APSIPA Transactions on Signal and Information Processing, 2022, 11, e16
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, for non-commercial use, provided the original work is properly cited.

Forum Paper

The Future of Video Coding

Nam Ling[1,*], C.-C. Jay Kuo[2], Gary J. Sullivan[3], Dong Xu[4], Shan Liu[5], Hsueh-Ming Hang[6], Wen-Hsiao Peng[7] and Jiaying Liu[8]

[1]*Department of Computer Science and Engineering, Santa Clara University, 500 El Camino Real, Santa Clara, CA 95053-0566, USA*

[2]*Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089-2564, USA*

[3]*Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA*

[4]*School of Electrical and Information Engineering, The University of Sydney, NSW 2006, Australia*

[5]*Tencent Media Lab, 2740 Park Ave., Palo Alto, CA 94306, USA*

[6]*Department of Electronics Engineering, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan*

[7]*Department of Computer Science, National Yang Ming Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan*

[8]*Wangxuan Institute of Computer Technology, Peking University, No. 128 Zhongguancun North Street, Haidian District, Beijing 100871, China*

ABSTRACT

This article summarizes the panel discussion on "The Future of Video Coding," organized by the U.S. Local Chapter of APSIPA on April 24, 2021. This panel brought together world leading experts in video coding to discuss and debate this hot topic. The speakers are leaders from various video coding fields, ranging from video coding standards, visual quality, deep learning approaches, screen content coding, reinforcement learning, rate-control, video coding for machines, and more. The panel discussed and debated different future issues related to video coding, including but not limited to, the emerging areas within the next 5 years, the trend and role of deep-based coding, the impact of visual quality

*Corresponding author: Nam Ling, nling@scu.edu.

Received 29 August 2021; Revised 01 January 2022
ISSN 2048-7703; DOI 10.1561/116.00000044
© 2022 N. Ling *et al.*

assessment, the role for academia, the advice for graduate students, and more. The key points of the panellists' opinions are highlighted in italics.

Keywords: Video coding, video compression, visual communications, visual quality, deep learning

1 Introduction

[Nam] In recent years, there have been significant advances in the research and development video coding technology and system. With the latest standardization of Versatile Video Coding (VVC) standard in July 2020 [8, 9], experts start wondering what the next standard could be that could achieve a similar coding efficiency gain beyond VVC/H.266. On the other hand, deep learning tools have gained much attention in recent years, including applying them to assist conventional video coding and to replace the conventional hybrid coding model in video coding [30, 31]. However, these tools come with very high computational complexity and high power consumption, which make them currently impractical in many cases. Overcoming such challenges is an important direction of research. On the visual quality side, there have been many proposals on developing models beyond the traditional computational models like peak-signal-to-noise ratio (PSNR), or even the multi-scale structural similarity (MS-SSIM) metric. Could there be a universally accepted computational, perceptual, or visual attention model for visual quality? Finally, with many participants from the academia, many wonder what key roles can the academia play and what advices we have for graduate students interested in pursuing a career in video coding. With so many challenges ahead the future of video coding research is still bright.

On April 24, 2021, the U.S. local chapter of the APSIPA brought together world leading experts to debate and to discuss "The Future of Video Coding." Members of the panel are all among the top experts in the video coding field. Given that we had a record-breaking attendance and a lively discussion covering many topics, we envision that there is a lot of interest in this subject and that many might not have had a chance to participate, we have therefore produced this article summarizing the opinions of the experts, with the members of the panel as the authors of this paper. Our panel discussions were based on five questions. (1) What do you think the hot emerging area(s) within video coding will be in the next 5 years? (2) Will deep-based-coding become the main trend in video coding? Yes, or No? Why? (3) Visual quality assessment is a hot research topic in academia. Many papers have been published. Will this effort have a real impact in video coding standardization? (4) Is there a role for academia in developing the new generation of video coding technology (or

standard)? (5) What advice do you have for graduate students who would like to pursue video coding research? Finally, time was allocated for the audience to ask questions and interact with the panel. A summary of the main points are presented in the conclusion section.

2 Hot Emerging Video Coding Areas

The first question posted to the panel is "What do you think the hot emerging area(s) within video coding will be in the next 5 years?"

2.1 Emerging Technology and Applications

2.1.1 Dual-track Approach

[Gary] Now we have a dual track approach. *In addition to looking at the traditional approach, experts are looking into using neural networks.* Immersive applications are also a big area, with things like six degree of freedom video (and even three-degree of freedom video is a challenge). Rapid streaming of "tile" regions depending on where one is looking within the scene is challenging. Even for ordinary coding technology in the so-called classic style, we are still making progress. In the JVET now, we have shown about 13–14% gain beyond the VVC version 1 standard of July 2020. However, complexity increases continue to be a challenge. The application space is growing as well, from broadcast to streaming to mobile, 8K video is now in the consumer application, and HDR and high frame rate are also seeing applications.

2.1.2 Motivation from Emerging Applications

[Shan] *The advent of new technologies is often motivated by emerging applications* [49]. As an example, screen content coding tools [48] are developed and used in applications involving computer-generated contents such as screen sharing and video conferencing for enhanced compression performance and bandwidth savings. During the Covid-19 pandemic, video conferencing has become an essential to our daily lives and helped numerous people to conduct school and work remotely. Nowadays screen content tools are adopted in almost all widely used video communication products and applications. Looking into the future, we expect that volumetric video, virtual reality, cloud gaming, and low power edge computing will have high demand. Another emerging area which has recently drawn attention is video coding for machines. Machine vision tasks such as object detection, segmentation and tracking have been used in many applications including intelligent transportation and smart city, etc. The volume of video consumed by machines is rapidly increasing and

hence, compressing video for machines and machine vision tasks has become important. MPEG created an ad-hoc group called VCM (video coding for machines) in 2019 and issued a call for evidence in January 2021 [13].

2.2 Deep Learning Approaches

[Wen-Hsiao] Currently there are three emerging areas related to deep-based approaches: (1) deep learning (DL)-assisted compression – by enhancing traditional codecs without changing the codecs; (2) deep learning-based compression – by using neural networks as the backbone of the compression system; and (3) hybrid systems – by incorporating DL-based tools or enhancement layers into traditional codecs.

2.2.1 DL-assisted Compression

[Wen-Hsiao] *In DL-assisted approaches, neural networks may be used for pre-processing the input video to repurpose a conventional codec without changing the codec.* For example, the input video can be pre-processed in such a way that after compression, the decoded video shows better results in MS-SSIM instead of the traditional PSNR or MSE (see Figure 1a). Likewise, it can be pre-processed such that the decoded video is suitable for vision tasks, such as in video coding for machines [24]. Another approach is to use neural networks in both pre-processing and/or post-processing [28]. In the pre-processing step, the neural network embeds useful information in the input image that can be extracted from the decoded image in the post-processing step to better achieve the goal of quality enhancement, rate saving, or adapting the decoded images/videos to vision tasks (see Figure 1b).

We can also apply reinforcement-learning to encoder control tasks [10, 16, 17, 42]. An example is rate-control. We have a neural-network based agent that interacts with the traditional codec to learn to control it (see Figure 1c). Remarkably, the work in [42] presents an example of video coding for machines by applying the same idea to training an agent that optimizes bit allocation for computer vision tasks. Another area is to apply lightweight neural networks to tasks like fast mode decision [25, 34, 40, 47].

