the enhanced motion prediction enables a reduction of the entropy of the prediction residual with respect to the residual generated by the HEVC motion prediction. Due to the preliminary nature of our experiments, the proposed method is not fully integrated into the HEVC codec and we do not have access to the inter-prediction residual encoder. However, it is possible to reliably assess potential gains via a proxy, where JPEG is used to encode the prediction residuals between the (enhanced) MC frames and the original uncompressed frame instead of the HEVC encoder. Before compression, an offset was added to the residual to make all the values positive. We argue that JPEG is a reliable proxy as compresses the residual following similar steps to the ones employed by HEVC, e.g., block-based quantization of the DCT coefficients. Moreover, we are only concerned with relative differences between the proposed method and the baseline technique, rather than absolute rate-distortion values, so any systematic discrepancy of the proposed proxy with respect to the HECV inter encoder has limited impact. The rates were compared for different QP values. To more realistically simulate the performance of HEVC, the residual images were compressed using quality values for the JPEG compression that approximately matched the level of distortion on the reconstructed frame generated by the respective QP values. The residuals were compressed using the *mjpeg* library inside the *ffmpeg* software using the following quality factors q $= \{4, 7, 10, 20\}$ to match QP $= \{22, 27, 32, 37\}$ respectively. The PSNR is not constant at fixed q though, and this is taken into account in the rate-distortion curves and computation of BD rate.

TABLE I
NETWORK PERFORMANCE ON JVET SEQUENCES:

Sequence Name	**BD Rate** [%]
BasketballDrive	-5.83
ParkScene	-4.14
Kimono1	-4.12
Cactus	-4.89
BQTerrace	-6.19

The test set is represented by common test sequences (CTS [21]) provided by JVET. In order to make a fair comparison, since the VIMEO 90K was built mostly from 720p and 1080p resolution videos, several 1080p sequences from the CTS were selected: BasketballDrive, Kimono1, ParkScene, Cactus, BQTerrace. In the same way as done for the training set, the sequences were compressed using the x265 library with default settings, forcing no B-frames and using QP $= \{22, 27, 32, 37\}$. The frames input to the optical flow network were zero-padded from the resolution 1920x1080 to the resolution to 1920x1088

in order to make the dimensions multiple of 64 and adapt them to the architecture of PWC-Net.

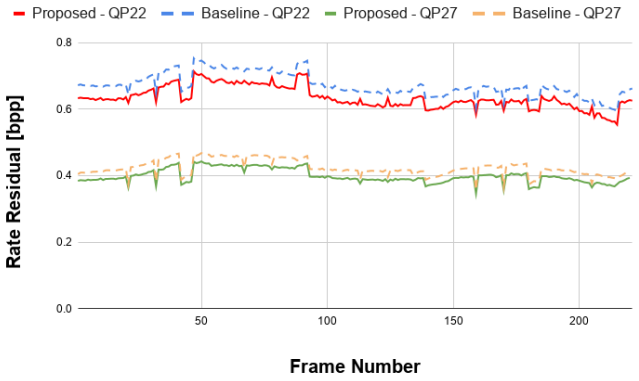

Fig. 3. Compression Rate for each frame before and after the enhancement on the ParkScene sequence. There is a rate reduction for all frames, but it gets smaller as the QP increases.

The PSNR was computed between the residual before and after the compression, and averaged across all frames of the sequences, again as a proxy to the actual quality of the reconstructed frame. The rate reduction was measured for every QP level and averaged over the whole rate-distortion curve using the Bjontegaard metric. The results of the test are shown in Table I. We can notice that the proposed approach provides up to 6.19% BD-rate reduction of the residual when compared to the traditional prediction based on block motion estimation and compensation performed by HEVC. Moreover, the reduction in rate is always accompanied by an improvement in terms of distortion on all sequences and at all the quantization levels we tested, as shown by the rate-distortion curves in Fig. 5.

