

Scalable Human-Machine Point Cloud Compression

Mateen Ulhaq
School of Engineering Science
Simon Fraser University
 Burnaby, BC, Canada
 mulhaq@sfu.ca

Ivan V. Bajić
School of Engineering Science
Simon Fraser University
 Burnaby, BC, Canada
 ibajic@ensc.sfu.ca

Abstract—Due to the limited computational capabilities of edge devices, deep learning inference can be quite expensive. One remedy is to compress and transmit point cloud data over the network for server-side processing. Unfortunately, this approach can be sensitive to network factors, including available bitrate. Luckily, the bitrate requirements can be reduced without sacrificing inference accuracy by using a machine task-specialized codec. In this paper, we present a scalable codec for point-cloud data that is specialized for the machine task of classification, while also providing a mechanism for human viewing. In the proposed scalable codec, the “base” bitstream supports the machine task, and an “enhancement” bitstream may be used for better input reconstruction performance for human viewing. We base our architecture on PointNet++, and test its efficacy on the ModelNet40 dataset. We show significant improvements over prior non-specialized codecs.

Index Terms—deep learning, point cloud compression, coding for machines, scalable coding, classification

I. INTRODUCTION

Point clouds representing 3D visual data are increasingly being used in many applications, including augmented reality, robotics, and autonomous driving. Advances in deep learning have led to the development of deep models for performing machine vision tasks on point cloud data, including classification, object detection, and segmentation. However, most deep learning models are computationally expensive. This poses a challenge for computationally-limited edge devices that want to perform machine vision tasks, and yet are limited in size, energy consumption, cost, and other factors.

One option for performing a machine task on the edge device is to limit the complexity of the model. Unfortunately, this usually comes at the cost of model accuracy. Another option is to transmit the input data to a server for machine analysis. In this approach, the edge device compresses the input data prior to transmission, often using a codec designed to reconstruct the input for human viewing. However, such *input reconstruction* codecs are not optimized for machine analysis. Thus, they often spend large amounts of bits on encoding information that is not relevant to the machine task. This results in a worse rate-accuracy trade-off than is possible with a more specialized codec. In situations where a low rate is desired — for instance, in areas of poor network connectivity, or congested networks — using a non-specialized codec may result in excessively high machine task latency [1].

In order to improve the rate-accuracy trade-off, we may instead use a codec that is specialized for the machine task.

Such codecs often perform part of the machine task on the edge device itself. In this hybrid approach, the edge device simultaneously compresses the input and performs part of the machine task. This allows the model to discard unnecessary information, thus reducing the rate, resulting in a system that is more robust to changing network conditions, and may reduce system latency over a certain range of available bitrates [2].

In [3], a novel codec for point cloud classification was proposed. This learned codec, based on PointNet [4], compresses the input point cloud into a highly compressed representation that is intended solely for machine analysis, in this case classification. This codec was shown to achieve a significantly better rate-accuracy trade-off in comparison with alternative methods using standard codecs that are not specialized for machine analysis. This was achieved by removing not only statistical redundancy, but also task-irrelevant information, during the compression process.

While the codec in [3] achieves a good rate-accuracy trade-off for point cloud classification, it is not suitable for other purposes. Most applications involving automated machine-based analysis are expected to run the machine task continuously, but may occasionally require human verification or review. Hence, it is important to develop codecs that are able to support both tasks — machine vision and human viewing — efficiently. In this paper, we present such a scalable codec, the first in the point cloud literature, which supports point cloud classification, while also providing a mechanism for human viewing. Our code is available online.¹

II. RELATED WORK

Point cloud classification is among the most researched point cloud analysis tasks. Related classification models accept different input formats, including point lists (e.g., PointNet [4] and PointNet++ [5]), 3D voxel grids (e.g., VoxNet [6]), and Octrees (e.g., OctNet [7]). Since our work builds on PointNet and PointNet++, we assume the input format is a point list.