2.2.2 DL-based End-to-end Compression

[Wen-Hsiao] *End-to-end learned codecs are catching up quickly these days.* In terms of PSNR, the state-of-the-art learned image codec performs comparably to VVC Intra (see Figure 2a); the best learned video codec achieves better PSNR than that of HEVC under the Low-delay P test condition with a group-of-pictures (GOP) size of 12, and approaches the performance of VVC at

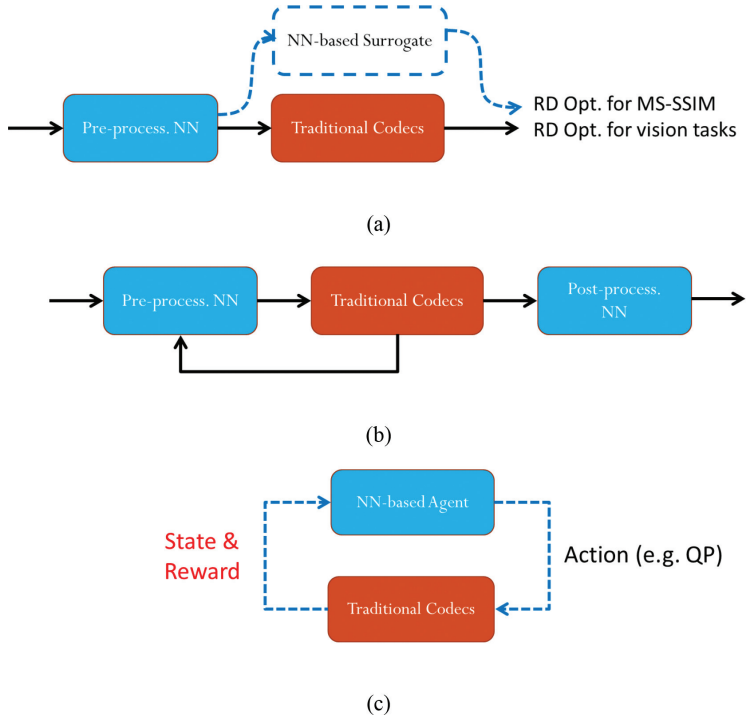

Figure 1: DL-assisted compression. (a) Using neural networks (NN) for pre-processing to repurpose conventional codecs. (b) Using neural networks for pre-processing and/or post-processing. (c) Using reinforcement learning to train a NN-based agent for video encoder control.

high bit-rates (see Figure 3a). In terms of MS-SSIM, the learned image codec shows a much higher MS-SSIM than VVC Intra (see Figure 3b); similarly, most learned video codecs have achieved better MS-SSIM than HEVC under the Low-delay P condition (GOP = 12), and also better than VVC at high bit-rates.

For image coding runtime, Figure 4 compares the encoding and decoding runtimes between an HEVC-based BPG encoder [7] and a learned codec [5]. For the learned codec, the encoding and decoding runtimes are about the same, which has to do with the symmetric architecture of the chosen autoencoder. For BPG [7], one can see that the traditional codec takes more time for encoding, mainly due to the rate-distortion optimization (RDO) process adopted by the encoder. For decoding time, the learned codec is about 8.6 times higher than the traditional BPG. For video coding, Figure 5 compares the encoding and decoding runtimes between X265 (an encoder for HEVC) [43] and a learned codec (DVC) [36]. For the learned codec, the encoding time is 1.6 times more

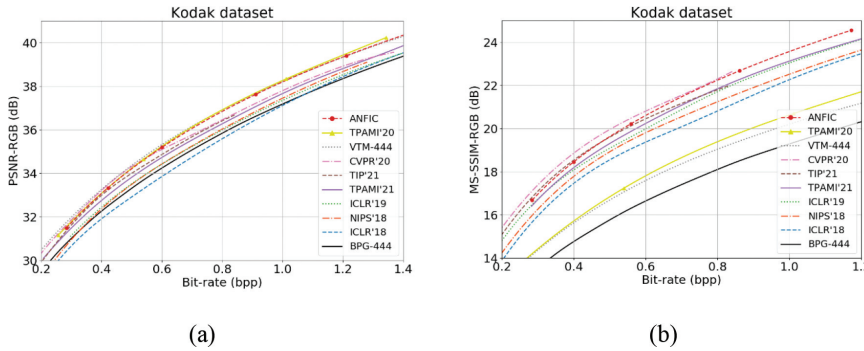

Figure 2: The rate-distortion comparison of end-to-end learned image codecs: BPG [7] (shown as BPG-444) and VVC Intra [44] (shown as VTM-444) in terms of (a) PSNR-RGB and (b) MS-SSIM-RGB. The learned image codecs include ANFIC [15], TPAMI'20 [38], TPAMI'21 [19], ICLR'19 [26], NIPS'18 [39], and ICLR'18 [5].

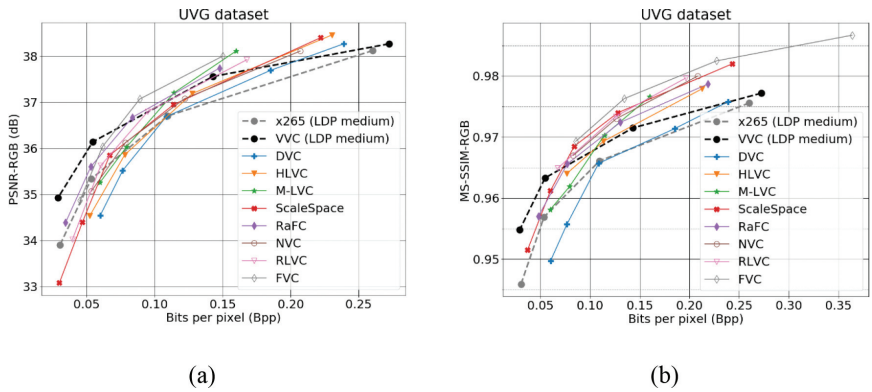

Figure 3: The rate-distortion comparison of end-to-end learned video codecs: x265 [43] and VVC [44] in terms of (a) PSNR-RGB and (b) MS-SSIM-RGB. The learned codecs include DVC [36], HLVC [52], M-LVC [29], ScaleSpace [1], RaFC [20], NVC [32], RLVC [53], and FVC [21]. Except for HLVC [52], which incorporates bi-prediction, all the other competing methods were evaluated under the Low-delay P condition.

than the decoding time due to the flow estimation network. In contrast, the encoding time of X265 is significantly higher (18×) than the decoding time. The reason is attributed to the fact that the RDO process of the traditional encoder adapts the coding modes according to the characteristics of every video frame. Obviously, the learned codec has not reached the same level of encoding optimization; there is still space for further improvement in the future. In terms of decoding, the learned decoder is still way more complex than the traditional one.

2.2.3 Open Issues

[Wen-Hsiao] For learned codecs, there are still several open issues to be addressed. (a) *Complexity* is way too high because heavy networks are used. (b) *Multi-rate encoding* is another issue. At present most experts still use separate models to achieve different bit rates. (c) *Rate control for end-to-end learned codecs* is an underexplored area. There has been little research on related topics. (d) *Encoder optimization* is certainly an issue; how to adapt

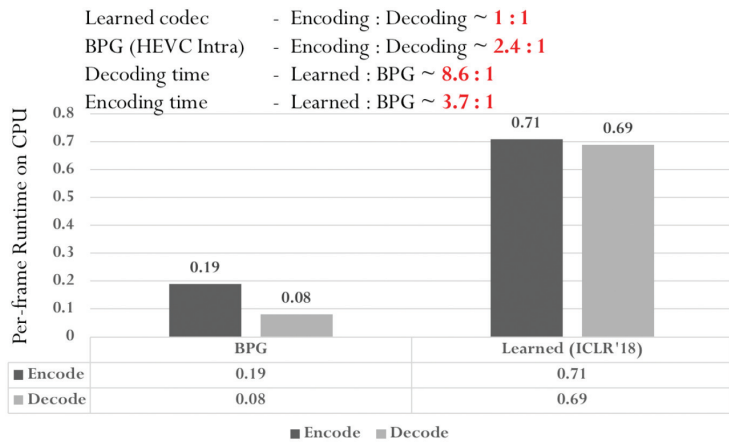

Figure 4: Comparison of encoding and decoding runtimes between BPG [7] and ICLR'18 [5]. The evaluation was done on CPU i7-9700K with 16GB RAM.

