Standard vs. Learning-based Codecs for Real Time Endoscopic Video Transmission

Aldo Marzullo[1][0000−0002−9651−7156], Martina Golini[2], Michele Catellani[3], Elena De Momi[2][0000−0002−8819−2734], Francesco Calimeri[1][0000−0002−0866−0834], Giuseppe Fiameni[4][0000−0001−8687−6609], and Iuri Frosio[4][0000−0002−7230−4287]

[1] Deparment of Mathematics and Computer Science, University of Calabria, Italy
https://www.mat.unical.it/demacs
{marzullo,calimeri}@mat.unical.it

[2] Dipartimento di Elettronica, Informazione e Bioingengeria, Politecnico di Milano, Milan, Italy http://www.polimi.it
{martina.golini, elena.demomi}@polimi.it

[3] Urology Department, Surgery Division, ASST Papa Giovanni XXIII, Bergamo, Italy
mcatellani@asst-pg23.it

[4] NVIDIA, NVIDIA AI Technology Center Italy http://www.nvidia.com
{gfiameni,ifrosio}@nvidia.com

Abstract. We compare traditional encoding/decoding methods for real time video streaming, like H264/AVC and H265/HEVC, and deep learning based methods, that are expected to deliver higher video quality at lower bandwidth in the next future. We concentrate our attention on the case of endoscopic videos, where streaming is part of a closed-loop system and robot-assisted minimally invasive surgery is performed on a patient in real time. Beyond low bandwidth and high video quality, such application also demands for low latency to guarantee the stability of the closed loop system and thus high safety standards. We analyze pros and cons of the deep learning approach in this domain, highlighting areas where deep neural networks overcome the traditional approach, and those that require further development. Our observations may be used as guidelines for the future research activity on video streaming in the surgical domain as well as in areas with similar requirements.

Keywords: Surgical Video · Latency · Bandwidth · Deep Learning · Compression · Codec.

1 Introduction

The advent of Deep Learning (DL) has allowed significant advances in many scientific and technological fields, including video compression and transmission. Learning based solutions have shown great effectiveness in reducing bandwidth

requirements while ensuring high quality of the transmitted video at the same time. For this reason, they are nowadays widely investigated, among other companies and in different application fields, by the principal streaming providers, like Disney [15] or Netflix [3], where the best compression rates can be achieved by resorting to specialized, per-title coding [1].

In recent years, advances in telecommunication technology and video coding systems have opened a new perspective also for surgical telementoring, remote diagnosis and teleoperation [4]. In this scenario, qualified surgeons give real-time supervision and technical help to the on-site physician during surgical procedures. This offers a transformative opportunity for accessing and delivering high-quality healthcare in resource-poor settings, particularly, but not exclusively, in disaster-affected and distant rural areas. However, in these contexts, high amounts of data (including high-resolution video frames) need to be transmitted, and bandwidth constraints constitute one of the primary bottlenecks for achieving real-time performances [4]. Indeed, it has been found that transmission errors (i.e., packet loss) can significantly reduce the perceived video quality for several surgical procedures, resulting in implications for the success of surgery tasks [9]. For these reasons, proper compression methods are crucial. Furthermore, when dealing with remote surgery, which often includes a closed-loop control mechanism and a surgery robot in the system, also latency has to be kept under strict control to guarantee the stability (and, consequently, the safety) of the entire system.

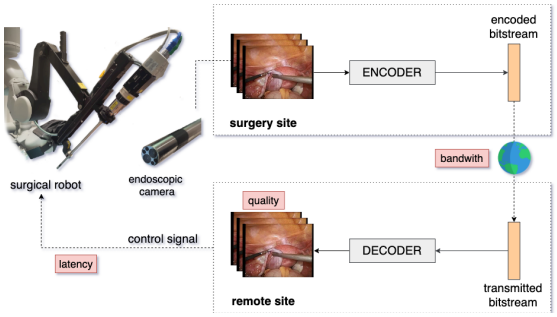

Fig. 1. In a typical minimally invasive surgery scenario, a remote surgeon controls a robot equipped with cameras (e.g. endoscopes) to frame the surgery field. Images are transmitted to the surgeon that takes action and controls the robot movements in a closed loop system. The latency introduced by data transmission has to be limited to guarantee fine and stable control of the robot. The amount of transmitted data per second have to be compatible with the bandwidth of the transmission channel. Finally, the quality of the images received at the remote site must be high enough to guarantee that no clinically significant information is lost in the data transmission process.