TABLE II
ABLATION: NETWORK PERFORMANCE WITHOUT ADDING THE WARPED FRAMES:

Sequence Name	**BD Rate** [%]
BasketballDrive	-2.96
ParkScene	-0.54
Kimono1	-2.15
Cactus	-1.39
BQTerrace	-1.44

For comparison we tried to evaluate the performance of a modified version of the network where the optical flow part is removed and the DnCNN is applied only on the motion-compensated frame. The BD-rate was evaluated between the original frames and the enhanced ones with the modified technique. The results seen in Table II show a great reduction in terms of rate gain, which highlights the importance of the multi-frame approach for the effectiveness of the proposed technique.

One interesting characteristic of the proposed approach is that it produces an improvement in terms of bit-rate which is very consistent across all frames of a sequence, as shown in Fig. 3. This can be attributed to the fact that the network manages to almost completely remove block-shaped artifacts that characterize standard MC frames, while providing a good estimate of motion (as it can be seen in Fig. 4). The figure also shows that gains tends to diminish as the QP level increases. This is due to the increased difficulty in estimating motion when the amount of distortion increases. We also remark that the network was trained primarily over frames compressed with $QP = 22$, and only finetuned for the other QP values. It is reasonable to expect a small performance improvement by training each quantization level from random initialization. On the other hand, improved training methods that are blind to the QP level are also an interesting future direction, as they would allow deployment of more compact models.

V. CONCLUSIONS

This paper described an approach to enhance motion estimation inside video codecs based on motion compensation, improving their compression performance. By using CNNs aided by the MC frame to improve the prediction model, we are able to significantly reduce the coding rate of the prediction residuals.

Preliminary results obtained on the HEVC codec are very encouraging and future work will focus on exploring the tradeoffs between the accuracy and cost of motion vectors and the effectiveness of the proposed scheme, as well as refining the network architectures, making it robust to multiple QPs and fully integrating it into HEVC.

REFERENCES

[1] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2020.

[3] Z. Chen, T. He, X. Jin, and F. Wu, "Learning for video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 566–576, 2020.

[4] T. Chen, H. Liu, Q. Shen, T. Yue, X. Cao, and Z. Ma, "Deepcoder: A deep neural network based video compression," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4.

[5] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao, "DVC: An End-To-End Deep Video Compression Framework," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang, and Z. Guan, "A Deep Learning Approach for Multi-Frame In-Loop Filter of HEVC," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5663–5678, 2019.

[7] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3236–3247, 2018.

[8] Y. Hu, W. Yang, S. Xia, W. Cheng, and J. Liu, "Enhanced intra prediction with recurrent neural network in video coding," in *2018 Data Compression Conference*, 2018, pp. 413–413.

[9] N. Yan, D. Liu, H. Li, B. Li, L. Li, and F. Wu, "Convolutional neural network-based fractional-pixel motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 840–853, 2019.

Fig. 4. Frame comparison, from left to right: original frame, HEVC MC frame, enhanced MC frame, residual using HEVC MC frame, residual using enhanced frame. The enhancement scheme manages to remove most of the block-shaped artifacts present in the HEVC MC frames.

