Editorial: Introduction to the Issue on Deep Learning for Image/Video Restoration and Compression

I. Introduction

THE HUGE success of deep-learning–based approaches in computer vision has inspired research in learned solutions to classic image/video processing problems, such as denoising, deblurring, dehazing, deraining, super-resolution (SR), and compression. Hence, learning-based methods have emerged as a promising nonlinear signal-processing framework for image/video restoration and compression.

Recent works have shown that learned models can achieve significant performance gains, especially in terms of perceptual quality measures, over traditional methods. Hence, the state of the art in image restoration and compression is getting redefined. This special issue covers the state of the art in learned image/video restoration and compression to promote further progress in innovative architectures and training methods for effective and efficient networks for image/video restoration and compression.

In the following, we provide a short overview of the state of the art in learned image and video processing Section II. Section III introduces the articles in this issue. Finally, we provide the outlook for future directions in Section IV.

II. Overview of the State-of-The-Art

A. Image/Video Restoration and Super-Resolution

Many researchers reported results that exceed the state of the art in image/video restoration and SR by a wide margin via supervised learning using pairs of ground-truth (GT) images/video and degraded or low-resolution (LR) images/video generated by known degradation models, such as bicubic downsampling. However, there is need for further research and room for improvement in at least three key areas: generalization of these results to real-world problems, efficiency of the solutions, and perceptual optimization of the results.

Most existing image restoration/SR methods assume a predefined degradation process from a GT image/video to a degraded/LR one, which can hardly hold true in real imaging with complex degradation types. To fill this gap, growing attention has been paid in recent years to approaches for unknown degradations, namely real-world SR or blind SR. We can roughly divide these methods into four groups: The first group of methods utilize an external dataset to learn a SR model well adapted to a large set of downsampling kernels, such as IKC [1], SRMD [2], or USRNet [3]. Another group of methods leverage the internal statistics within a single image derived from the degradation model, thus requiring no external dataset for training, like ZSSR [4] and DGDMLSR [5]. The third group resorts to implicit modeling, which defines the degradation process implicitly through a data distribution [6]–[8]. Particularly, these methods utilize data-distribution learning with Generative Adversarial Networks (GANs) [9] to grasp the implicit degradation model possessed within dataset, like WESPE [6], FSSR [10] and CinCGAN [11]. The last group directly builds real image datasets with input-output pairs for specific applications, such as DPED [12], RealSR [13], Zurich RAW-to-RGB [14] and DRealSR [15]. These new datasets make it possible to take advantage of existing supervised-learning methods in real-world applications.

For real-world applications besides dealing with data captured in uncontrolled or challenging conditions, the restoration/SR solutions need to be run-time, memory and energy efficient and to run on constrained hardware [16]. In a pioneering work, Ronneberger *et al.* [17] introduced U-Net, a widely adopted efficient neural design for image to image mapping. Since then tremendous progress has been achieved in Neural Architecture Search (NAS) [18]. However, while very efficient architectures have been optimized for tasks such as image classification (MobileNetV3 [19]), solutions are sought for image restoration/SR tasks as shown in the AIM 2020 challenge on efficient SR [20].

Another active area of research is perceptual image restoration and SR. Variations of the GAN architecture have been proposed for various low-level–vision tasks to obtain perceptually better results with more texture details. In a pioneering work, Ledig *et al.* [21] have proposed SRGAN model that could generate photo-realistic images in SR tasks. Ignatov *et al.* [6], [12] proposed to use perceptual losses and GANs to learn from paired or unpaired images to enhance the images from a smartphone camera to a DSLR target camera quality. In [22], the authors made the observation that there is a trade-off between fidelity (measured by full-reference metrics) and perception (measured by no-reference metrics). In the PIRM 2018-SR Challenge [23], ESRGAN [24] achieved state-of-the-art performance by improving the network architecture for the generator and loss functions. Benefiting from a learnable ranker, RankSRGAN [25] can optimize the generative network in the direction of any image quality assessment (IQA) metrics and achieves state-of-the-art performance. Although remarkable progress has been made, Gu *et al.* [26] reveal that existing IQA method cannot objectively

Digital Object Identifier 10.1109/JSTSP.2021.3053364

evaluate perceptual SR methods. In the newly-proposed IQA dataset, there is still a large gap between IQA methods and human labels.