Conventional handcrafted point cloud codecs include Draco [8] and G-PCC [9] (implemented as TMC13 [10]). More recently, the research focus has shifted towards learned codecs, following the seminal work of Ballé *et al.* [11] who proposed a variational autoencoder (VAE) based architecture

¹ <https://github.com/multimedialabsfu/learned-point-cloud-compression-for-classification>

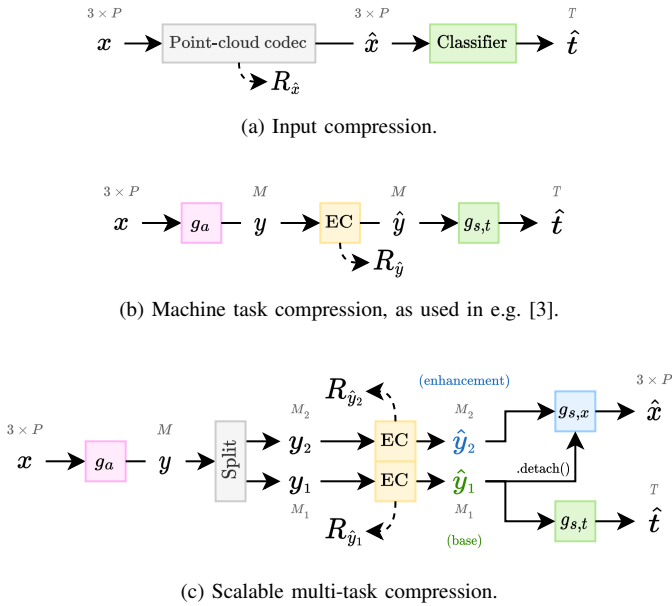


Fig. 1. High-level comparison of codec architectures.

for image compression. In their architecture, the input x is first transformed into a latent representation y , which is then quantized and entropy-coded using a learned entropy model. Such an architecture can be trained end-to-end using the loss

$$\mathcal{L} = R + \lambda \cdot D(x, \hat{x}), \quad (1)$$

where $D(x, \hat{x})$ is the distortion measure between the input x and decoded \hat{x} , and R is the estimate of the entropy of \hat{y} . This mechanism has been successful in various fields of learned compression, including point cloud compression [12]–[16].

Specialized compression for machine tasks — often referred to as *coding for machines* — has been explored for images [17] and video [18] and, recently, for point clouds [3]. Moreover, scalable multi-task coding approaches [19], [20] have shown that one can perform a machine vision task at a fairly low bitrate, while enabling other tasks, such as input reconstruction for human viewing, with an additional enhancement bitstream.

While quality-scalable point cloud coding has been studied before [21], this paper presents the first scalable multi-task point cloud codec: the base bitstream supports point cloud classification, while the enhancement bitstream allows point cloud reconstruction. Unlike our earlier work [3], which presented a classification-optimized codec based on PointNet [4], our scalable codec is based on PointNet++ [5], which is a hierarchical extension of PointNet.

III. PROPOSED CODEC

A. Preliminaries

In Fig. 1a, we show an abstract representation of an *input compression codec*. The input point cloud x is encoded and decoded as \hat{x} by any desired point cloud codec, including non-learned codecs such as G-PCC [9]. Then, the reconstructed point cloud \hat{x} is fed into a classification model (e.g., PointNet)

in order to obtain the class prediction \hat{t} . This approach provides a baseline for comparison with our proposed codec.

In Fig. 1b, we show an abstract representation of a *machine task codec*, as was explored by [3] for point cloud classification. Using the same terminology as in [11], g_a refers to the *analysis transform*, and g_s refers to the *synthesis transform*. In this codec, the input point cloud x is first encoded into a latent representation $y = g_a(x)$. This is then quantized as $\hat{y} = Q(y)$, and then losslessly compressed using a learned entropy model. For instance, in [3], a fully-factorized entropy model was used. The reconstructed latent representation \hat{y} may then be used to predict the classes $\hat{t} = g_{s,t}(\hat{y})$.