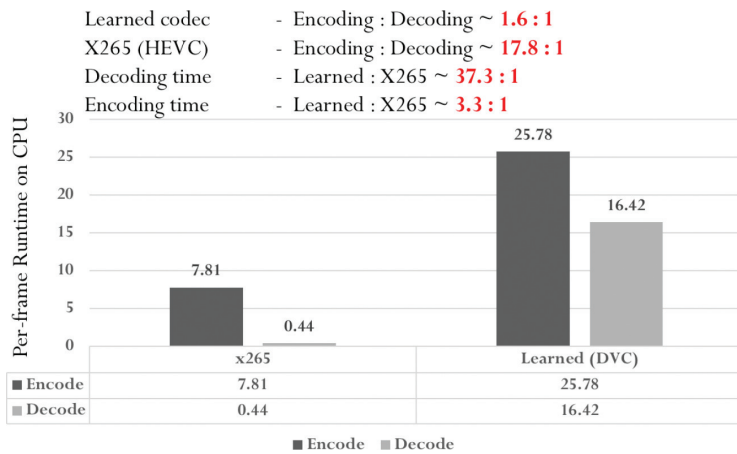

Figure 5: Comparison of encoding and decoding runtimes between x265 [43] and DVC [36]. The evaluation was done on CPU i7-9700K with 16GB RAM.

the encoding process to every input video frame or image needs more work for learned codecs. (e) *Generalization* is another issue especially on how to make learned codecs less dependent on training data. (f) *Lossless or nearly lossless coding* is also an issue as most learned codecs are unable to achieve such performance.

2.2.4 Visual Quality for End-to-end Learned Codec

[**Hsueh-Ming**] In recent years, the deep neural network (DNN) shows outstanding performance for pattern extraction, storage, and retrieval, particularly when the DNN models are trained with proper training datasets. Such properties (pattern extraction and retrieval) are important to the success of the end-to-end learned codecs to compress and reconstruct the images. The learned codecs usually have good subjective image/video quality. This is known even at the early stage of DNN-based image codec development. For example, in a 2018 paper [2], one can see a comparison between a learned codec and a BPG codec (H.265 single image compression) at very low bit-rate of around 0.035–0.039 bpp (Figure 6 is cited from [2]). Subjectively, the resulting pictures produced by the learned codec clearly have better visual quality; however, the PSNR value of the learned codec is not higher. The learned codec re-synthesizes the image to produce good subjective quality but it may not be a faithful reproduction of the original image.

Figure 6: Figure 6 (cited from [2]): Visual comparison of images produced by a DNN-based scheme (ours) in [2] and the standard codec H.265 (BPG). BPG (Better Portable Graphics) is essentially the still image codec of H.265. The reconstructed image of the DNN-based codec (ours) is sharper and has more realistic texture but it has a lower PSNR (MSE) value.

For the hot research areas, my observations are as follows. (1) There have been many publications in the last 4–5 years on intra-frame image coding. Up to now, leading intra-frame learned codecs (e.g., [11] and [27] in 2020) can achieve a better visual quality (MS-SSIM) than the best standard codec (H.266 or VVC) with comparable PSNR values. (2) However, inter-frame video

coding is much more complicated and thus the development of end-to-end DNN-based codec started only about 3 years ago, and there is still much work to be done on inter-frame coding, especially on video predictor or interpolator (e.g., to extract the pattern of object movement). Therefore, learned video coding system has a lot of room for improvement. (3) *There is a need for a well-accepted visual quality assessment method for learning-based codecs because they produce good subjective image quality but lower objective-based quality such as mean-squared error.*

3 The Trend of Deep-Based Video Coding

The second question posted to the panel is "Will deep-based coding become the main trend in video coding? Yes or No? Why?"

3.1 General Trend and Issues on Network Depth

[Dong] *Deep learning will pay an important role in future video coding standard, whether in DL-assisted, DL-based end-to-end, hybrid, or in other ways.* Google and NYU have published three important works [4, 6, 39] on deep learning-based image compression. Currently learned-based codec works in some cases has have achieved comparable or even better results than VVC/H.266 [8].

[Jiaying] One of the challenges on learning-based codec is the complexity due to the depth of the network. *It is possible that we do not need a very deep network, depending on applications.* Based on our observation in the deep learning loop filter, we achieved a 10% BD rate improvement [45]. We have also seen not-so-deep end-to-end learned codecs that are faster than VVC, where encoder and decoder can be implemented with only four convolutional layers, which is lightweight. We have also seen a one-layer method that learned from data but without an optimizer. We should also consider the trade-off between complexity and practicality.

3.2 Counter Viewpoint on Deep-Based Video Coding

[Jay] I have a different view on deep-based video coding. The current video coding framework has been fine-tuned for several decades. Now, people are willing to consider a very different approach. If there is no good alternative, deep-based coding is the one that receives attention. But, if there is an alternative, things could be different. If we do not understand why deep learning works, it would be difficult to come up with a competitive alternative. We have tried in developing a new video coding method that can capture the essence of deep learning yet it is not deep-based.

One key emphasis is green video coding, which means low power consumption. Deep learning is apparently not a green solution [55].

New video coding standards target at squeezing 50% bandwidth by 10× complexity. With the broadband infrastructure such as optical fibre, 5G, etc. in place, bit-rate reduction could be less critical. On the other hand, with more video Internet of things (IoT) devices being deployed, there is a great need to save power, and video coding with lower power consumption becomes important. For example, VVC is excellent in optimizing the rate-distortion trade-off. But, we may need to work furthermore to lower its complexity, say, 1% of today's VVC complexity while keeping the same level of rate-distortion performance. Since the deep-based coding technique deviates from this green principle, I am less enthusiastic.

3.3 Understanding of Why/How Things Work and Visual Quality

[Gary] Understanding why things work is important; knowing the coherent theory can help us move beyond what we know today. This is the natural role of academia. Better measurement of visual quality will also help us go to the next step. We are still at the beginning stage of being able to measure quality. *PSNR is not the thing, but alternatives so far are not working as much better as we would like.* We need to build quality into the optimization algorithms. Quality measurement and its optimization would be an important trend.

4 Visual Quality Assessment and Its Impact

The third question is "Visual quality assessment is a hot research topic in academia. Many papers have been published. Will this effort have a real impact on video coding standardization?"

4.1 Focusing on Visual Quality Assessment for New Media

[Jay] Compression artifacts are not general artifacts. They are special artifacts coming from image/video coding. There is little interest in measuring artifacts at low-bit-rate video since its quality is too poor to attract people. On the other hand, when the bit rate is high, the differences between different quality metrics are small. I have no problems in using PSNR as a metric to measure the quality of high-bit-rate image/video; the value of developing new quality metrics is questionable. The main battlefield should lie in the mid-bit-rate image/video.

For supervised quality assessment, the main challenge is how to do supervision. It is expensive to conduct a large-scale human subjective test. As to full-reference or no reference quality metrics, it is common to have full-reference at the encoder but no-reference at the decoder. Due to the vast diversity of

the decoder environment, it is difficult to calibrate the decoder environment. Due to these variabilities, there are many research problems and we have seen many papers published on visual quality assessment.