B. Image/Video Compression

Much of the early work in applying learned models to compression focused on image compression, starting with approaches that solely learned non-linear transformations of image inputs without learning corresponding probability models [27]. Subsequent, more effective approaches jointly learned models of non-linear auto-encoders with models of the latent-variable probability distributions [28], [29]. The model coupling was done by minimizing the Lagrangian formulation of the rate-distortion loss, using the learned probability model for the rate estimation and the decoded reconstruction from the scalar-quantized latent variables for the distortion. Adding side information (“hyper priors”) to allow the probability models themselves to adapt locally resulted in a learned models that exceeded the performance of traditional image encoders (e.g., BPG) [30]. Extensions to the adaptive probability modeling include additional layers of side information about the probability models for the hyper-priors themselves, as well as autoregressive context models [31].

Much of the previous work in learned video compression addressed the problem of replacing parts of standard compression systems (e.g. HEVC) with learned components [32]–[38]. End-to-end optimized fully learned models have also shown promise. Some learned models based on uni-directional motion-compensation (low-latency) [39], [40] have outperformed H.264 in PSNR and HEVC in MS-SSIM. The best performance to date in fully learned low-latency video compression uses a learned scale-space motion-compensation model [41]. Recently, end-to-end optimized learned models based on bi-directional motion compensation have also shown competitive performance [42], [43].

C. Point Cloud Denoising and Compression

3-D point clouds (PC) are a collection of millions of points, where each point represents a specific 3-D coordinate of a scene, and associated color features. Raw 3-D PCs can be obtained by various acquisition devices or as output of 3-D reconstruction algorithms. Among many other applications, they are used as a scene representation for free view-point imaging and video. Compared to 3D meshes, they offer a simpler, denser, and closer-to-reality representation. However, raw 3-D PCs are typically contaminated with noise and outliers and the size of raw 3-D PC data is huge for storage and/or streaming. Hence, denoising and compression of 3-D PC are recent topics of significant interest.

Traditional methods for 3-D PC denoising typically rely on local surface fitting, local or non-local averaging, or on statistical modeling of the data and noise [44]. In contrast, deep learning offers a simple and universal data-driven approach for removing outliers and denoising 3-D PCs, corrupted with potentially very high levels of structured noise.

Among many existing traditional approaches to 3-D PC compression, the MPEG-3DG (3D Graphics group) has standardized two different frameworks: i) Video-based Point Cloud Compression (V-PCC), and ii) Geometry-based Point Cloud Compression (G-PCC) [45]. V-PCC considers compression of 2-D projections of 3-D PC to leverage the existing and future video compression technologies, as well as the established video eco-system. The reference model encoder achieves compression rates of 125:1 with good perceptual quality. G-PCC considers 3-D geometry-driven approaches to provide efficient lossless and lossy compression. Recently, deep-learning based data-driven methods have started to achieve state of the art 3-D PC compression performance.

III. Overview of the Articles

This special issue consists of 20 papers on recent advances in deep learning for image/video restoration and compression: 13 papers on image/video restoration and super-resolution, 5 papers on image/video compression, and 2 papers on point cloud processing. We provide a short introduction to these papers in the following.

A. Image/Video Restoration and Super-Resolution

The paper “Degradation aware approach to image restoration using knowledge distillation” is the first journal paper on application of knowledge distillation on image restoration. The authors present a new approach to handle image-specific and spatially-varying degradations that occur in practice, such as rain-streaks, haze, raindrops, and motion blur. They decompose the restoration task into two stages of degradation localization and degraded region-guided restoration, unlike existing methods that directly learn a mapping between the degraded and clean images.

In “Color image restoration exploiting inter-channel correlation with a 3-stage CNN” Cui *et al.* propose a 3-stage CNN for color image restoration tasks. In this framework, first the green component is reconstructed, followed by the red and blue channels with parallel networks, then all the intermediate reconstructions are concatenated to generate the final result. This method is successful in three typical color image restoration tasks: color-image demosaicking, color compression artifact reduction, and real-world color image denoising.