Lossless compression algorithms are not well suited for these tasks, due to the large bandwidth and the high latency requested. Conversely, lossy compression methods provide a valid alternative for real-time streaming, as they consume less bandwidth, but video quality and latency need to be carefully controlled to guarantee a good user experience. The H.264/AVC codec, which is widely adopted in several applications and can be hardware accelerated on many

devices, represents nowadays
the most viable choice also in the field of Minimally Invasive Surgery (MIS) [2, 10]. Its successor, H.265/HEVC, overcomes it in terms of quality, but it is computationally more demanding and, for this reason, not as widely spread as H.264. Both H.264 and HEVC are based on the hybrid prediction/transform coding method, proposed for the first time in 1979 by Netravali and Stuller [17], and can introduce block artefacts and other forms of quality compression degradation because of the quantization step applied before data transmission.

However, neither H.264 nor HEVC leverage the potential of learning methods in general, and DL in particular, that have just been started to be exploited for off-line video compression and streaming [15, 3, 1]. In this context, solutions for increasing the performance of one of the five main modules of the traditional codecs (intra-prediction, inter-prediction, quantization, entropy coding and loop filtering) have been proposed, as well as brand new codecs [17].

Here we perform a careful analysis of one of the state-of-the-art DL methods (Deep Video Compression, DVC [16]) for real-time video compression and transmission, and its comparison against H.264 and HEVC. We perform our analysis in the context of MIS, as this is the one of the most challenging applications with strict requirements in terms of quality and latency at the same time. More specifically, to realize a stable and clinically effective system, three constrains must necessarily be met (see Fig.1):

- Quality: the quality of the transmitted frames has to be good enough to guarantee that the surgeon can detect any detail which is clinically relevant, such as a small bleeding or unexpected tumor masses [2, 10].
- Latency: the images of the surgery field acquired by the camera, as well as the control signal going back to the robotic arm, must be received with the smallest possible delay, to guarantee the stability (no oscillations) of the closed-loop system [11]; this aspect is particularly critical if the surgeon is remote. Notice that both average and maximum latency are important in this context.
- Bandwidth: as for any video transmission system, the bandwidth required to transmit the data does not have to exceed the bandwidth allowed by the transmission system.

Our experiments highlight the points in favor of DL and the aspects that deserves more attention in future research for its adoption in real time video streaming of endoscopic videos, but our findings can be of use in other application fields with similar requirements. More in detail, for the tested DL-based codec (named DVC), we measured higher image quality and reduced bandwidth in comparison to H.264 and HEVC, at the cost of an increase in the latency, that can be however reduced with optimized implementations. In general, when compared to traditional codecs, DVC preserves better the high frequency components while introducing some light color shift, which is perceptually not too relevant. A more refined analysis revealed that the quality, as well as the latency in the transmission of the frames with DVC, have a large range of variation, which is detrimental for its practical adoption. Moreover, we found that I-Frames

transmitted with DVC are characterized by high latency, and their high quality strongly influence (in a positive way) the quality of the following frames that use prediction to be transmitted with limited bandwidth. This suggests that one of the key aspect for the adoption of DL in video transmission systems is the development of computationally efficient, single frame compression methods.

The paper is organized as follows: we perform a detailed review of the state-of-the-art video compression and transmission methods in the next section, then we introduce the experimental setup adopted to compare H.264, HEVC, and DVC, and we present and discuss the results towards the end of the paper.

2 Related Work

Among the many traditional video codecs, H.264 and HEVC are the most adopted and diffused [22]. Both these codecs are based on the hybrid prediction/transform coding method, first proposed in 1979 [17]. Despite HEVC overcomes H.264 in several aspects, the latter is hardware friendly, and thus easily implemented and distributed in its accelerated version. This makes it the de facto standard for most of the existing video streaming applications, including the surgical domain [10]. As a consequence of the optimized implementation, it is also characterized by small latency. Counter side, as both H.264 and HEVC use block-based coding (i.e., square blocks in the transmitted frames are processed independently one from each other), these schemes can introduce block artifacts; quality degradation can also be associated to the quantization of the compressed data stream. For these reasons, DL-based codec have started to be explored as a promising alternative.

Researchers developed several DL-based methods in recent years [17], where neural networks have been used for building brand new end-to-end schemes for compression, enhancement and restoration of the video quality, or as a tool to increase the performance of one of the five main modules of the traditional codecs: intra-prediction, inter-prediction, quantization, entropy coding and loop filtering [17].

