JMPNET: JOINT MOTION PREDICTION FOR LEARNING-BASED VIDEO COMPRESSION

Dongyang Li, Zhenhong Sun, Zhiyu Tan, Xiuyu Sun, Fangyi Zhang, Yichen Qian, Hao Li*

Alibaba Group, China

ABSTRACT

In recent years, more attention is attracted by learning-based approaches in the field of video compression. Recent methods of this kind normally consist of three major components: intra-frame network, motion prediction network, and residual network, among which the motion prediction part is particularly critical for video compression. Benefiting from the optical flow which enables dense motion prediction, recent methods have shown competitive performance compared with traditional codecs. However, problems such as tail shadow and background distortion in the predicted frame remain unsolved. To tackle these problems, JMPNet is introduced in this paper to provide more accurate motion information by using both optical flow and dynamic local filter as well as an attention map to further fuse these motion information in a smarter way. Experimental results show that the proposed method surpasses state-of-the-art (SOTA) rate-distortion (RD) performance in the most data-sets.

***Index Terms*—** deep learning, video compression, joint motion prediction, optical flow, dynamic local filter

1. INTRODUCTION

With the development of deep neural networks, learning-based video compression approaches have shown competitive performance compared with traditional codecs [1, 2, 3]. Existing works first try compressing video sequences with the help from implicit motion information either contained in a frame interpolation network [4] or 3D auto-encoder convolutional layers [5, 6]. These methods show promising video compression performance, but the temporal redundancy may not be effectively removed for the lack of explicit motion information. To improve the performance of motion prediction, explicit motion information is introduced into the framework in the form of optical flow or its variants [7, 8, 9, 10]. In the most recent work, the researcher proposed a feature-space video coding network which performs all major operations in the feature space [11].

However, optical flow based methods have problems such as tail shadow of moving objects and background distortion in predicted frames (see the warped patches in Fig. 1), therefore resulting in worse RD performance.

*Corresponding author: xiuyu.sxy@alibaba-inc.com.

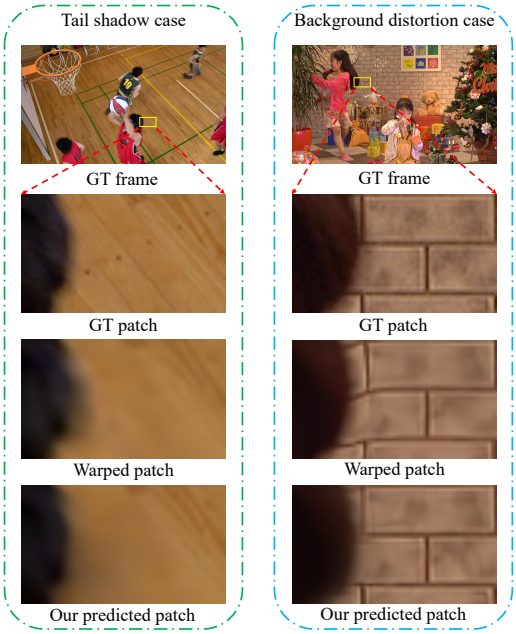

Fig. 1. **Visualization of the predicted frames.** Compared to the optical flow based warped patches, tail shadow and background distortion are significantly reduced in our proposed method.

To address these issues, we propose JMPNet to generate more accurate motion information. Dynamic local filters [12] are introduced to provide complementary motion information in addition to optical flow. An attention map is also introduced to fuse these two motion information to further improve the accuracy of predicted frames.

In particular, this paper has three major contributions: *1*. A JMPNet is proposed to tackle the tail shadow and background distortion problems in optical flow based video compression methods, which combines optical flow and dynamic local filter to generate more accurate predicted frames. *2*. An attention map is introduced to select optimal motion information for each pixel in a smarter way. *3*. Together with a novel scheme for motion information estimation (i.e., estimating horizontal and vertical motion information separately in an orthogonal way via two heads), our method achieves the SOTA performance in the most benchmarks.