I collaborated with Netflix in developing the VMAF (Video Multimethod Assessment Fusion) full-reference video quality metric [46] from 2013 to 2015. Since Netflix has a closed user community, the company can adopt VMAF in its video delivery system and, furthermore, make it an open-source platform. For this reason, VMAF becomes a de facto video quality assessment standard used by other companies. This is nevertheless a rare case.

Although there are many published papers on image/video quality assessment, it is difficult to have an actual impact. *In my opinion, we should pay more attention to quality assessment of new media such as stereo video, 360-degree video, and AR/VR.* Clearly, PSNR is not working for them. There is an urgent need for better quality metrics. There are more research opportunities.

[**Shan**] Besides providing objective measurement of user visual experience, visual metrics are also used in video encoding such as in the mode decision process. PSNR or MSE has been used for tuning video encoders for decades and proved to be effective. On the other hand, conventional visual metrics such as PSNR are not suitable for evaluating visual quality of some emerging media formats, for example, point cloud and light field, etc. The demand for developing new metrics for those is rising.

4.2 Experience from the Standardization Committee

[**Gary**] So far in the international standardization committees we have not seen using PSNR lead to us making wrong decisions in video format designs. We were sometimes confused when we tried to use other metrics, and on the whole they usually all point in the same direction. More recently we have seen JPEG-XL getting quite different results for different metrics, and that design has not been studied fully yet. *So far we have not seen a huge impact on standardization from improved metrics, but there is a huge potential there and it could change the whole game.*

4.3 Subjective Quality with an End-to-End Learning Based Codec

[**Hsueh-Ming**] For the entertainment video, the target is subjective quality. As shown in [27], in terms of PSNR, an end-to-end learning-based codec produces comparable values to that of VVC. However, in terms of MS-SSIM (probably a better subjective quality metric than PSNR) the end-to-end learning-based codec trained to optimize MS-SSIM performs much better than VVC (Figure 7 is cited from [27]). On the other hand, similar to many other DNN-based schemes, *the end-to-end learning-based codec may fail to produce*

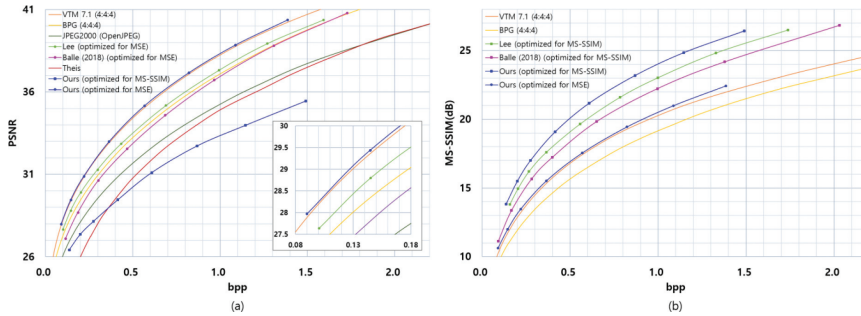

Figure 7: (cited from [27]): The rate-distortion curves of the proposed scheme (ours) in [27] versus the other image coding schemes including H.266 (VTM). VTM is the standard committee software for H.266/VVC. PSNR is the distortion metric in (a), and MS-SSIM is the distortion metric in (b). It is clear that the learned image codecs including the ones proposed in 2018 have superior MS-SSIM values than VTM.

a good approximation to the original image in certain cases. Hence, this may become a concern in certain applications such as medical imaging.

5 The Role of Academia

The fourth question posted to the panel is "Is there a role for academia in developing the new generation of video coding technology (or standard)?"

5.1 Possible Research Areas for the Academia

[Wen-Hsiao] *Academia is more flexible in exploring new and innovative ideas, which may seem to be more long-termed and less mature from the industry viewpoint.* If one looks at the contributors of the challenge on learned image compression (CLIC) at CVPR from 2018 to 2021 (see Figure 8), most contributors were initially from the academia. Then more industrial contributors jumped in. More recently, we saw much joint collaboration between industry and academia. Hence there is ample room for the academia to collaborate with the industry to contribute to the future of video coding.

[Hsueh-Ming] JPEG-AI issued a "called for evidence" last year, and the results were reported at the MMSP 2020 conference [22]. There were only six responses for this call for evidence (challenge). In the end, the evaluation was based on human observers. After that, JPEG-AI formed an ad-hoc group focusing on the possibility of standardizing it using an end-to-end learning-based coding approach. Its scope and framework go beyond human visualization, and thus it also targets at image processing and computer vision tasks; that is, the output may be consumed by machines instead of humans. In other words,

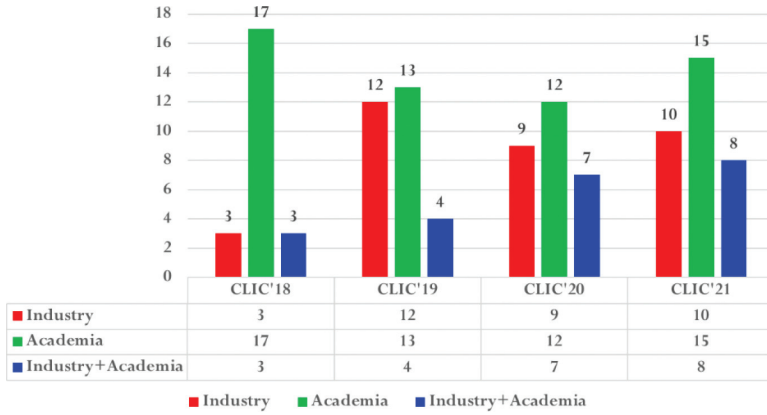

Figure 8: The distribution of CLIC contributors from 2018 to 2021.

the compressed images may be later processed by image processing tasks such as super-resolution or denoising, but they may also be used for computer vision tasks such as classification, object detection/recognition, and semantic segmentation. *Therefore, the goal is not limited to achieving good subjective quality for human viewers but also for producing good results for image processing and computer vision.* This standard activity is currently ongoing.

[**Gary**] JPEG-AI is also looking for not just compression but for what else can be done with a neural network-based coding approach, and images may be less complicated to deal with than video.

5.2 Understanding Why Things Work

[**Shan**] Deep learning has demonstrated its effectiveness in solving a variety of computer vision and image processing problems, and thus there has recently been increased enthusiasm to apply deep learning to video coding. One observation is that some people use neural networks as a black box and seem to think that good results could be obtained easily by tuning parameters. It is worthwhile to remind that any research, if to succeed, needs in-depth understanding about fundamentals. *Technology development needs healthy ecosystem to foster and nourish.* Academia and industry collaboration plays an important role.

5.3 Collaboration with Industry

[**Jay**] *I would like to emphasize the importance of collaborations between industry and academia.* Academic people can write a paper whenever they have a new idea. People may publish 100 papers based on various ideas at a

low cost since most of them involve computer simulation only. If one of them is eventually adopted by industry, it is a great achievement.

6 Advice for Graduate Students

The fifth and last question posted to the panel is "What advice do you have for grad students who would like to pursue video coding research?"

6.1 Areas and New Framework

[**Dong**] First, we encourage graduate students to *know not just deep learning or compression but need to know both areas*, as many areas are interdisciplinary. Second, we encourage students to *work on new frameworks*. For example, in our DVC paper [36, 37], initially we worked on the pixel space to perform motion estimation and extract optical flow information, but now look into the feature space by using deformable convolution. This CVPR 2021 work [21] is a totally different framework and results are improved significantly. We also worked on deep compression on point-cloud sequences [41], and we were one of the earliest groups to work on this direction as well. Third is to try to solve fundamental problems on what is beyond a black box. It is hard to try to explain why it works for image/video compression. We do not have much progress on this but it is important. Fourth, AI and deep video compression will still have a future because eventually we will have an AI chip which can be used for video compression, face recognition, and many other applications. It is different from before like wavelet. We believe eventually deep learning based approaches will replace DCT. Wavelet-based approaches cannot be supported by the existing chips, and new chips need to be designed to support wavelet-based approaches. However, the current chips designed for face/image recognition can be readily used to support deep learning-based image/video compression. When the performance of deep video compression (DVC) approaches is comparable with the DCT-based standards, it is likely that these DVC technologies will be quickly supported in the existing chips.