For instance, Lu *et al.* [16] proposed Deep Video Compression (DVC), one of the first end-to-end DL-based video codec. In DVC, each step of the traditional compression pipeline was substituted by DL models which were jointly trained at minimizing the reconstruction error while reducing the bits used for compression, reaching state-of-the-art results at the time of publication. As this is one of the most comprehensive DL-based approach, covering all the aspects of video encoding, transmission and decoding, it is also the one we considered for our experiments.

When used for replacing single modules of the codec pipeline, DL-based intra/inter prediction solutions, as well as post processing filtering techniques, have shown good performance especially in a low bit-rate scenario. Li *et al.* proposed a five layers CNN-based block up-sampling scheme for intra-frame coding [12]. Each Coding Tree Unit (the basic processing unit of the HEVC) is firstly down-sampled, then coded by HEVC, and eventually decoded and up-sampled to its

original resolution and post-processed by the CNN. This scheme achieved an important reduction in terms of required bandwidth (5.5% for HEVC common test sequences and 9.0% for Ultra High Definition test sequences). On the other hand, compression noise due to the dependency of the CNN from the Quantization Parameters (QPs) used in compressed training videos has been highlighted in some cases. Moreover, the CNN encoding/decoding time was significantly higher when compared to HEVC (although without any optimization for speed on the CNN side [12]). The same authors proposed an extension of their scheme for inter-frame predictions in 2019 [13]. Feng *et al.* [6] developed a dual network structure to improve the reconstruction quality of videos compressed at low resolution. Here, an enhancement module operates before a super-resolution network to deal with sampling and compression artifacts separately. The model achieved about 31.5% bit-rate saving when compared to HEVC. Zhang *et al.* proposed a residual convolutional neural network for loop filtering in HEVC [26]. In this scheme, the QP range is divided into several bands and a dedicated network is trained with a progressive training scheme for each of them. This framework achieved substantial coding gains, especially for low bit rates, but the encoding time heavily increased [17]. Another approach based on transmitting frames using the traditional H.264 codec and then refining the transmitted frames was proposed in [24], where a small encoder network is first trained to generate a binary code that is transmitted together with the frame data, whereas on the decoder side a small DNN decoder applies a residual correction to the frames decoded by H.264.

The literature about the use of DL for video compression and transmission in the surgical domain is, on the other hand, limited. It has been shown that the detection of clinically relevant spatio-temporal information can be exploited to save compression time, while maintaining high quality in the transmitted frames. To this aim, CNNs are used to segment the input frames and detect Regions of Interest (ROI) whose quality needs to be better preserved. In [21], Munzer *et al.* identified domain-specific features of endoscopic videos that can be exploited for an efficient compression. In [7], Ghamsarian *et al.* proposed a cataract surgery video compression approach, based on HEVC, with the aim of preserving high quality in meaningful regions. Two separate networks were employed for the classification and segmentation of such regions and larger distortion on irrelevant content was allowed. Hassan *et al.* proposed a CNN-based segmentation network (S-CNN) which has been demonstrated useful for real time applications in a limited bandwidth scenario [9]. It is composed by four convolution layers that identify the surgical regions that need to be transmitted in high quality, differently from the background. Low QP values - corresponding to high quality outputs - are then used for SR regions, whereas high QP values are used for the background. In comparison to the standard HEVC scheme, S-CNN achieved an average bit-rate reduction of 88.8% at HQ settings (QP in range of 0–20) [9]. Differently from DVC, none of the aforementioned approaches adopted in the surgery context covers all the five components of a codec system.

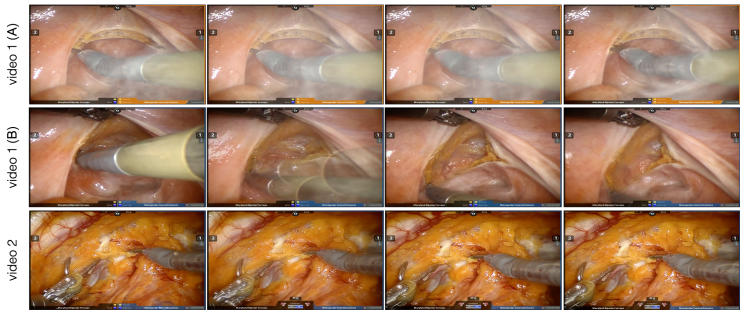

Fig. 2. Subset of frames extracted from the set of 40s clips used for our experiments. Video 1 (A) contains smoke. Video 1 (B) shows fast movements of the surgical instrument and biological tissues in the scene. Video 2 contains fast movements of the surgical instrument together with large camera drifts.