Another image restoration work “A deep primal-dual proximal network for image restoration” borrows idea from image classification tasks and proposes a primal-dual proximal network. Specifically, it reformulates a specific instance of the primal-dual hybrid gradient (PDHG) algorithm as a deep network with fixed layers. Each layer corresponds to one iteration of the primal-dual algorithm. Two learning strategies – Full learning and Partial learning – are also proposed for better optimization. The proposed DeepPDNet shows excellent performance on the several benchmark datasets for image restoration and super resolution.

The paper “Semi-supervised landmark-guided restoration of atmospheric turbulent images” considers restoration of atmospheric turbulent (AT) images. As there is no paired training dataset for AT images, especially with faces, this work proposes a semi-supervised method for jointly extracting facial landmarks and restoring degraded images. The proposed approach learns to generate AT images by combining the content from a clean image and turbulence information from AT images in an unpaired manner. It adopts heatmaps from the landmark localization network as an additional prior. Experiments demonstrate the effectiveness of the proposed network on both AT image restoration and landmark localization.

In the rain-removal task, Kui *et al.* ("Multi-level memory compensation network for rain removal via divide-and-conquer strategy") leverage the divide and-conquer strategy by decomposing the learning task into several subproblems according to levels of texture richness. It produces a high-quality rain-free image by subtracting the predicted rain information from multiple subnetworks. Each subnetwork processes a specific sub-sampled image, sampled from the original rainy ones via the Gaussian kernel. Experiments show that the proposed MLMCN outperforms existing deraining methods on several benchmark datasets, and the high-level object detection task.

Also, Yasarla *et al.* ("Exploring Overcomplete Representations for Single Image Deraining using CNNs") proposes a deraining solution called Over-and-Under Complete Deraining Network (OUCD). OUCD consists of two branches: one employing an overcomplete convolutional network architecture for learning local structures by restraining the receptive field of filters and another one employing U-net targeting global structures. The solution significantly improves over state-of-the-art on synthetic and real benchmarks.

Ning *et al.* ("Accurate and Lightweight Image Super-Resolution with Model-Guided Deep Unfolding Network") propose an explainable approach toward SISR named model-guided deep unfolding network (MoG-DUN). MoG-DUN unfolds the iterative process of model-based SISR into a multi-stage concatenation of building blocks with three interconnected trainable modules (denoising, nonlocal-AR, and reconstruction). Experiments show improvements over existing model-based methods.

In "Multi-scale image super-resolution via a single extendable deep network" Zhang *et al.* propose a solution (MSWSR) addressing efficiency and arbitrary upscaling factors. MSWSR implements multi-scale SR simultaneously by learning multi-level wavelet coefficients of the target image. Structurally, MSWSR is composed of one CNN part for low frequencies and one extendable RNN part for high frequencies and multiscale SR. A side window kernel is proposed for efficiency.

In "WDN: A Wide and Deep Network to Divide-and-Conquer Image Super-resolution," Singh and Mittal propose to divide the SISR problem into multiple sub-problems and then solve/conquer them within a neural network design. Their introduced network architecture is much wider and is deeper than existing networks and employs a new technique to calibrate the intensities of feature map pixels. The advantages are demonstrated through extensive experiments.

In "Multi-Grid Back-Projection Networks" Navarrete Michelini *et al.* demonstrate the power of the Mult–Grid Back–Projection (MGBP) network architecture on image and video super-resolution tasks with fidelity and/or perceptual quality targets. MGBP combines a novel cross-scale residual block inspired by the iterative back–projection (IBP) algorithm and a multi-grid recursion strategy inspired by multi–grid PDE solvers to scale computational complexity efficiently with increasing output resolutions.

In "LSTM-DNN Based Autoencoder Network for Nonlinear Hyperspectral Image Unmixing," Zhao *et al.* address the problem of blind hyperspectral unmixing by proposing a non-symmetric autoencoder network to fully exploit the spectral and spatial correlation information. LSTM captures spectral correlation information, a spatial regularization improves the spatial continuity of results, while an attention mechanism further enhance the unmixing performance. The effectiveness of the proposed method is validated on synthetic and real data.