B. Scalable human-machine compression codec

In Fig. 1c, we show a high-level representation of a *scalable multi-task codec*. Following the principle of *latent space scalability* [20], the scalable multi-task codec splits the latent space into two parts, $[\hat{y}_1, \hat{y}_2] = \text{split}(\hat{y})$, along the channel dimension. The first part \hat{y}_1 is called the “base” layer and the second part \hat{y}_2 is called the “enhancement” layer. The base layer is used for the machine task (i.e., classification) to predict the class $\hat{t} = g_{s,t}(\hat{y}_1)$. Both base and enhancement layers are used for input reconstruction: $\hat{x} = g_{s,x}(\text{concat}[\text{detach}(\hat{y}_1), \hat{y}_2])$. This approach allows for scalability: the enhancement bitstream only needs to be computed and transmitted when human viewing is desired. Note that the detach operation is used to disable gradient propagation of $D(x, \hat{x})$ backwards through \hat{y}_1 . This improves the specialization of \hat{y}_1 towards the machine task; otherwise, \hat{y}_1 may end up with some enhancement-layer information [22]. Two separate bitstreams (base and enhancement) are produced, whose total rate is $R_{\hat{y}} = R_{\hat{y}_1} + R_{\hat{y}_2}$.

C. Proposed architecture

Our complete proposed codec architecture, based on PointNet++ [5], is shown in Fig. 2, and the details of each block are shown in Fig. 3. Our codec is provided in “full” and “lite” configurations, as detailed in Table I. The input point cloud of P points is represented as a matrix $x \in \mathbb{R}^{3 \times P}$ of (x, y, z) coordinates. u is an additional set of attribute features (e.g., normals, color, etc.) that optionally may also be compressed. At the beginning of the encoder, x and u are concatenated along the channel dimension so that $u^{(0)} = \text{concat}[x, u]$. This ensures that the same encoding capabilities are available for both. We feed $x^{(0)} = x$ and $u^{(0)}$ into a sequence of downsampling blocks. Each i -th downsampling block takes in $x^{(i-1)}$ and $u^{(i-1)}$, and outputs a smaller set of centroids $x^{(i)}$ and features $u^{(i)}$, along with features $u_g^{(i-1)}$ that are grouped by the centroids. The final $u^{(3)}$ is compressed in a multi-task scalable manner, from which $\hat{u}^{(3)}$ is derived. Each $u_g^{(i-1)}$ is compressed using a standard transform-encoder-decoder-transform compression pipeline to obtain $\hat{u}_g^{(i-1)}$. A sequence of upsampling blocks is then applied, where each i -th upsampling block takes in $\hat{u}^{(i)}$. For $i > 0$, the output of the i -th upsampling block is concatenated with $\hat{u}_g^{(i-1)}$ to give $\hat{u}^{(i-1)}$. For $i = 0$, the output is \hat{x} .