3 Testing of Standard and DL-based Codecs

In our experiments we performed a comparison in the surgical domain between traditional (H.264, HEVC) codecs and DVC [16], one of the first completely DL-based video codecs.

We selected robotic assisted radical prostatectomy (RARP) as a representative procedure, as it constitutes one of the most performed robotic assisted MIS operation [5]. RARP includes three phases. The first one is the pelvic lymphadenectomy, where the focus is concentrated at the level of the iliac vessels; in this phase, the most delicate structures in the center of the field are the blood vessels that must be freed from the lymph nodes; the surgical field is small, the surgical movements are slow and more delicate. In the following step, called "demolition phase", the prostate is isolated posteriorly from the bladder, from the nerve bands laterally and anteriorly from the urethra; here the surgical field is wider, movements are faster and the organ of interest, the prostate, is in the center of the visual field; the peripheral area is occupied by the iliac vessels laterally and by the pubic bone over. In the last reconstructive phase, the bladder neck is sutured to the urethra; the surgical field is tight since the anastomosis between bladder and urethra is performed in the small pelvis; movements are small and mostly in the center of the surgical field.

We extracted ten clips (40 seconds each) from 94 minutes, high quality (1200×720) youtube video with the endoscopic view captured during RARP (Fig. 2). The clips were selected to maximize their diversity from different phases of RARP, and thus they include different anatomical sections, surgery instruments, levels of illumination and degrees of action performed in the surgery field.

For each 40s clip, our aim was to measure quality, bandwidth and latency as a function of the adopted codec. To do so, we compressed and decompressed each clip using the H.264 and HEVC implementations provided by ffmpeg [23]. Each clip was encoded at different bitrates, ranging from 1 to 30 Mb/s (corresponding approximately to 0.30 to 9.25 Bit Per Pixels, BPP). Compressing at different bit

rate allowed investigating the codec performance as a function of the transmission bandwidth. In ffmpeg, H.264 and HEVC come with predefined presets that achieve different compression ratio / frame quality / compression time (latency) compromises. More specifically, some preset is designed to compress the frame in a short amount of time (low latency) at the cost of decreased quality and larger bandwidth, while others achieve the highest compression rate and frame quality, but require more processing time. In our experiment, we considered the *Ultrafast*, *Medium* and *Slow* presets, whose interpretation in terms of quality / bandwidth / latency should be clear to the reader.

For each frame in the encoded/decodec clip, and each BPP/preset pair, we measured then the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), where the original, uncompressed video is the ground truth. Furthermore, we measured the encoding and decoding time that, once summed to the transmission time, gives the total latency.

To perform our comparison, we encoded and decoded the same clips with DVC [16], where all the encoding and decoding components (motion estimation, motion compensation, residual compression, motion compression, quantization, and bit rate estimation) are implemented by end-to-end neural networks. In particular, a learning based optical flow estimation network obtains the motion information; two auto-encoders compress and reconstruct the corresponding motion and residual information. All the components are jointly trained to optimize a bit rate - distortion (measured by PSNR or SSIM) trade-off through a single loss function, controlled by the hyperparameter λ. We considered the DVC model optimized for PSNR, trained on Vimeo-90k (composed of a large variety of real world scenes and actions) and using $\lambda = 2048$ to achieve the best reconstruction quality. It is worth noting that in our DVC implementation, only differential frames are encoded using the DL models, while I-frames are encoded through the BPG compression scheme. We measured PSNR, SSIM and compression / decompression times for DVC (as well as for BPG) and compared them with those obtained for H.264 and HVEC. To evaluate the impact of the I-Frame coding on the overall performance of DVC, we tested three DVC configurations, where the I-Frame was encoded every 5, 10 and 150 frames.

4 Results

4.1 Average quality and encoding / decoding time

The first row of Fig. 3 shows the average PSNR and SSIM computed over the entire set of clips, together with the encoding and decoding time, as a function of the BPP, for H.264 and different codec presets. As expected, the quality of the decoded frames increases with the bandwidth. On the other hand, also the encoding and decoding time increases with BPP (as more time is required to process a larger amount of data), up to the point (at least for the case of the *slow* preset, BPP > 2: Fig. 3, first row, third panel, green line) where encoding is not feasible in real time anymore, at least for the video resolution considered here. HEVC shows a similar behavior (second row in the same figure), although,

when compared to H.264, it achieves a slightly better frame quality in terms of both PSNR and SSIM, but at the cost of a higher encoding time, that makes it unsuitable for real time streaming (at least for the video resolution considered here), if not using the *fast* preset.