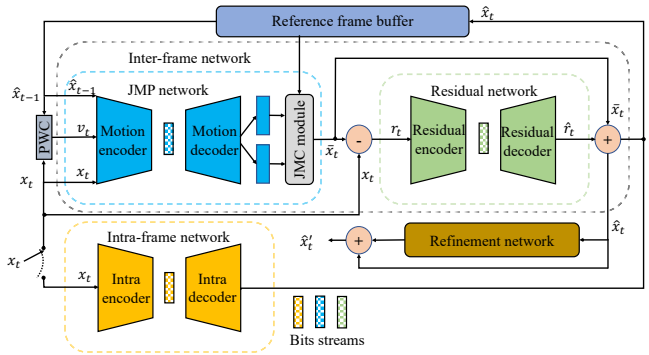

Fig. 2. Overview of our proposed video compression system. Intra-frame network, JMPNet, and residual network share the similar auto-encoder style network structure. The joint motion compensation (JMC) module is the operation of generating the predicted frame.

2. PROPOSED METHOD

As shown in Fig. 2, the reconstructed frame $\hat{x}_{t-1}$ and the original frame x_t are fed into PWC-net [13] to extract optical flow v_t. Let $v = \{x_0, x_1, ..., x_t, ...\}$ denotes a GOP in the target video sequence. x_0 is the key-frame, which is compressed via intra-frame network. At time step $t > 0$, motion information and residual are compressed via inter-frame network. $\overline{x}_t$ is the final predicted frame, residual r_t is set as $x_t - \overline{x}_t$. $\hat{r}_t$, and $\hat{x}_t$ are the corresponding reconstructed version, $\hat{x}'_t$ is the final reconstructed frame coming from refinement network.

The diagram of JMPNet is shown in Fig. 3, given the input information (v_t, $\hat{x}_{t-1}$, x_t), the encoder network compress the motion information into latent representation and the decoder network decompress the reconstructed motion information. At the last deconvolution layer, decoder network is split into two sub-networks. Each sub-network decompress the horizontal motion information $\hat{m}^h_t$ or the vertical motion information $\hat{m}^v_t$. $\hat{m}^h_t$ and $\hat{m}^v_t$ are split into $\{\hat{v}^h_t, K^h_t, Amap^h_t\}$ and $\{\hat{v}^v_t, K^v_t, Amap^v_t\}$ along the dimension of channel, then rebuilding $\{\hat{v}_t, K_t, Amap_t\}$ as shown in Equation (1), (4), and (6).The optical flow $\hat{v}_t$ is reconstructed as follow:

$$\hat{v}_t = concat(\hat{v}^h_t, \hat{v}^v_t) \quad \text{with} \quad dim = ch, \tag{1}$$

where $dim = ch$ denotes that $\hat{v}^h_t$ and $\hat{v}^v_t$ are concatenated along the dimension of channel.

After rebuilding the optical flow $\hat{v}_t$, the warped frame $\overline{x}^w_t$ are generated via warping operation:

$$\begin{aligned} &\overline{x}^w_t = Warp(\hat{x}_{t-1}, \hat{v}_t) \quad s.t. \\ &\overline{x}^w_t(i,j) = \hat{x}_{t-1}(i + \hat{v}^h_t(i,j), j + \hat{v}^v_t(i,j)), \end{aligned} \tag{2}$$

where $\hat{x}_{t-1}(i,j)$ represents the pixel of reference frame $\hat{x}_{t-1}$ at the coordinate (i,j). $\hat{v}^h_t(i,j)$ and $\hat{v}^v_t(i,j)$ denote the horizontal and vertical displacement in optical flow field.