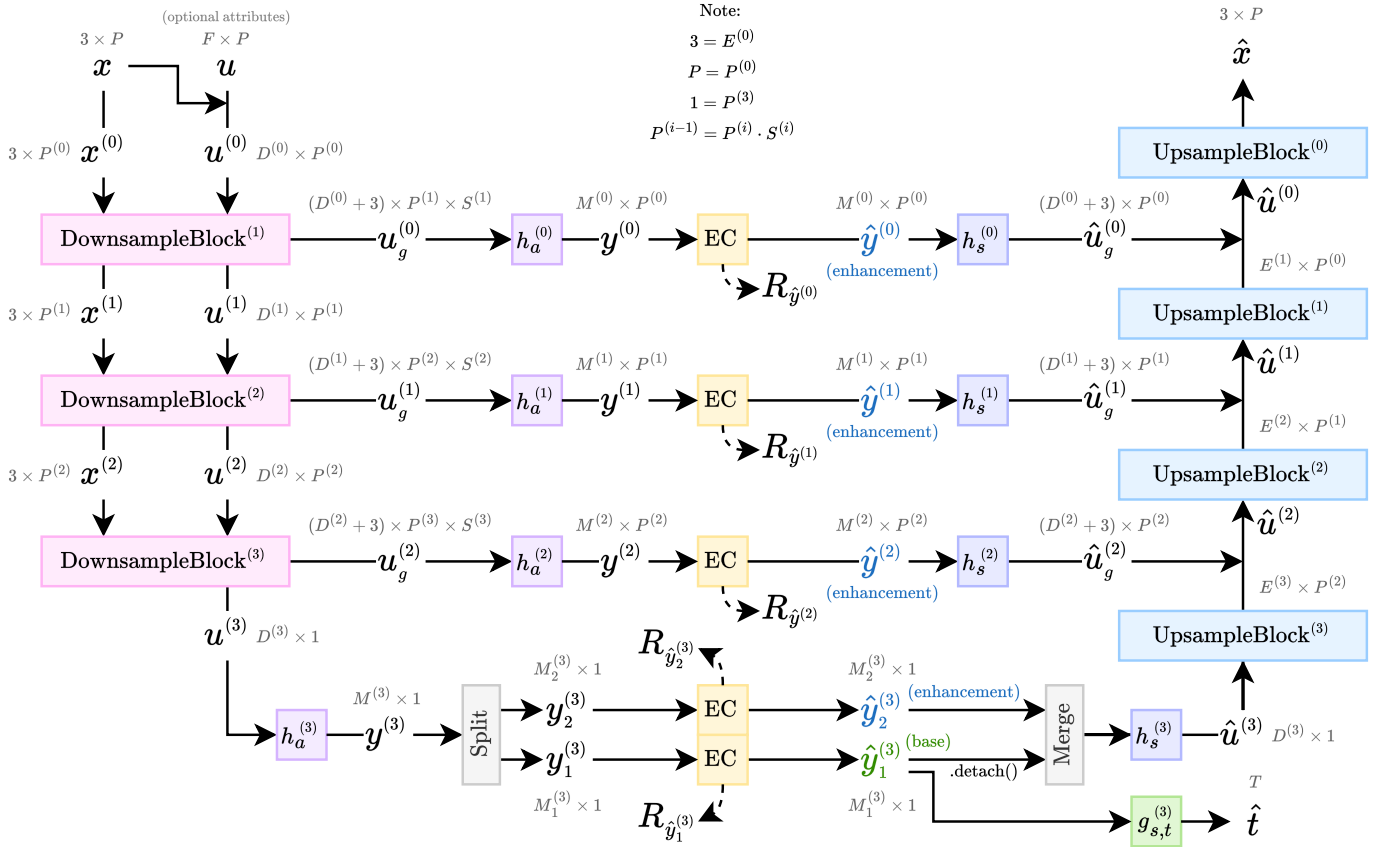


Fig. 2. Proposed codec architecture.

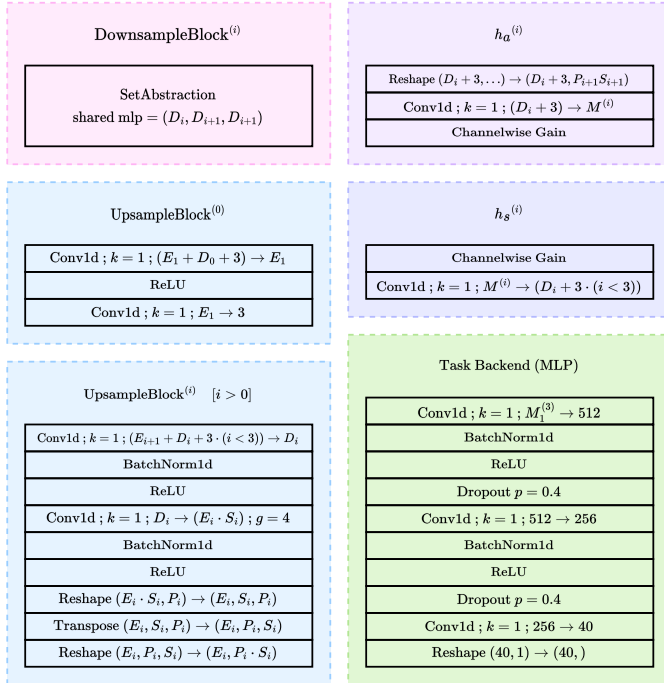


Fig. 3. Proposed codec architecture (details).