"Uncertainty-Aware Semantic Guidance and Evaluation for Image Inpainting" address the problem of filling in missing irregularly shaped areas of an image, a problem that arises in practice when trying to recover an image that has an overlay (e.g., super-imposed text) or a foreground object that is being synthetically removed or when trying to create a different viewpoint of a scene (in newly dis-occluded areas). The approach that is taken is to iteratively evaluate inpainted visual contents as well as a structural segmentation mask. The approach surpasses other state-of-the-art approaches, in terms of clear boundaries and photo-realistic textures.

The paper "Deep energy: Task driven training of deep neural networks" offers an unsupervised training approach using task-specific energy functions, where the proposed solution is better than the one obtained by a direct minimization of the energy function due to added regularization property of deep neural networks.

B. Image/Video Compression

Breakthroughs in modeling latent-variable probability distributions jointly with parameterized non-linear transformations [16] [17] were what was needed to allow learning-based approaches to image compression to quickly surpass the performance achieved by more traditional models. "Nonlinear transform coding" is the first journal paper with comprehensive coverage of latent-variable RD-curve optimization for nonlinear-transform coding.

The next two papers focus on different approaches to intra-frame block compression. "Intra-frame coding using a conditional autoencoder" introduces an auto-encoder approach to mode-selection for predicting intra-frame image/video blocks. The learned latent-space variable is itself the prediction-function index (replacing the mode-index used in classic intra-frame block coding) and the context pixels condition both parts of the auto-encoder architecture. Cross-channel prediction is provided between the luma and chroma encodings, to avoid the need to separately send the latent-variables for the chroma channels. The results improve the Bjøntegaard delta rate (BD-rate) for both luma and chroma channels compared to previous state-of-the-art. The second of these papers, "Attention-based neural networks for chroma intra prediction in video coding," also looks at intra-frame chroma prediction but does so with a very different approach. In this paper, a purely convolutional network is used with an attention layer [46] to cross-index between the (known) chroma boundary pixels and the (previously decoded) luma pixels within the block. Since the network is purely convolutional, it is able to handle all block sizes (4 × 4, 8 × 8, and 16 × 16), reducing both the space required for the models and the average computation used across the video duration.

The next paper, "MFRNet: A new CNN architecture for post-processing and in-loop filtering," also looks at leveraging convolutional neural-networks within the framework of a traditional video compressor, but this time for in-loop and post-processing filtering. The paper introduces a new neural-network architecture that allows reuse of early-layer representations throughout the remaining layers of the network. The results show significant PSNR gains for both in-loop and post-processing.

Finally, "Learning for video compression with recurrent auto-encoder and recurrent probability model" presents a fully learning-based approach to video compression that outperforms the default-speed setting for ×265, using recurrent probability models for the latent variables of the recurrent auto-encoder network that is used to encode the motion-compensated video frames.

C. Point Cloud Denoising and Compression

Finally, there are two papers on deep learning for point cloud processing, one on denoising and one on compression.

The paper entitled "Learning robust graph-convolutional representations for point cloud denoising" proposes a deep learning method that can simultaneously denoise a point cloud and remove outliers in a single model. The core of the proposed method is a graph-convolutional neural network able to efficiently deal with the irregular domain and the permutation invariance problem typical of point clouds.

The paper entitled "Adaptive deep learning-based point cloud geometry coding" is the first journal paper on point cloud compression. It presents a novel deep learning-based solution for point cloud geometry coding that divides the point cloud into 3D blocks and selects the most suitable available deep learning coding model to code each block, thus maximizing the compression performance.

IV. Outlook

There are many compelling future research challenges still remain to be addressed. These include: i) learned models contain millions of parameters, which makes real-time inference on common devices a challenge, ii) it is difficult to interpret learned models or to provide performance bounds on results, iii) it is important to provide perceptual loss functions, for training, that accurately reflect human preferences, iv) the performance of learned models trained on synthetically generated data drops sharply on real-world images/video, where the quantity and quality of training data is limited, and v) exploiting temporal correlations for efficient and effective video restoration and compression is challenging.

We hope that this special issue broadly summarizes the current state of the art in learned methods for image/video restoration and compression, and inspires researchers to work on numerous future directions calling for deeper investigation.

A. Murat Tekalp, *Lead Guest Editor*
Department of Electrical and Electronics
Engineering
Koç University
Istanbul, Turkey

Michele Covell, *Guest Editor*
Google Research
Mountain View
CA 94043 USA

Radu Timofte, *Guest Editor*
ETH Zurich
Switzerland

Chao Dong, *Guest Editor*
Shenzhen Institute of Advanced
Technology
China

References

[1] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1604–1613.