6.2 Entry Barrier in Industry –Deep Learning versus Video Coding

[**Jay**] The cultures in China and U.S. are different. In China, if one has a professor who is an expert in coding, his/her students will follow and do research along the same direction. It is a free market in the U.S. If a professor works on a topic that is not fashionable, he/she may not be able to attract students. There are also other reasons, for example, the entry barrier. Machine learning tools such as Python and TensorFlow have a lower entry barrier. Students may claim that they know deep learning well in three months by getting familiar

with these coding tools. In contrast, compression has a high barrier. It may take 1 or 2 years for a new person to get familiar with C codes and coding reference codes. *A high barrier is a protection since people cannot acquire the same skill in a short period and experts are difficult to replace.* Deep learning programmers are easier to replace since low barrier is not only good for them but also good for others.

Many companies look for people in machine learning, yet the number of applicants is also high. Although there are fewer job opportunities in video coding, the number of applicants is fewer. From the demand side, the multimedia industry does need coding engineers. The supply does not match the demand. Many students are career driven. The supply could be more than the demand in deep learning.

6.3 Knowledge of Signal Processing and Deep Learning

[**Shan**] Conventional or non-learning-based video coding is still the backbone of video communication today. For example, in real-time video conferencing, video coding provides the core capability while deep learning may be used in some add-on features, for example, virtual background and face touch up. In the last 3–4 years I have not had a shortage with resumes or job applicants from deep learning background, but candidates with video coding expertise are not many. The supply of video coding engineers seems to be extremely low, compared with the high market demand. A good video coding engineer can almost guarantee to have multiple offers from top-notch companies to choose. Video coding is built on signal processing theory. *One with both solid signal processing background and in-depth knowledge of deep learning would obviously be highly desirable.* One reminder for students working in deep learning subjects and may consider pursuing an industry job is that the deep learning algorithms we develop often need to run real-time on client and mobile devices. This is where Python and other scripting languages would face challenge and good programming language (e.g., C/C++) skills could help.

6.4 The Importance of Fundamental Knowledge

[**Jiaying**] I remember when I was an exchange Ph.D. student in Jay's group at USC in 2007–2008, Prof. Kuo told us that during that time people who work on graphics were just like standing in the Manhattan, video coding people were on the east side, but the bioinformatics guys were on the west side. It has shown some situations for different research areas that time. After my graduation in 2010, I switched my research from video compression to low level computer vision. My group (students) were back, working on video coding again in 2018, because we found some key technologies in computer vision, especially at the low level, which could be naturally utilized in the video

compression scheme. It helps us get accurate predictions or fit more complex relationships.

We also feel very excited about video coding for machines (VCM), where our group has published several papers last year [12, 18] to discuss potential paradigm. This brings signal processing and computer vision together, and tries to use deep neural network to bridge the two domains.

All these new topics have broadly extended the video coding research scope. This year as area chair for CVPR I took care of 28 papers and found two-thirds of which were related video coding. Our group has three out of six students working in this area. *I also agree that fundamental knowledge is very important, although my students are very familiar with tuning parameters, our knowledge of signal processing methods can inspire us a lot.*

7 Questions and Answers

Questions from the audience and responses from the panellists are presented in this section.

7.1 Video Coding in Decentralized Computation

The first question: Is there a future for video coding in decentralized computation?

[**Gary**] Decentralization for video coding means lots of interacting optimizations. *Decentralizing things and making them parallel ends up in an inability to effectively optimize a complex system like a video coder.* One can parallelize across different pictures, but in a classical design there is a limit on how much one can parallelize.

7.2 Application/Market-Driven versus Technology-Driven

The second question: In early days, video coding standards were mainly driven by applications, with mass market. With applications come characteristics which can be used to optimize things. Later, standards were more technology-driven. Standards like H.264 and H.265 are successful and have an impact because they are able to achieve 50% or more improvement. On the application side, people mentioned high resolution and compression for machines, but what are the new characteristics in these new applications that can be use to further technology improvement? Also, are there mass market applications that can get lots of people interested and involved? On the technology side, can using technology like deep learning based approach achieve like 30%, 40%, or more improvement? With that level of improvement then we have an impact and

be successful and people are willing to invest money; and for the academia, only with high improvement it is worthy of a PhD work.

[**Gary**] On the market and requirement questions, for some of the things that we are working on that are interesting, we really do not know how big the market is. This also complicates the requirements analysis. Without knowing how big the market is, it is hard to see what the requirement is to fill market needs. On gain, 3% is a lot in video coding now, and is publishable. The way we got to VVC was primarily with many 1% or 3% gains that collectively turned into a 50% gain.

[**Shan**] *The market needs for video coding have never faded.* During the pandemic the whole world relied on video communication for work, school, and everything. The bandwidth pressure was (and is still) on global networks and regional disconnections happened from time to time during the peak months. Furthermore, emerging applications such as cloud gaming, VR, freeview and immersive video require much higher bitrate than conventional video applications. Hence the need of video compression is always there.

[**Wen-Hsiao**] With regard to applications and rate savings, we need to think about the visual quality metrics. For example, if we use MS-SSIM, learning based video compression already achieved significant gain over traditional compression methods. *I believe learning based video compression opens up new applications because the objective function, if differentiable, can be optimized for different purposes, which cannot be easily achieved with traditional codecs.* It is also worth noting that JPEG AI [23] is seeking specifically learning-based image compression technologies for both human perception and computer vision tasks.

7.3 Training Data – Crucial for Learned Based Codec

[**Wen-Hsiao**] On the other hand there is a danger in using learning-based compression. We need to be very careful about the training data. *Training data have a crucial effect on its coding performance.* For example, we do not normally use white noise as training data. Figure 9 illustrates that in coding white noise, the rate-distortion curve of the end-to-end learned codec actually plateaus after some bit-rate and is much worse than that of the traditional codec, for example, BPG [7]. This is because an autoencoder implements a non-linear transform; when training data are not well represented, the codec learns an incomplete basis and hence cannot represent any type of signals. On the other hand, discrete cosine transform (DCT) often used in traditional codecs realizes a complete basis that can represent any type of signals when enough bits are given.

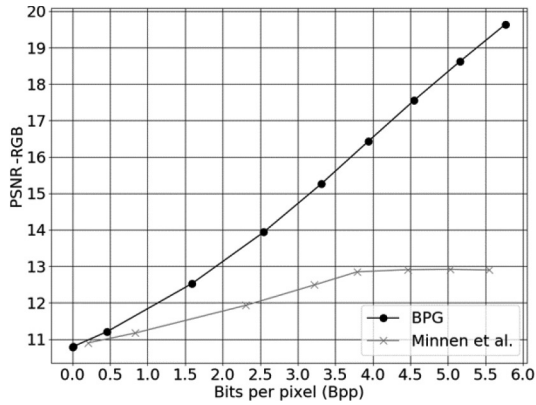

Figure 9: Rate-distortion comparison of BPG [7] and Minnen *et al.* [39] in coding white noise.

The third question: About generalization capability of learning based models, if the test sequences are completely unseen from training data, then the visual quality is quite bad. Do you have similar experience?