1) *Downsampling*: Each downsampling block is a PointNet++ “set abstraction” layer (see [5] for more details), with minor modifications. We used single-scale grouping (SSG) of points for our proposed codec, though it may likely be improved with multi-scale grouping (MSG) or multi-resolution grouping (MRG) as described in [5]. The “set abstraction” layer takes as input a (subsampled) point cloud $\mathbf{x}^{(i-1)}$ of shape $3 \times P^{(i-1)}$ and features $\mathbf{u}^{(i-1)}$ of shape $D^{(i-1)} \times P^{(i-1)}$ containing information about each corresponding point. Using farthest point sampling (FPS), a set of $P^{(i)}$ centroids $\mathbf{x}^{(i)}$ of shape $3 \times P^{(i)}$ is selected from $\mathbf{x}^{(i-1)}$. Then, the points in $\mathbf{x}^{(i-1)}$ are grouped into $P^{(i)}$ groups of $S^{(i)}$ points each. Then for each centroid, a ball query² [5] is performed to find the first $S^{(i)}$ points that are within a radius of $R^{(i)}$ from the given centroid. The relative positions of each group of points is then computed with respect to its associated centroid, resulting in the residuals $\mathbf{r}^{(i)} = \text{ball_query}(\mathbf{x}^{(i-1)}, \mathbf{x}^{(i)}, R^{(i)}) - \text{repeat}(\mathbf{x}^{(i)}, S^{(i)})$ of shape $3 \times P^{(i)} \times S^{(i)}$. Each point in $\mathbf{r}^{(i)}$ is concatenated with its respective feature vector in $\mathbf{u}^{(i-1)}$, resulting in a grouped feature tensor $\mathbf{u}_g^{(i-1)}$ of shape $(D^{(i-1)} + 3) \times P^{(i)} \times S^{(i)}$. Then, $\mathbf{u}_g^{(i-1)}$ is fed into a miniature group-wise “PointNet encoder” block consisting of a shared

²Note that due to the ball query, the same point may be assigned to multiple centroids. Also, centroids are not always present within their own group of points. Nonetheless, the ball query has the benefit of scale-invariant grouping.

TABLE I
ARCHITECTURE HYPERPARAMETER CONFIGURATIONS

Codec	i	$P^{(i)}$	$S^{(i)}$	$R^{(i)}$	$D^{(i)}$	$E^{(i)}$	$M^{(i)}$
full	0	1024			3	3	0
	1	256	4	0.2	128	64	0
	2	64	4	0.4	192	32	64
	3	1	64		256	16	(48, 16)
lite	0	1024			3	3	0
	1	256	4	0.2	32	32	0
	2	64	4	0.4	48	16	16
	3	1	64		64	8	(48, 16)

multi-layer perceptron (MLP) with a max pooling operation at the end. That is, the same PointNet encoder is applied to each group of features independently and identically. This results in $\mathbf{u}^{(i)}$ of shape $D^{(i)} \times P^{(i)}$. Three downsampling blocks are used in our proposed codec. The last downsampling block groups all the remaining points into a single group, i.e., $P^{(3)} = 1$.