Fig. 3 also reports the performance of the DL-based DVC codec, that achieves a higher average frame quality while encoding data at a lower bit rate. On the other hand, the encoding and decoding time for the unoptimized DVC implementation considered here are significantly (one order of magnitude or more) higher than those measured for the highly optimized for speed and hardware accelerated H.264.

By increasing the frequency with which the I-Frames are encoded in DVC, a overall higher average frame quality can be achieved. This trend is clearly explained by the fact that the pure BPG codec, which is characterized by minimal compression loss, achieves the best PSNR/BPP compromises (and the overall best SSIM), but at the price of a much higher encoding time, which renders BPG unsuitable for real time applications.

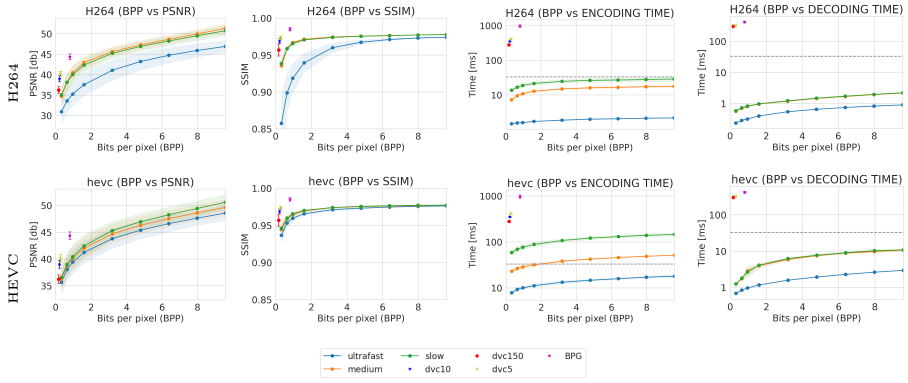

Fig. 3. Average PSNR, SSIM, encoding, and decoding time as a function of the BPP for H.264 (first row) and HEVC (second row), for the set of 10 endoscopic clips considered here. The shaded areas represent one standard deviation. The colored points and bars report the average value and standard deviation of DVC, when I-frames are encoded with BPG every 5, 10, or 150 frames. The values for pure BPG encoding are also reported. The black, dashed line represents the real time threshold (30 fps here). Best viewed in color.

4.2 Per frame quality and encoding / decoding time

To investigate more in detail the performances of the three codecs, we also conducted a per-frame analysis. Figs. 4 and 5 report the per-frame PSNR and SSIM for two 40s clips extracted from the RARP video. The two clips are characterized by different content, luminance level and dynamic conditions observed by

the endoscopic camera (Fig. 2). On *video 1*, DVC achieves an average frame quality comparable to that obtained by H.264 and HEVC under the *medium* and *slow* preset, while it performs better than H.264 under the *ultrafast* preset. Both the traditional codes and DVC show an oscillatory pattern in terms of quality that are associated with the transmission of the I-Frames in high quality (PBG compression in the case of DVC). Fig. 4 also shows a sudden performance decrease in terms of PSNR (and a less significant decrease in terms of SSIM) for DVC around frame 540 (denoted as (A)). Visual inspection (Fig. 2, video 1 (A)) reveals that smoke is present at this point in the clip, which alters the colors of the scene in a way that is not well captured by the DVC compression algorithm. Furthermore, we can observe that H.264 and HEVC achieve better quality (compared to DVC) in the first part of the clip, characterized by slow motion and static scenes, while DVC performs better in the second part ((B) in Fig. 4), where the surgical instruments execute fast incisions and drifts of the camera field of view are visible. While the quality of the frames compressed and transmitted by H.264 and HEVC drifts significantly along the clip, DVC performances in terms of quality are more stable. The quality of the frames treated by DVC remains pretty constant over time also in *video 2*, where the RARP procedure is significantly more dynamic compared to *video 1* (Fig. 5). In this conditions, DVC overcomes H.264 and HVEC in terms of quality in almost every frame.