[Wen-Hsiao] Yes, training data really matter. There are some new types of autoencoders; for example, a few early attempts [14, 15, 38] use invertible or reversible neural networks, which can actually represent any type of signals and make the learned codecs less sensitive to training data. We have done some studies [15] and found that there is a potential for these kinds of networks to be used for compression, but there is still much work to be done.

[Dong] If we have good training data we can achieve significant improvement over the state-of-the-art learning-based codec and may outperform the best conventional codec. In earlier stage we used datasets Vimeo-90k [51] for low-level computer vision tasks that are not targeted for video compression, but it is important for industry and academia to collect large scale datasets to train video compression models. We also worked on how to reduce data distribution mismatch between training and testing data (see our ECCV 2020 work [35]). For different videos we can fine tune the model parameters and meanwhile fixing the decoder. We have worked on visual domain adaptation for many years (see our survey work [54] for more recent progress in this area); when training and testing data come from different domains, the model cannot generalize well on the testing data. Perhaps it is more difficult to solve this problem in video compression; perhaps we need to collect more data to represent different distributions as much as possible, to partially solve this problem.

[Shan] Having realized the importance of video training data for research of learning-based video coding technologies, Tencent built a video dataset

(TVD) which is free for research and standardization usage [50]. This video dataset has been contributed to JVET for its exploration activities on neural network-based video coding (NNVC) [3, 33].

7.4 Image Coding Job Market

The fourth question: Are image coding people using deep learning desirable in the job market?

[**Gary**] Students should have a full background of classical technical knowledge and understand how things work.

[**Shan**] Industry welcomes students with in-depth knowledge and understanding of both image coding and deep learning. Proficiency in at least one programming (in additional to scripting) language would be a big plus.

7.5 Reducing the Complexity and Different Codec Framework

The fifth question: What is the limit of video compression? How much more do we have to go? Or is it the time for us to focus on reducing the complexity of the codec?

[**Jay**] Deep-based image/video coding has stimulated lots of interests in recent years. Yet, we need to understand the principle behind deep learning, and do something that is transparent and explainable. *I am less concerned with coding efficiency but with coding complexity. I hope we can reduce the complexity of VVC while maintaining its coding efficiency with a novel framework.*

The sixth question: From H.261 to H.266 (about 30 years), we have been following the same hybrid block-based codec framework (transform, quantization, motion compensation, etc.). Does the standardization committee have a plan to look at a totally different framework, not to target on further reducing compression ratio, but perhaps on reducing the complexity by at least 50%?

[**Gary**] We would be welcoming if we have approaches not doing incremental tasks in the hybrid block-based framework. *The reason that we have used this block-based model till today is because it seemed to work very well.* We have tried to investigate other approaches but this was what we ended up with. We do not want to abandon things on the table if we do not see a clear path in a different direction. About how much more gain is there, I think no one knows. We will keep trying until we run out of ideas for how to do better or come up with a theory that shows we cannot do better. There is potential; part of that is using alternative quality metrics, or alternative architectures. Even in using classical signal processing methods we seem to be able to squeeze out a bit more gain; less than a year after VVC we already know how to do about 14% better with ordinary methods, with some complexity problems.

7.6 More on the Problems of Data-Driven Approach

The seventh question: Since the nature of training data makes a significant difference on the outcome and it is hard to fit into every kind of cases, shouldn't machine learning approaches be more special purpose or application dependent, not trying to fit the training data to generally every case?

[Jay] *There are two powerful ideas behind deep-based image/video coding. One is the hierarchical representation of images.* Most video codecs are fundamentally a single-scale method although there are blocks of variable sizes. If we can leverage the hierarchical representation well, there will be a significant gain. *The other is the data-driven representation.* One well known example is vector quantization (VQ). We may learn the distribution of image samples. Deep learning can yield a good coding gain if training and testing images are correlated. Similarly, the learned codebooks of VQ do depend on the training image set.

My lab is working on green coding to compete with the deep-learning coding nowadays. These are tools of consideration. I spent 6–7 years in understanding deep learning and we are now ready for a new paradigm. In my opinion, the black box approach will not last long and we cannot build the knowledge on the black box. We need things to be more transparent and modularized. Mathematical tools such as linear algebra, probability, statistics, optimization, and information theory should be the foundation for future image/video coding techniques.

[Gary] The problem of over-fitting for training data for neural networks is not a new one. We used to study VQ, but the same over-fitting problem occurs there if we do not have a robust training dataset or general capability. *Hence we need a generalized approach.*

[Shan] We are still in the exploration stage of applying deep learning in video coding. While impressive research progress has been made, many practical challenges need to be solved before it can be used in real-world products.

7.7 Graph Signal Processing

The eighth question: For new media how to represent data is the core problem. Is the combination of graph signal processing and neural network feasible to deal with coding task?

[Jiaying] Graph signal processing is very hot but data dependent, for specific applications like point cloud and other applications it can be very useful. But we are not sure if it is good for general case video coding. We adopt video coding for machines (VCM) in some computer vision tasks with very sparse feature representations. Points and edges or other feature maps may be good

for graphs; and adding more information like residues to key frames can be used to reconstruct video information pixel-wise. *Graph signal processing might be useful to try in some new areas like VCM.*

7.8 Generative Models

The ninth question: How would generative models like generative adversarial network (GAN) play in video coding?

[**Dong**] GAN-based approach can work reasonably well for low bit-rates, but for high bit-rates it may not work that well. For image compression, the pipeline from Google's group [4, 6, 39] achieves promising results, which are generally better than those of GAN-based approaches. For video compression, perhaps we should follow the hybrid compression approach, with residue compression and motion compression [36, 37]. If you have low bit-rates maybe GAN-based approach can work to some extent.

[**Hsueh-Ming**] Two or three years ago, there were a few papers on GAN-based approach to compress images, for example, [2]. More recently, the autoencoder approach seems to be more popular. *It is observed that the autoencoder or CNN based codecs can do better for most pictures and at the lower bit-rates.*

8 Conclusion

[**Nam**] In summary, from the panel the future of video coding seems to be very bright. There are many more challenges ahead for researchers to deal with and solve. The key points are summarized here.

1. *Hot Emerging Areas.* The standardization committee is looking at a dual-track approach, the traditional hybrid block-based track (which has been doing very well) and a track using neural network. The advent of new technologies is often motivated by emerging applications (such as immersive applications); hence we expect that volumetric video, virtual reality, cloud gaming, and low power edge computing will have high demand. Another emerging area is video coding for machines (VCM). Most experts acknowledge the role of deep learning in the future of video coding and the main challenge is its high complexity and the need for a well-accepted visual quality assessment method.

2. *The Trend of Deep-based Video Coding.* On this issue, most experts feel that deep learning will play an important role in the future video coding standard, whether in DL-assisted, DL-based end-to-end, hybrid, or in other ways. It is possible that we do not need a very deep network for some applications, to reduce complexity. There is also a counter

viewpoint pointing out that a black-box approach will not grow much further and there should be an alternative, especially with the need to focus on reducing complexity and power consumption instead of increasing coding efficiency (i.e., green video coding).

3. *Visual Quality Assessment.* PSNR seems to work pretty well from the standardization committee's viewpoint for the traditional approach. End-to-end learned codecs show better results on metrics like MS-SSIM and subjective ones, although not so much on PSNR; there are many works devoted to developing other metrics. For learning-based approach, there is the need for a well-accepted visual quality assessment method, and the main battlefield could lie in the mid-bit-rate range. The panel also feels that it is important to develop the right metrics for new media such as point cloud, 360-degree video, stereo video, AR/VR, etc. In addition, the goal is not limited to achieving good subjective quality for human viewers but also for producing good results for machine use such as in many computer vision and image processing tasks.