2) *Feature compression*: At each level $i < 3$, a latent representation $\mathbf{y}^{(i)} = h_a^{(i)}(\mathbf{u}_g^{(i)})$ of shape $M^{(i)} \times P^{(i)}$ is computed, then quantized as $\hat{\mathbf{y}}^{(i)} = Q(\mathbf{y}^{(i)})$, and then entropy-coded using a learned entropy model. The resulting bitstream is transmitted and then decoded as $\hat{\mathbf{y}}^{(i)}$. From this, a set of grouped features $\hat{\mathbf{u}}_g^{(i)} = h_s^{(i)}(\hat{\mathbf{y}}^{(i)})$ is computed for usage during upsampling. (Note that $\hat{\mathbf{u}}_g^{(i)}$ and $\mathbf{u}_g^{(i)}$ may inhabit entirely different feature spaces, and are unrelated.)

The final feature vector $\mathbf{u}^{(3)}$ is fed into the scalable multi-task compression pipeline described in Section III-B. The latent representation $\mathbf{y}^{(3)} = h_a^{(3)}(\mathbf{u}^{(3)})$ is computed, then quantized as $\hat{\mathbf{y}}^{(3)} = Q(\mathbf{y}^{(3)})$, and then entropy-coded to generate a base bitstream for $\hat{\mathbf{y}}_1^{(3)}$ and an enhancement bitstream for $\hat{\mathbf{y}}_2^{(3)}$. The base bitstream is decoded as $\hat{\mathbf{y}}_1^{(3)}$, and then fed into a decoder-side task backend to obtain the class prediction $\hat{\mathbf{t}} = g_{s,t}^{(3)}(\hat{\mathbf{y}}_1^{(3)})$. The enhancement bitstream is decoded as $\hat{\mathbf{y}}_2^{(3)}$, and then its concatenation with $\hat{\mathbf{y}}_1^{(3)}$ (detached) is fed into a decoder-side transform to obtain $\hat{\mathbf{u}}_g^{(3)} = h_s^{(3)}(\text{concat}[\text{detach}(\hat{\mathbf{y}}_1^{(3)}), \hat{\mathbf{y}}_2^{(3)}])$.

Note that we split only $\mathbf{y}^{(3)}$ into $\mathbf{y}_1^{(3)}$ and $\mathbf{y}_2^{(3)}$. This is because $\mathbf{y}_1^{(3)}$ suffices for the classification task, while $\mathbf{y}_2^{(3)}$ and $\mathbf{y}^{(j)}$ for $j < 3$ are only needed for input reconstruction. For brevity, in the experiments we only test input reconstruction with all enhancement layers, but the proposed architecture also allows for quality- or density-scalable input reconstruction.

3) *Upsampling*: An upsampling block takes in $E^{(i+1)} + D^{(i)} + 3 \cdot (1 - \delta_{i,3})$ channels and outputs $E^{(i)}$ channels, where $\delta_{i,j}$ is the Kronecker delta ($\delta_{i,j} = 1$ if $i = j$, 0 otherwise). Each upsampling block consists of a simple two-layer point-wise MLP. For upsampling blocks at levels deeper than $i > 0$, batch normalizations are also present. Furthermore, each such block is followed by a reshape-transpose-reshape combination that converts the resulting shape $(E^{(i)} \cdot S^{(i)}) \times P^{(i)}$ into the output shape $E^{(i)} \times (S^{(i)} \cdot P^{(i)})$. To ensure compatibility of shapes among various components during reconstruction, we enforce the constraint $P^{(i-1)} = P^{(i)} \cdot S^{(i)}$.

IV. EXPERIMENTS

We trained our models on the ModelNet40 [23] dataset, sampling $P = 1024$ points per object, and reconstructed the same number of points. Our implementation uses the PyTorch, CompressAI [24], and CompressAI Trainer [25] libraries.