Finally, Fig. 6 shows the residual error between the original frames and the same frames compressed and reconstructed with H.264, HEVC and DVC, for three video sequences randomly extracted from our set. When H.264 is adopted, errors are mostly concentrated in the high frequency domain, i.e. mostly around edges and small image details. HEVC clearly achieves higher quality, with residual errors in the middle / high frequency domain. When DVC is adopted, on the other hand, the error is dominated by low frequency residuals, that can be seen as slight color shift of large objects, while edges and small details appear to be well reconstructed. The result from visual inspection is thus consistent with the PSNR and SSIM metrics measured in our previous experiments: while DVC does not always overcome H.264 and HEVC in terms of PSNR, because of a generalized color shift that creates a numerically large error with (likely) poor clinical significance, it also preserves edges better, which leads to a better quality perceived by a human observer, as captured by SSIM (a metric that was designed to vaguely resemble the response of the human visual system).

5 Discussion and Conclusion

We have analyzed the problem of video transmission in the specific case of surgical operations, which is characterized by peculiar constraints: to guarantee the stability of the system, the latency must be kept under control, while bandwidth may be limited; at the same time, the quality of the transmitted frames has to be sufficient to guarantee that no significant clinical information is lost in the process of compressing, transmitting and reconstructing frames.

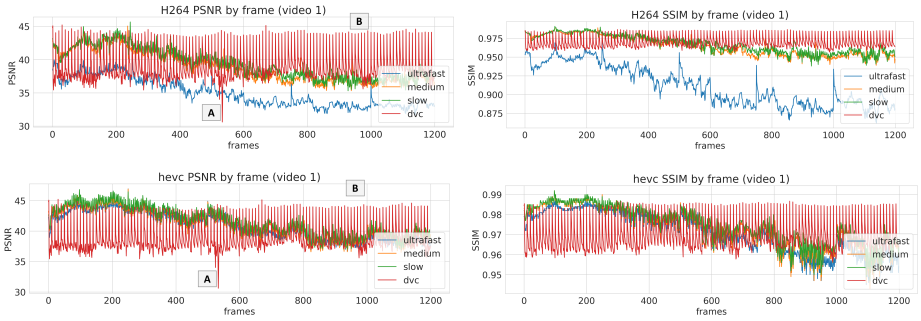

Fig. 4. Per-frame PSNR and SSIM as a function of BPP for H.264, HEVC and DVC on **video 1** (see Fig. 2 for sample frames extracted around A and B).

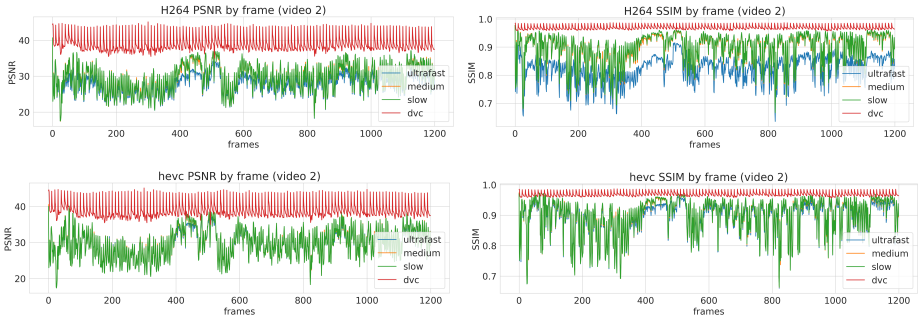

Fig. 5. Per-frame PSNR and SSIM as a function of BPP for H.264, HEVC and DVC on **video 2** (see Fig. 2 for sample frames extracted from the video).

Our results show that, despite some of the existing codecs are already capable of (and, indeed, already used for) transmitting frames at low latency and good quality, DNNs (or, at least, the DVC method herein considered) are potentially capable of transmitting video frames at higher quality while consuming less bandwidth. Furthermore, it is worth noticing that the considered network was trained on a real-world scenes dataset and higher reconstruction quality can be expected if the network is trained on endoscopic videos. However, the naive implementation of DVC considered here (and likely those of many other DNNs) does not satisfy the latency constraint — in other words, whereas traditional approaches based on H.264 and HEVC codecs are already largely optimized for speed and quality, DNNs have a much larger margin of improvement that is not completely explored yet. Some of the possible optimizations in terms of both quality and speed are therefore mentioned in the following.

We observed that, when using DVC, the average quality of the transmitted frames significantly increases when many I-Frames are encoded through BPG.