4. *The Role of Academia.* Academia is more flexible in exploring new and innovative ideas, which may seem to be more long-termed and less mature from the industry viewpoint. There is ample room for the academia to collaborate with the industry to contribute to the future of video coding, this collaboration plays an important role. The panel agrees that understanding why things work is important, instead of obtaining good results just by tuning parameters, especially when discussing the role of academia.

5. *Advice for Graduate Students.* The panel feels that students should know not just deep learning or compression using signal processing, but need to know both areas, and the panel also encourages students to work on new frameworks. Deep learning has a lower barrier and can be easily mastered in a short time, whereas video coding has a higher barrier and needs a much longer time to master, hence leading to better job security. Engineers in deep learning with the knowledge of Python and TensorFlow are more easily replaced compared to those in video coding with the knowledge of C or C++.

6. *Other Key Points at Q & A.* There are other new key points brought up during Q & A: (a) Decentralizing and making the coding process parallel ends up in an inability to effectively optimize a complex system like a video coder. (b) The market needs for video coding have never faded and will continue to be bright. On gain, 3% is a lot in video coding now, and the way we got to VVC was primarily with many 1% or 3% gains that collectively turned into a 50% gain. Learning-based video

compression opens up new applications because the objective function, if differentiable, can be optimized for different purposes. (c) Another major issue brought up was the importance of having the right training dataset as it is crucial for coding performance. Generalization is another issue, especially on how to make learned codec to be training data independent, it is important to reduce data mismatch between training and testing. (d) On graph signal processing, it may be useful to try it on some new areas like VCM. (e) On GAN-based approach, experts feel that it can work reasonably well for low bit-rates, but not so for high bit-rates. Autoencoder or CNN-based codecs may be better.

Audience interacted with questions and the panellists addressed them.

Acknowledgement

We had more than 130 participations on Zoom, which was a record high. The authors acknowledge the support of the APSIPA U.S. Local Chapter, the active participations of the audience, APSIPA and many individuals/groups who have helped to promote the panel, Yijing Yang and the University of Southern California for their technical support, Yifan Wang and Qingyang Zhou of the University of Southern California for their help in proofreading of the manuscript, and several people who provided advice and other support.

Biographies

Nam Ling received the B. Eng degree in electrical engineering from the National University of Singapore and the M.S. and Ph.D. degrees in computer engineering from the University of Louisiana at Lafayette, USA. He is currently the Wilmot J. Nicholson Family Chair Professor of Santa Clara University, California, USA, and the Chair of its Department of Computer Science and Engineering, since 2010. From 2002 to 2010, he was an Associate Dean for its School of Engineering. He has been an IEEE Fellow since 2008. His main research interests are in video/image coding, 3D/stereoscopic video/image, rate control, and the use of deep learning in image/video.

C.-C. Jay Kuo received his B.S. degree in electrical engineering in 1980 from National Taiwan University and received his M.S. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology, USA, in 1985 and 1987, respectively. He is currently the William M. Hogue Professor in electrical and computer engineering, a Distinguished Professor of electrical engineering and computer science, and the Director of the Multimedia Communications Laboratory at the University of Southern California, USA. He is a Fellow of

the National Academy of Inventors (NAI), IEEE, AAAS, and SPIE. His main research interests are in multimedia data compression and computing.

Gary J. Sullivan received the B.S. and M.E. degrees in electrical engineering from the University of Louisville, Kentucky, USA, in 1982 and 1983, respectively, and received the Engineer and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, in 1991. He is currently a Video and Image Technology Architect with the Research division of Microsoft Corporation, Washington, USA. He has been a longstanding Chair or Co-Chair of various video and image coding standardization activities in ITU-T VCEG, ISO/IEC MPEG, ISO/IEC JPEG, and in their joint collaborative teams such as JVT, JCT-VC, and JVET, which developed the H.264/AVC, HEVC/H.265, and VVC/H.266 standards, respectively. He is a Fellow of IEEE and SPIE.

Dong Xu received the B.Eng. and Ph.D. degrees in electronic engineering and information science from the University of Science and Technology of China, in 2001 and 2005, respectively. Currently he is the Chair in Computer Engineering and Australian Research Council (ARC) Future Fellow at the School of Electrical and Information Engineering, The University of Sydney, Australia. He is a Fellow of the IEEE and IAPR. His main research interests are in the areas of image and video processing, computer vision, and multimedia.

Shan Liu received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, USA, respectively. She is currently a Tencent Distinguished Scientist and General Manager of Tencent Media Lab. She served as Co-Editor of the H.265/HEVC SCC and H.266/VVC standards. Her research interests include audio-visual, volumetric, immersive and emerging media compression, intelligence, transport and systems. She is a Fellow of IEEE.

Hsueh-Ming Hang received the B.S. and M.S. degrees in control engineering and electronics engineering from National Chiao Tung University (NCTU), Taiwan, in 1978 and 1980, respectively, and received the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, New York, in 1984. He is currently a Professor with the Department of Electronics Engineering at National Yang Ming Chiao Tung University, Taiwan. He served as the Dean of the ECE College at NCTU in 2014–2017. He is a Fellow of IEEE. His current research interests include spherical image/video processing and deep-learning based image/video processing.

Wen-Hsiao Peng received his B.S., M.S., and Ph.D. degrees, all in electronics engineering, from National Chiao Tung University, Taiwan, in 1997, 1999, and 2005, respectively. He is currently a Professor with the Computer Science Department, National Yang Ming Chiao Tung University. He serves as the

Chair of the IEEE Circuits and Systems Society Visual Signal Processing and Communications Technical Committee. His research interests include learning-based video/image coding, multimedia analytics, and computer vision.

Jiaying Liu received her B.E. degree in computer science from Northwestern Polytechnic University, Xi'an, China, in 2005, and the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010. She is currently an Associate Professor and Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University, China. Her current research interests include multimedia signal processing, compression, and computer vision.

References

[1] E. Agustsson, D. Minnen, N. Johnston, J. Ballé, S. J. Hwang, and G. Toderici, "Scale-space Flow for End-to-end Optimized Video Compression," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, 8503–12.

[2] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative Adversarial Networks for Extreme Learned Image Compression," *arXiv:1804.02958*, 2018, 4.

[3] E. Alshina, S. Liu, W. Chen, F. Galpin, Y. Li, Z. Ma, and H. Wang, "Exploration Experiments on Neural Network-based Video Coding," July 2021, Document JVET-W2023, 23rd JVET meeting, online meeting.

[4] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational Image Compression with a Scale Hyperprior," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational Image Compression with a Scale Hyperprior," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[7] BPG image format, URL: https://bellard.org/bpg/.

[8] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, "Versatile Video Coding (draft 10)," Document JVET-S2001, 19th JVET meeting, online meeting, June 2020.

[9] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y. K. Wang, "Development in International Video Coding Standardization after AVC, with an Overview of Versatile Video Coding (VVC)," in *Proc. of the IEEE (Early Access)*, Vol. 1, 2021, 1–31.

[10] L. C. Chen, J. H. Hu, and W. H. Peng, "Reinforcement Learning for HEVC/H.265 Frame-level Bit Allocation," in *Proc. IEEE 23rd International Conference on Digital Signal Processing (DSP)*, Nov 2018.

[11] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics," *IEEE Trans. on Image Processing (TIP)*, 29, 2020, 8680–95.

[13] W. Gao, S. Liu, X. Xu, M. Rafie, Y. Zhang, and I. Curcio, "Recent Standard Development Activities on Video Coding for Machines," 2021, arXiv:2105.12653.

[14] L. Helminger, A. Djelouah, M. Gross, and C. Schroers, "Lossy Image Compression with Normalizing Flows," arXiv:2008.10486.