In our proposed method, the dynamic local filters K_t are estimated by the JMPNet in a dense pixel-wise manner for each pixel:

$$K_t = \begin{pmatrix} k_{t,0,0} & k_{t,0,1} & \cdots & k_{t,0,W} \\ k_{t,1,0} & k_{t,1,1} & \cdots & k_{t,1,W} \\ \vdots & \vdots & \ddots & \vdots \\ k_{t,H,0} & k_{t,H,1} & \cdots & k_{t,H,W} \end{pmatrix}, \tag{3}$$

where H and W are the width and height of the video frame. $k_{t,i,j}$ is the dynamic local filter at coordinate (i,j). The shape of $k_{t,i,j}$, 1D filter $k^h_{t,i,j}$, and $k^v_{t,i,j}$ are (N,N), $(1,N)$, and $(N,1)$. We set N as 5 in this paper. $k_{t,i,j}$ is the product of $k^v_{t,i,j}$ and $k^h_{t,i,j}$:

$$\begin{aligned} k_{t,i,j} &= k^v_{t,i,j} \times k^h_{t,i,j} \\ &= \begin{pmatrix} k^{v,0,0}_{t,i,j} \\ k^{v,1,0}_{t,i,j} \\ \vdots \\ k^{v,N-1,0}_{t,i,j} \end{pmatrix} \times \begin{pmatrix} k^{h,0,0}_{t,i,j} & k^{h,0,1}_{t,i,j} & \cdots & k^{h,0,N-1}_{t,i,j} \end{pmatrix} \\ &= \begin{pmatrix} k^{0,0}_{t,i,j} & k^{0,1}_{t,i,j} & \cdots & k^{0,N-1}_{t,i,j} \\ k^{1,0}_{t,i,j} & k^{1,1}_{t,i,j} & \cdots & k^{1,N-1}_{t,i,j} \\ \vdots & \vdots & \ddots & \vdots \\ k^{N-1,0}_{t,i,j} & k^{N-1,1}_{t,i,j} & \cdots & k^{N-1,N-1}_{t,i,j} \end{pmatrix}, \end{aligned} \tag{4}$$

where $\times$ is the matrix multiplication.

In order to reduce the artifacts such as background distortion and tail shadow in the warped frame $\overline{x}^w_t$, the other predicted frame $\overline{x}^c_t$ is generated via dynamic local filters:

$$\begin{aligned} &\overline{x}^c_t = \hat{x}_{t-1} \circledast K_t \quad s.t. \\ &\overline{x}^c_t(i,j) = \hat{x}^p_{t-1,i,j} \circledast k_{t,i,j}, \end{aligned} \tag{5}$$

where $\hat{x}^p_{t-1,i,j}$ is a $N \times N$ patch of the reference frame $\hat{x}_{t-1}$ centering coordinate (i,j) and $\circledast$ is the local convolution. We treat each color channel equally and apply the same dynamic local filters K_t to each of RGB channels of $\hat{x}_{t-1}$ to synthesize $\overline{x}^c_t$.

Finally, x^w_t and x^c_t are fused by an attention map $Amap_t$:

$$\begin{aligned} &\overline{x}_t = x^w_t \times Amap_t + x^c_t \times (1 - Amap_t) \quad s.t. \\ &Amap_t = sigmoid(Amap^h_t + Amap^v_t). \end{aligned} \tag{6}$$

The goal of video compression is to keep distortion between the reconstructed and original frames to the minimum for a given number of bits, In other words, it is to consume the least bits for a given distortion using Lagrange RD cost function:

$$\begin{aligned} \mathbb{L} &= R + \lambda D \\ &= R_m + R_r + \lambda d(x, \hat{x}) \\ &= R_m + R_r + \lambda d(x, \overline{x} + \hat{r}), \end{aligned} \tag{7}$$