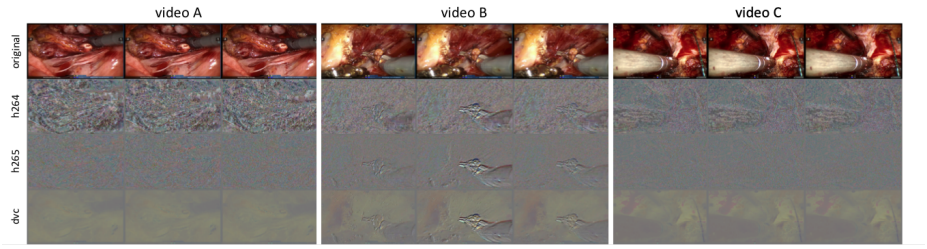

Fig. 6. Residual errors after compression and decompression of frames with H.264, HEVC, and DVC for three different clips randomly extracted from the set of 40s clips used in our experiment. While errors are concentrated in the high frequency domain for H.264 and HEVC, DVC shows reconstruction errors mostly in the low frequency domain (e.g. color shift).

On the other hand, the BPG encoding time (around one second per frame in our experiments) largely exceeds the maximum allowed by surgical practice and consequently obliges surgeons to change their workflow (e.g. adopting a *move-and-wait* strategy) while also affecting the effectiveness of the surgical operation [11]. Therefore, exploring more efficient methods to compress and transmit I-Frames at high quality is needed, so that the latency introduced while transmitting them is less critical. Solutions that exploit both past and future frames (e.g. [25]) cannot be applied in the case of real-time streaming, whereas other directions like partial transmission (block-based) of I-Frames or the adoption of ad-hoc vocabolaries [1] promise to deliver significant improvements in the future.

In the case of the non optimized DVC implementations considered here, the latency (even without considering the transmission of the I-Frames) still exceeds the limit that allows a surgeon to effectively operate without being affected by it (around 160ms [11]); even more important, as the encoding time is larger than 33ms per frame[1], it does not even allow real time transmission at 30Hz, as a significant delay would be accumulated over time. There are however several methods to accelerate DNNs that can be easily exploited, ranging from lightweight learning based solutions for image compression [17] to software based solutions that prune the DNN to make them more efficient [20], hardware-aware optimizations of the network implementation [19] and up to the adoption of GPU accelerators for DNN like Tensor cores [18] or even ad-hoc hardware DNN implementations [8] that can be seen as equivalent to hardware-accelerated H.264 encoders and decoders.

[1] It is worthy noticing that the total latency is given by the sum of the encoding, transmission, and decoding times; on the other hand, the maximum among (encoding + transmission) and (transmission + decoding) defines the working frequency of the system — e.g., if encoding and transmission take overall 100ms, the maximum number of frames transmitted per second and without accumulating delays will be 1s / 100ms / frame = 10 frames.

In the end, it is worth noticing that often surgery procedures also make use of stereo images to enable the perception of 3D information in the surgical field. This complicates the transmission procedure, as the size of the data doubles, but it also offers the possibility to take advantage of redundancy in left and right views for the development of novel stereo codecs (see for instance in [14]) that can find application in fields even very distant from the medical ones, such as videogames or virtual reality.