The loss function used for training was

$$\mathcal{L} = \sum_{i=0}^3 R_{\hat{\mathbf{y}}^{(i)}} + \lambda_x D(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_t D(\mathbf{t}, \hat{\mathbf{t}}). \quad (2)$$

The rate estimate $R_{\hat{\mathbf{y}}^{(i)}}$ for each level is the negative log likelihood outputted by the respective entropy models. The distortion $D(\mathbf{t}, \hat{\mathbf{t}})$ is the cross-entropy between the one-hot encoded labels \mathbf{t} and the softmax of the model’s prediction $\hat{\mathbf{t}}$. The distortion $D(\mathbf{x}, \hat{\mathbf{x}})$ is the Chamfer distance between the input and reconstructed point clouds. We trained different models to operate at different base/enhancement rate points by varying the hyperparameters $\lambda_t \in \{2^{-7}, 2^{-6}, \dots, 2^0\}$ and $\lambda_x \in [1, 8000]$. These values are reported for rates in units of bits per point (bpp), i.e., the rates are divided by $P = 1024$. To improve stability during training, we disabled the first two levels by setting $M^{(0)} = 0$ and $M^{(1)} = 0$. We also set $M^{(2)} = 64$ and $M^{(3)} = 64$, and used a partitioning ratio of 0.75 for the last layer, i.e., $M_1^{(3)} = 48$ (base) and $M_2^{(3)} = 16$ (enhancement).

The input compression codecs were evaluated using the same methodology for varying P and input scaling as in [3]. Since codecs like TMC13 [10] and OctAttention [15] are tailored to point clouds with a large number of points, the coding overhead they produce may be significant when applied to smaller point clouds such as those used in classification. To improve the fairness when measuring the rate, we ignored such overhead within the bitstream when possible.

V. RESULTS

Figs. 4a and 4b show the rate-accuracy curves for the base task of point-cloud classification. Only the rate of the base bitstream is measured for the base task. Scalable codec curves are grouped by their worst-case reconstruction ability measured in terms of Chamfer distance [26]. Our proposed codec achieves better classification performance than all existing codecs, including [3] (denoted “UIhaq2023”), which in turn is better than input compression codecs (“ICC”) that use a point-cloud codec (TMC13 [10] and OctAttention [15]) followed by a PointNet/PointNet++ classifier. Only the best ICC curves are shown. Note that “P=*” indicates the best results attained when varying $P \in \{8, 16, \dots, 1024\}$.

There is a large gap in base task performance between task-specialized codecs (the proposed one and [3]) and non-specialized codecs because, in addition to statistical redundancy, the former codecs remove a significant amount of task-irrelevant information. Meanwhile, the proposed codec outperforms [3] due to its reliance on the more powerful PointNet++, rather than PointNet.

Fig. 4c show the rate-distortion curves for the reconstruction task in terms of the rate versus the Chamfer distance. For

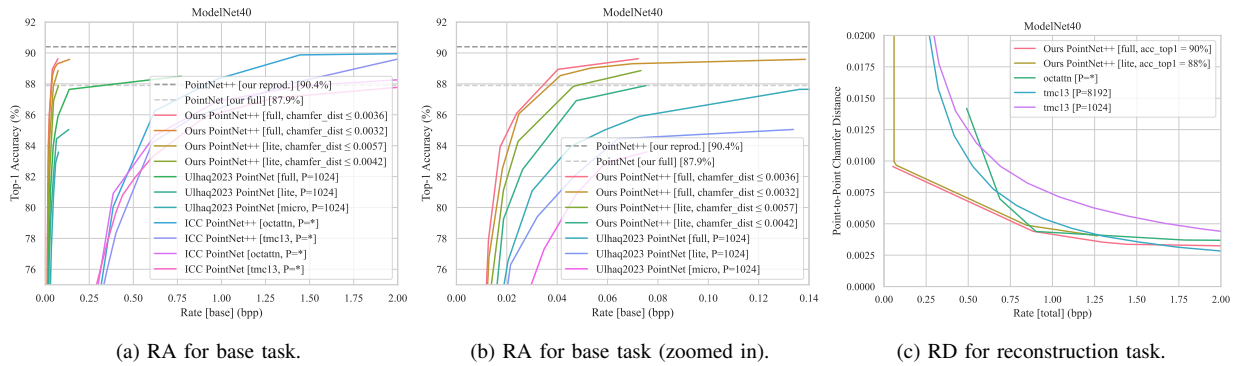


Fig. 4. Rate-accuracy (RA) and rate-distortion (RD) curves on the ModelNet40 dataset, with rate units of bits per point (bpp) scaled for 1024 points.