[2] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271.

[3] K. Zhang, L. Van Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3217–3226

[4] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3118–3126.

[5] Xi Cheng, Z. Fu, and J. Yang, "Zero-shot image super-resolution with depth guided internal degradation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[6] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "WESPE: Weakly supervised photo enhancer for digital cameras," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 691–700.

[7] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 5958–5964.

[8] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised learning for real-world super-resolution," in *Proc. Int. Conf. Comput. Vis. Workshop*, 2019, pp. 3408–3416.

[9] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014.

[10] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2019, pp. 3599–3608.

[11] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super resolution using cycle-in-cycle generative adversarial networks," in *IEEE/CVF Conf. Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2018, pp. 814–81409.

[12] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3277–3285.

[13] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3086–3095.

[14] A. Ignatov, L. Van Gool, and R. Timofte, "Replacing mobile cmera ISP with a single deep learning model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 2275–2285.

[15] P. Wei *et al.*, "Component Divide-and-Conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comp. Vis.*, 2020, pp. 101–117.

[16] A. Ignatov *et al*, "AI benchmark: Running deep neural networks on Android smartphones," in *Proc. IEEE/CVF Int. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 3617–3635.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. IEEE Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[18] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, 2019.

[19] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[20] K. Zhang *et al.*, "AIM 2020 Challenge on Efficient Super-Resolution: Methods and Results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2020.

[21] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE/CVF Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[22] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *IEEE/CVF Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6228–6237.

[23] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[24] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018.

[25] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: Generative adversarial networks with ranker for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3096–3105.

[26] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[27] G. Toderici *et al.*, "Variable rate image compression with recurrent neural networks," in *Proc. Int. Conf. Learning Representations*, 2016.

[28] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Representations*, 2017.

[29] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Proc. Int. Conf. Learn. Representations*, 2017.

[30] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Representations*, 2018.

[31] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10794–10803.

[32] X. Zhao, J. Chen, A. Said, V. Seregin, H. E. Egilmez, and M. Karczewicz, "NSST: Non-separable secondary transforms for next generation video coding," in *IEEE Picture Coding Symp.*, 2016, pp. 1–5.

[33] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *Proc. Int. Conf. Multimedia Model.*, Springer, 2017, pp. 28–39.

[34] Y. Li, *et al.*, "A hybrid neural network for chroma intra prediction," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 1797–1801.

[35] D. Wang, S. Xia, W. Yang, Y. Hu, and J. Liu, "Partition tree guided progressive rethinking network for in-loop filtering of HEVC," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2671–2675.

[36] P. Helle *et al.*, "Intra picture prediction for video coding with neural networks," in *Proc. IEEE Data Compression Conf.*, 2019, pp. 448–457.

[37] M. Gorriz, S. Blasi, A. F. Smeaton, N. E. O'Connor, and M. Mrak, "Chroma intra prediction with attention-based CNN architectures," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 783–787.

[38] L. Murn, S. Blasi, A. F. Smeaton, N. E. O'Connor, and M. Mrak, "Interpreting CNN for low complexity learned sub-pixel motion compensation in video coding," 2020, *arXiv:2006.06392*.

[39] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11006–11015.

[40] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3453–3462.

[41] 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. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8500–8509.

[42] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *IEEE/CVF Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6627–6636.

[43] M. A. Yílmaz and A. M. Tekalp, "End-to-end rate-distortion optimization for bi-directional learned video compression," *IEEE Int. Conf. Image Process.*, Abu Dhabi, UAE, Nov. 2020, pp. 1311–1315.

[44] X.-F. Han, J. S. Jin, M.-J. Wang, W. Jiang, L. Gao, and L. Xiao, "A review of algorithms for filtering the 3D point cloud," *Signal Process.: Image Commun.*, vol. 57, 2017, pp. 103–112.

[45] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai, "An overview of ongoing point cloud compression standardization activities: Video-based (V-PCC) and geometry-based (G-PCC)," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. E13, 2020, doi: 10.1017/ATSIP.2020.12.

[46] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.