References

1. Per-title encode optimization (2015)
2. Chaabouni, A., Gaudeau, Y., Lambert, J., Moureaux, J.M., Gallet, P.: H. 264 medical video compression for telemedicine: A performance analysis. IRBM **37**(1), 40–48 (2016)
3. Chen, L.H., Bampis, C.G., Li, Z., Norkin, A., Bovik, A.C.: Proxiqa: A proxy approach to perceptual optimization of learned image compression. IEEE Transactions on Image Processing **30**, 360–373 (2021). https://doi.org/10.1109/tip.2020.3036752, http://dx.doi.org/10.1109/TIP.2020.3036752
4. Collins, J.W., Ma, R., Beaulieu, Y., Hung, A.J.: Telementoring for minimally invasive surgery. In: Digital Surgery, pp. 361–378. Springer (2021)
5. Dasgupta, P., Kirby, R.S.: The current status of robot-assisted radical prostatectomy. Asian journal of andrology **11**(1), 90 (2009)
6. Feng, L., Zhang, X., Zhang, X., Wang, S., Wang, R., Ma, S.: A dual-network based super-resolution for compressed high definition video. In: Pacific Rim Conference on Multimedia. pp. 600–610. Springer (2018)
7. Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M., Schöffmann, K.: Relevance-based compression of cataract surgery videos using convolutional neural networks. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3577–3585 (2020)
8. Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M.A., Dally, W.J.: Eie: Efficient inference engine on compressed deep neural network. ISCA '16, IEEE Press (2016). https://doi.org/10.1109/ISCA.2016.30, https://doi.org/10.1109/ISCA.2016.30
9. Hassan, A., Ghafoor, M., Tariq, S.A., Zia, T., Ahmad, W.: High efficiency video coding (hevc)–based surgical telementoring system using shallow convolutional neural network. Journal of digital imaging **32**(6), 1027–1043 (2019)
10. Kumcu, A., Bombeke, K., Chen, H., Jovanov, L., Platisa, L., Luong, H., Looy, J., Nieuwenhove, Y., Schelkens, P., Philips, W.: Visual quality assessment of h.264/avc compressed laparoscopic video (01 2015). https://doi.org/10.1117/12.2044336
11. Kumcu, A., Vermeulen, L., Elprama, S.A., Duysburgh, P., Platiša, L., Van Nieuwenhove, Y., Van De Winkel, N., Jacobs, A., Van Looy, J., Philips, W.: Effect of video lag on laparoscopic surgery: correlation between performance and usability at low latencies. The International Journal of Medical Robotics and Computer Assisted Surgery **13**(2), e1758 (2017)
12. Li, Y., Liu, D., Li, H., Li, L., Wu, F., Zhang, H., Yang, H.: Convolutional neural network-based block up-sampling for intra frame coding. IEEE Transactions on Circuits and Systems for Video Technology **28**(9), 2316–2330 (2018). https://doi.org/10.1109/TCSVT.2017.2727682

13. Lin, J., Liu, D., Yang, H., Li, H., Wu, F.: Convolutional neural network-based block up-sampling for hevc. IEEE Transactions on Circuits and Systems for Video Technology **29**(12), 3701–3715 (2019). https://doi.org/10.1109/TCSVT.2018.2884203
14. Liu, J., Wang, S., Urtasun, R.: Dsic: Deep stereo image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3136–3145 (2019)
15. Lombardo, S., HAN, J., Schroers, C., Mandt, S.: Deep generative video compression. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/f1ea154c843f7cf3677db7ce922a2d17-Paper.pdf
16. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: Dvc: An end-to-end deep video compression framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
17. Ma, S., Zhang, X., Jia, C., Zhao, Z., Wang, S., Wang, S.: Image and video compression with neural networks: A review. IEEE Transactions on Circuits and Systems for Video Technology **30**(6), 1683–1698 (2020). https://doi.org/10.1109/TCSVT.2019.2910119
18. Markidis, S., Chien, S.W.D., Laure, E., Peng, I.B., Vetter, J.S.: Nvidia tensor core programmability, performance and precision. 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) (May 2018). https://doi.org/10.1109/ipdpsw.2018.00091, http://dx.doi.org/10.1109/IPDPSW.2018.00091
19. Molchanov, P., Hall, J., Yin, H., Kautz, J., Fusi, N., Vahdat, A.: Hant: Hardware-aware network transformation (2021)
20. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J.: Importance estimation for neural network pruning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11256–11264 (2019). https://doi.org/10.1109/CVPR.2019.01152
21. Münzer, B., Schoeffmann, K., Böszörmenyi, L.: Domain-specific video compression for long-term archiving of endoscopic surgery videos. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS). pp. 312–317 (2016). https://doi.org/10.1109/CBMS.2016.28
22. Punchihewa, A., Bailey, D.: A review of emerging video codecs: Challenges and opportunities. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE (2020)
23. Tomar, S.: Converting video formats with ffmpeg. Linux Journal **2006**(146), 10 (2006)
24. Tsai, Y.H., Liu, M.Y., Sun, D., Yang, M.H., Kautz, J.: Learning binary residual representations for domain-specific video streaming. In: AAAI (2018)
25. Yang, R., Mentzer, F., Gool, L.V., Timofte, R.: Learning for video compression with hierarchical quality and recurrent enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6628–6637 (2020)
26. Zhang, Y., Shen, T., Ji, X., Zhang, Y., Xiong, R., Dai, Q.: Residual highway convolutional neural networks for in-loop filtering in hevc. IEEE Transactions on Image Processing **27**(8), 3827–3841 (2018). https://doi.org/10.1109/TIP.2018.2815841