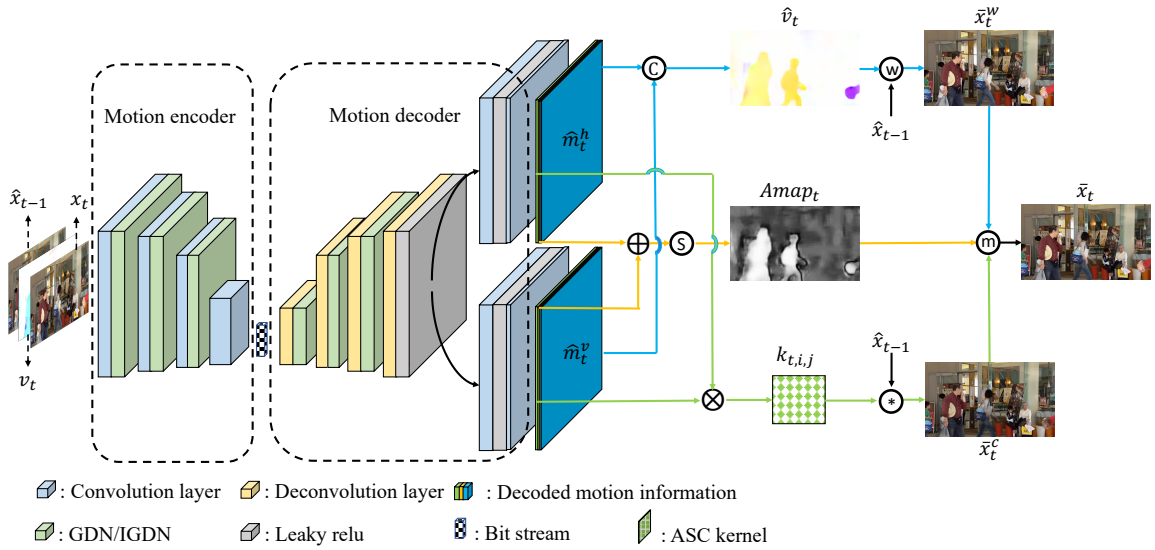

Fig. 3. **Overview of the joint motion prediction network.** Note that ⓒ represents the concatenation operation, ⓢ denotes the sigmoid function, ⊛ is the local convolution, ⓦ represents the warping operation, and ⓜ denotes fusing $\overline{x}_t^w$ and $\overline{x}_t^c$ via attention map $Amap_t$.

where R_m and R_r are bit per pixel (bpp) which are consumed to compress the motion information and residuals, respectively. λ is Lagrange multiplier determining the expected RD trade-off. d is distortion of reconstructed frame concerning the original frame measured by MSE or MS-SSIM.

3. EXPERIMENTS

Data-sets. Our proposed model is trained on the Vimeo-90K [15] with training samples randomly cropped into 256×256. In order to compare the RD performance of our proposed method with other SOTA methods, we evaluate the quantitative results on the UVG data-sets and the HEVC common test sequences class B and C, that are widely used to measure the RD performance of video compression methods.

Metrics. We follow the recent learning-based video compression work and use PSNR and MS-SSIM between the original and the final reconstructed frames in RGB color-space as the distortion metrics. The bpp is used to measure the number of consumed bits for compressing the video frames.

Quantitative Results on the test data-sets. As shown in Fig. 4, the experiments are performed on the UVG and HEVC class B and C data-sets. For the PSNR metric, our model is optimized for the MSE. On the UVG data-sets, our proposed method exceeds FVC [11], SSF [10],and HLVC [9] ∼0.4 dB, ∼0.96 dB, and ∼0.9 dB at ∼0.15 bpp. Despite our method achieve a distinct gain on UVG and class B data-sets, the performance of our method does not reach the level of FVC [11] on the class C data-sets, because the authors employ multi-frames which may have the better RD performance than the single reference frame in our method on this data-sets. When optimized for MS-SSIM, our proposed method shows a similar RD performance with the version optimized for PSNR and achieves the best RD performance in the most data-sets. Especially, despite multi-frames are used in the FVC [11], our method shows competitive performance on class C data-sets.