[15] Y. H. Ho, C. C. Chan, W. H. Peng, H. M. Hang, and M. Domanski, "ANFIC: Image Compression Using Augmented Normalizing Flows," *IEEE Open Journal of Circuits and Systems*, 2021.

[16] Y. H. Ho, G. L. Jin, Y. Liang, W. H. Peng, X. B. Li, and Y. K. Chen, "A Dual-critic Reinforcement Learning Framework for Frame-level Bit Allocation in HEVC/H.265," in *Proc. Data Compression Conference*, March 2021.

[17] J. H. Hu, W. H. Peng, and C. H. Chung, "Reinforcement Learning for HEVC/H.265 Intra-frame Rate Control," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018.

[18] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, "Towards Coding for Human and Machine Vision: A Scalable Image Coding Approach," in *Proc. of IEEE International Conf. on Multimedia & Expo (ICME)*, 2020.

[19] Y. Hu, W. Yang, Z. Ma, and J. Liu, "Learning End-to-end Lossy Image Compression: A Benchmark," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021, 1.

[20] Z. Hu, Z. Chen, D. Xu, G. Lu, W. Ouyang, and S. Gu, "Improving Deep Video Compression by Resolution-Adaptive Flow Coding," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[21] Z. Hu, G. Lu, and D. Xu, "FVC: A New Framework Towards Deep Video Compression in Feature Space," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] "JPEG AI Challenge on Learning-based Image Coding," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2020.

[23] "JPEG AI Second Draft Call for Proposals," in *ISO/IEC JTC 1/SC29/WG1 N92014*, July 2021.

[24] J. P. Klopp, K. C. Liu, L. G. Chen, and S. Y. Chien, "How to Exploit the Transferability of Learned Image Compression to Conventional Codecs," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 16165–74.

[25] W. Kuang, Y. Chan, S. Tsang, and W. Siu, "DeepSCC: Deep Learning-Based Fast Prediction Network for Screen Content Coding," *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, 30(7), 2020, 1917–32.

[26] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive Entropy Model for End-to-end Optimized Image Compression," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[27] J. Lee, S. Cho, and M. Kim, "An End-to-end Joint Learning Scheme of Image Compression and Quality Enhancement with Improved Entropy Minimization," *arXiv: 1912.12817v2*, 2020, 3.

[28] Y. L. Lee, Y. C. Chen, M. Y. Tseng, Y. H. Tsai, and W. C. Chiu, "Learning to Hide Residual for Boosting Image Compression," in *Proc. The British Machine Vision Conference (BMVC)*, Nov 2021.

[29] J. Lin, D. Liu, H. Li, and F. Wu, "M-LVC: Multiple Frames Prediction for Learned Video Compression," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, 3546–54.

[30] D. Liu, Z. Chen, S. Liu, and F. Wu, "Deep Learning-Based Technology in Responses to the Joint Call for Proposals on Video Compression with Capability Beyond HEVC," *IEEE Trans. on Circuits and Systems for Video Technology*, 30(5), 2020, 1267–80, DOI: 10.1109/TCSVT.2019.2945057.

[31] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep Learning-Based Video Coding: A Review and a Case Study," *ACM Computing Surveys*, 53(1), 2020, 1–35.

[32] H. Liu, M. Lu, Z. Ma, F. Wang, Z. Xie, X. Cao, and Y. Wang, "Neural Video Coding Using Multiscale Motion Compensation and Spatiotemporal Context Model," *IEEE Trans. on Circuits and Systems for Video Technology*, 31(8), 2021, 3182–96.

[33] S. Liu, A. Segall, E. Alshina, and R. Liao, "Common Test Conditions and Evaluation Procedures for Neural Network-based Video Coding Technology," Jan 2021, Document JVET-U2016, 21st JVET meeting, online meeting.

[34] Z. Liu, X. Yu, Y. Gao, S. Chen, X. Ji, and D. Wang, "CU Partition Mode Decision for HEVC Hardwired Intra Encoder Using Convolution Neural Network," *IEEE Trans. on Image Processing (TIP)*, 25(11), 2016, 5088–103.

[35] G. Lu, C. Cai, X. Zhang, L. Chen, W. Ouyang, D. Xu, and Z. Gao, "Content Adaptive and Error Propagation Aware Deep Video Compression," in *Proc. of the European Conference on Computer Vision*, 2020.

[36] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An End-to-end Deep Video Compression Framework," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, 11006–15.

[37] G. Lu, X. Zhang, W. Ouyang, L. Chen, Z. Gao, and D. Xu, "An End-to-end Learning Framework for Video Compression," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, DOI: 10.1109/TPAMI.2020.2988453.

[38] H. Ma, D. Liu, N. Yan, H. Li, and W. F., "End-to-end Optimized Versatile Image Compression with Wavelet-like Transform," *IEEE Trans. on Pattern Analysis and Machine Intelligence (Early Access)*, 2020, 1.

[39] D. Minnen, J. Ballé, and G. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, 10794–803.

[40] Z. Pan, P. Zhang, B. Peng, N. Ling, and J. Lei, "A CNN-Based Fast Inter Coding Method for VVC," *IEEE Signal Processing Letters*, 28, 2021, 1260–4.

[41] Z. Que, G. Lu, and D. Xu, "VoxelContext-Net: An Octree Based Framework for Point Cloud Compression," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] J. Shi and Z. B. Chen, "Reinforced Bit Allocation Under Task-driven Semantic Distortion Metrics," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Oct 2020.

[43] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, 22(12), 2012, 1649–68.

[44] VVC official test model VTM, URL: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftwareVTM/tree/VTM-9.0.

[45] D. Wang, S. Xia, W. Yang, and J. Liu, "Combining Progressive Rethinking and Collaborative Learning: A Deep Framework for In-loop Filtering," *IEEE Trans. on Image Processing*, 30, 2021, 4198–211.

[46] Wikipedia, URL: https://en.wikipedia.org/wiki/Video_Multimethod_Assessment_Fusion.

[47] M. Xu, T. Li, Z. Wang, X. Deng, R. Yang, and Z. Guan, "Reducing Complexity of HEVC: A Deep Learning Approach," *IEEE Trans. on Image Processing (TIP)*, 27(10), 2018, 5044–59, DOI: 10.1109/TIP.2018.2847035.

[48] X. Xu and S. Liu, "Overview of Screen Content Coding in Recently Developed Video Coding Standards," *IEEE Trans. on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2021.3064210.

[49] X. Xu and S. Liu, "Recent Advances in Video Coding beyond the HEVC Standard," *APSIPA Trans. on Signal and Information Processing*, 8, 2019, e18, DOI: 10.1017/ATSIP.2019.11.

[50] X. Xu, S. Liu, and Z. Li, "Tencent Video Dataset (TVD): A Video Dataset for Learning-based Visual Data Compression and Analysis," 2021, arXiv:2105.05961.

[51] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video Enhancement with Task-oriented Flow," *International Journal of Computer Vision*, 127(8), 2019, 1106–25.

[52] R. Yang, F. Mentzer, L. V. Gool, and R. Timofte, "Learning for Video Compression with Hierarchical Quality and Recurrent Enhancement," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, 6628–37.

[53] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for Video Compression with Recurrent Auto-Encoder and Recurrent Probability Model," *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 15(2), 2021.

[54] J. Zhang, W. Li, P. Ogunbona, and D. Xu, "Recent Advances in Transfer Learning for Cross-dataset Visual Recognition: A Problem-oriented Perspective," *ACM Computing Surveys*, 52(1), 2019, 1–38.

[55] Z. Zhou, "MCL Technology Outlook: Green Learning," 2020, URL: http://mcl.usc.edu/news/2020/12/27/mcl-technology-outlook-green-learning/.