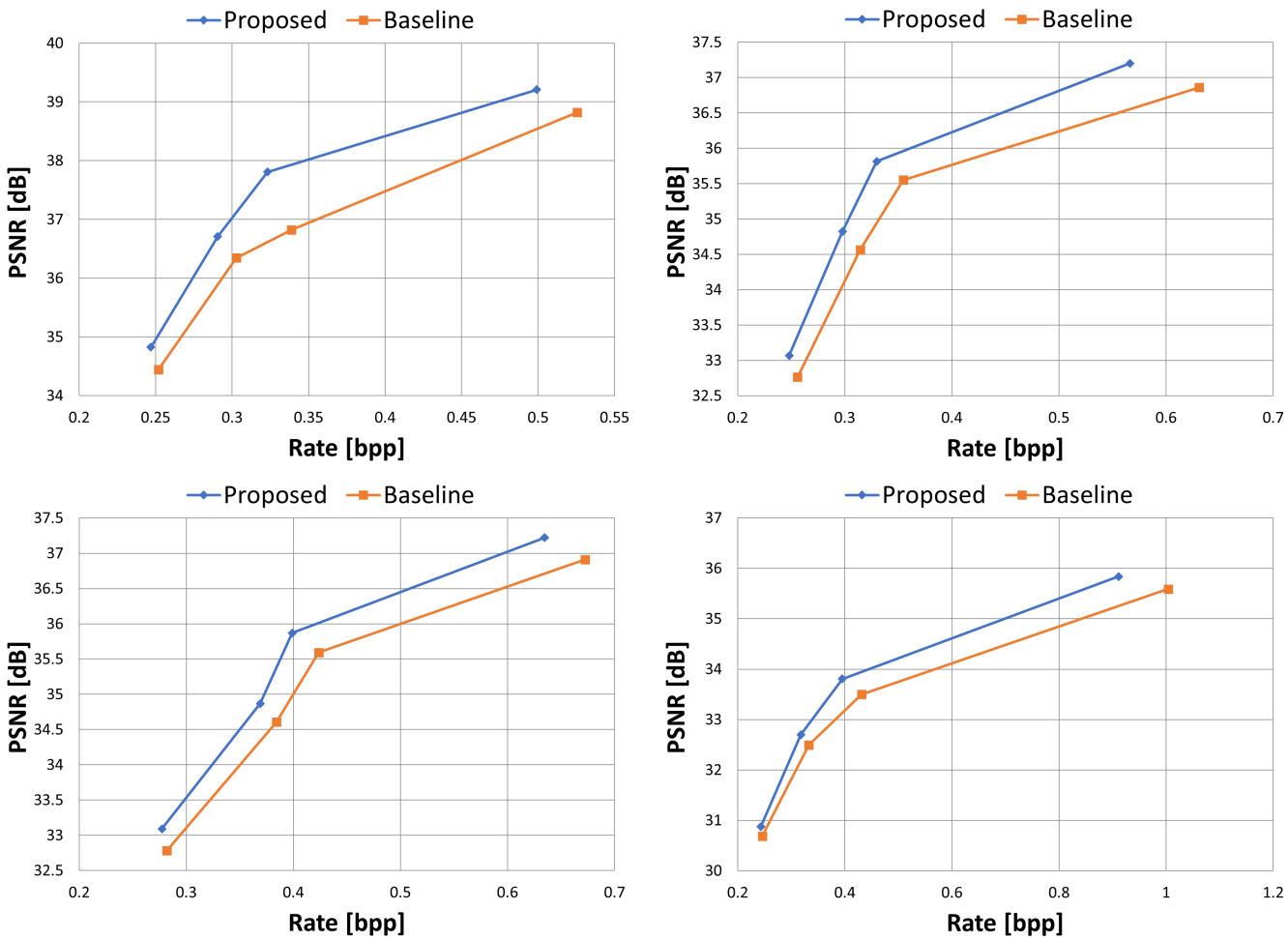

Fig. 5. Rate-distortion curves on the test sequences Kimono (top left), BasketBallDrive (top right), ParkScene (bottom left), BQTerrace (bottom right).

[10] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, and J. Yang, "CNN-Based Bi-Directional Motion Compensation for High Efficiency Video Coding," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–4.

[11] S. Huo, D. Liu, F. Wu, and H. Li, "Convolutional neural network-based motion compensation refinement for video coding," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–4.

[12] H. Choi and I. V. Bajić, "Deep frame prediction for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[13] J. Tong, X. Wu, D. Ding, Z. Zhu, and Z. Liu, "Learning-based multi-frame video quality enhancement," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 929–933.

[14] R. Yang, M. Xu, Z. Wang, and T. Li, "Multi-frame quality enhancement for compressed video," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6664–6673.

[15] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision (IJCV)*, vol. 127, no. 8, pp. 1106–1125, 2019.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 2017–2025. Curran Associates, Inc., 2015.

[17] D. Sun, X. Yang, M. Liu, and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[18] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[19] Phil Ferriere, "Optical Flow Prediction with Tensorflow," https://github.com/philferriere/tfoptflow.

[20] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[21] Frank Bossen et al., "Common test conditions and software reference configurations," *JCTVC-L1100*, vol. 12, pp. 7, 2013.