Ablation Study and Analysis. In order to validate the performance of each component in our proposed method, an ablation study is conducted as Fig. 5 and optimized for the MSE. Because the size of dynamic local filters are set as 5×5, and these small kernels can not handle large motion efficiently, *W/O Flow & RN* has the worst RD performance. When only the optical flow is used as the motion information, *W/O DFL & RN* achieves remarkable gain compared to *W/O Flow & RN*. It means that the optical flow is more efficient than the small dynamic local filter in the motion prediction. In *W/O RN*, the JMPNet can handle the large motion and keep visually pleasing in tail shadow and background distortion regions, and achieves excellent RD performance. In order to further improve the quality of reconstructed frames, we add an Unet [16] as the out-loop refinement network in *Proposed* version.

In order to demonstrate the influence of attention map $Amap_t$, the structure of network, and the size of dynamic local filters, we conduct a series of experiments shown in Fig. 5. We let $\overline{x}_t$ be equal to $0.5 \times (\overline{x}_t^w + \overline{x}_t^c)$ in *W/O Attention Map & RN*. Compared to *W/O RN*, this naive way of fusing two different motion information shows worse RD performance. It demonstrates that an incorrectly fusing method of two different motion information indeed brings negative gain. In *W/O RN & ONE PATH*, all motion information is decoded from a single sub-network and the RD performance is also worse than *W/O RN*. It demonstrates that decoding the motion information in

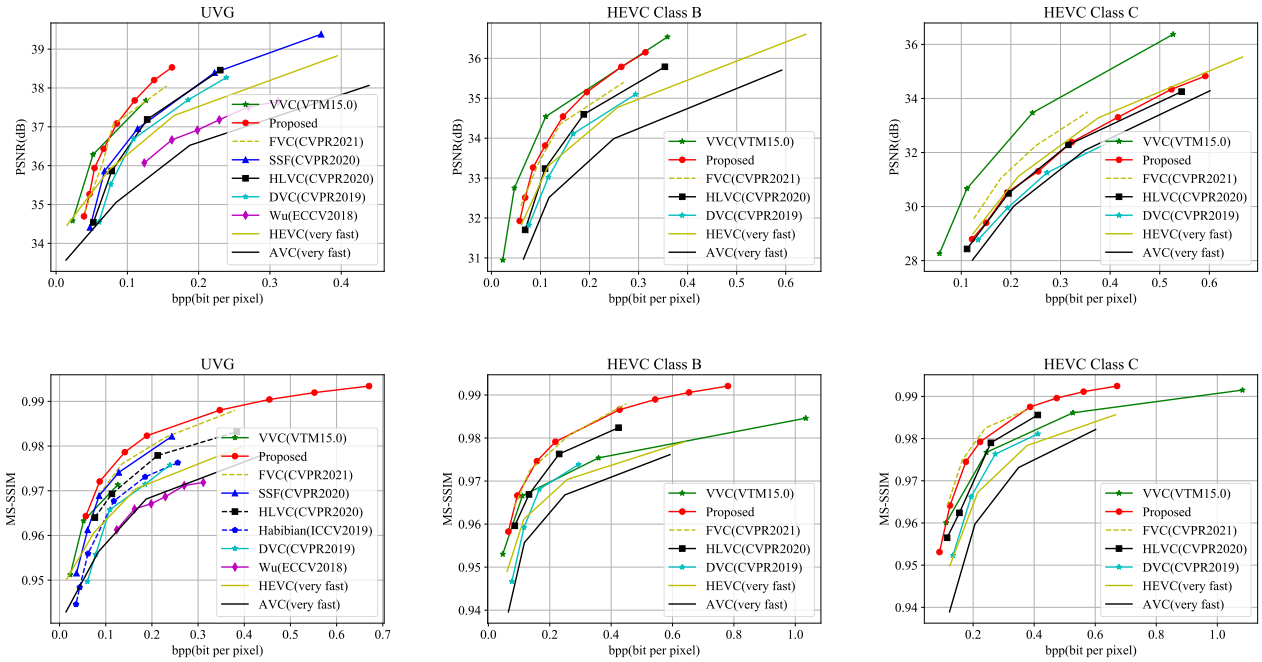

Fig. 4. **RD curve.** AVC [1], HEVC [2], VVC[14], Wu [4], DVC [8], HLVC [9], SSF [10], FVC [11], and our proposed method are compared on UVG, HEVC Class B , and C data-sets.

an orthogonal way is more efficient. When we use the bigger learned kernels in *W/O RN & N(11)* and *W/O RN & N(21)*, it can not achieve the better RD performance. It can be explained that bigger learned kernels can handle the larger motion, but more bits need to be consumed and the total RD performance may be non-ideal.

4. CONCLUSION

In this paper, we propose a joint motion prediction network for learning-based video compression. Multiple motion information such as optical flow and dynamic local filters are introduced to generate a more accurate predicted frame. The optical flow is responsible for large motion prediction while the dynamic local filters keep the visually pleasing in the tail shadow and background distortion regions which may appear in warped frames. An attention map is introduced to fuse the two motion information for each pixel. Qualitative results (see Fig. 1) demonstrate the effectiveness of our method in overcoming the background distortion and tail shadow problems. Quantitative results (see Figure 4) show that our method outperforms recently published SOTA learning-based methods [4, 8, 9, 10, 11] for both PSNR and MS-SSIM in the most data-sets. In conclusion, this paper introduces a novel scheme for motion prediction in learning-based video compression. The fusion of multiple motion information enables a more accurate motion prediction module and achieve better video compression performance.

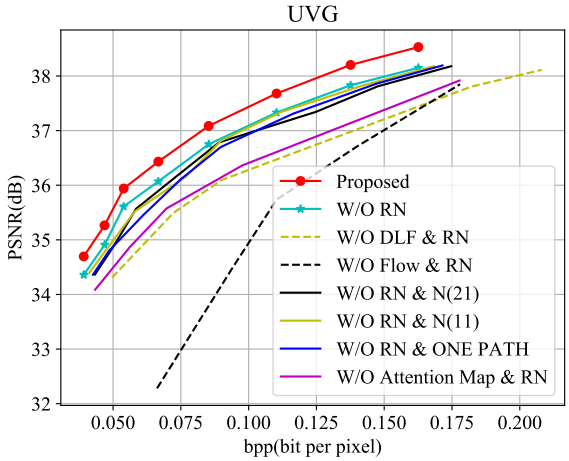

Fig. 5. **Ablation study.** The RD performance of the proposed method with different settings on UVG data-sets are reported. *W/O RN*: refinement network is removed. *W/O Flow & RN*: optical flow is removed. *W/O DLF & RN*: dynamic local filters are removed. *W/O Attention Map & RN*: attention Map is replaced by fixed weight 1/2. *W/O RN & ONE PATH*: compared to Fig. 3, at the last deconvolution layer, the feature map is directed into one sub-networks and the motion information is decoded from the single sub-network. *W/O RN & N(11)* and *W/O RN & N(21)*: the size of dynamic local filters are set as 11 and 21.

5. REFERENCES

[1] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra, "Overview of the h. 264/avc video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas 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.

[3] Jianjun Lei, Dongyang Li, Zhaoqing Pan, Zhenyan Sun, Sam Kwong, and Chunping Hou, "Fast intra prediction based on content property analysis for low complexity hevc-based screen content coding," *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 48–58, 2016.

[4] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl, "Video compression through image interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 416–431.

[5] Jorge Pessoa, Helena Aidos, Pedro Tomás, and Mário AT Figueiredo, "End-to-end learning of video compression using spatio-temporal autoencoders," 2018.

[6] Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen, "Video compression with rate-distortion autoencoders," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7033–7042.

[7] Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers, "Neural inter-frame compression for video coding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6421–6429.

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

[9] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6628–6637.

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

[11] Zhihao Hu, Guo Lu, and Dong Xu, "Fvc: A new framework towards deep video compression in feature space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1502–1511.

[12] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, pp. 667–675, 2016.

[13] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[14] Jens-Rainer Ohm and Gary J Sullivan, "Versatile video coding–towards the next generation of video compression," in *Picture Coding Symposium*, 2018, vol. 2018.

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

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.