the proposed codec, the rate includes both the base and enhancement bitstreams. The figures show that our proposed codec is also capable of achieving better reconstruction quality than conventional codecs at low rates, below 1.6 bpp. As the rate increases, the conventional codecs eventually catch up in terms of reconstruction quality.

VI. CONCLUSION

In this paper, we proposed a scalable multi-task compression codec for point clouds. Our proposed codec produces a base bitstream that supports point cloud classification and an enhancement bitstream that, together with the base bitstream, supports reconstruction of the point cloud for human viewing. The proposed codec showed a significant improvement in rate-accuracy performance for the base task over existing codecs, while also achieving competitive performance against conventional codecs on input reconstruction at reasonably low rates. In the future, we hope that this work will inspire further research into scalable multi-task compression codecs for point clouds on more complex real-world datasets, and for other tasks such as semantic segmentation.

REFERENCES

- [1] N. Shlezinger and I. V. Bajić, "Collaborative inference for AI-empowered IoT devices," *IEEE Internet of Things Magazine*, vol. 5, no. 4, pp. 92–98, Dec. 2022.
- [2] I. V. Bajić, W. Lin, and Y. Tian, "Collaborative intelligence: Challenges and opportunities," in *Proc. IEEE ICASSP*, 2021, pp. 8493–8497.
- [3] M. Ulihaq and I. V. Bajić, "Learned point cloud compression for classification," in *Proc. IEEE MMSP*, 2023.
- [4] C. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF CVPR*, 2017, pp. 77–85.
- [5] C. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NIPS*, 2017.
- [6] D. Maturana and S. A. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.
- [7] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE/CVF CVPR*, 2017, pp. 6620–6629.
- [8] Google, "Draco: 3D data compression," 2017. [Online]. Available: <https://google.github.io/draco/>
- [9] K. Mammou, P. A. Chou, D. Flynn, M. Krivokuća, O. Nakagami, and T. Sugio, "G-PCC codec description v2," 2019, ISO/IEC JTC1/SC29/WG11 N18189.
- [10] D. Flynn and K. Mammou, "MPEG-PCC-TMC13," 2021. [Online]. Available: <https://github.com/MPEGGroup/mpeg-pcc-tmc13>
- [11] J. Ballé, D. C. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.
- [12] M. Quach, G. Valenzise, and F. Dufaux, "Learning convolutional transforms for lossy point cloud geometry compression," in *Proc. IEEE ICIP*, 2019, pp. 4320–4324.
- [13] W. Yan, Y. shao, S. Liu, T. H. Li, Z. Li, and G. Li, "Deep autoencoder-based lossy geometry compression for point clouds," *arXiv preprint arXiv:1905.03691*, 2019.
- [14] Y. He, X. Ren, D. Tang, Y. Zhang, X. Xue, and Y. Fu, "Density-preserving deep point cloud compression," *arXiv preprint arXiv:2204.12684v1*, 2022.
- [15] C. Fu, G. Li, R. Song, W. Gao, and S. Liu, "OctAttention: Octree-based large-scale contexts model for point cloud compression," in *Proc. AAAI*, vol. 36, no. 1, Jun. 2022, pp. 625–633.
- [16] K.-S. You, P. Gao, and Q. T. Li, "IPDAE: Improved patch-based deep autoencoder for lossy point cloud geometry compression," in *Proc. 1st Int. Workshop on Advances in Point Cloud Compression, Processing and Analysis*, 2022.
- [17] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, and E. Rahtu, "Image coding for machines: an end-to-end learned approach," in *Proc. IEEE ICASSP*, 2021, pp. 1590–1594.
- [18] L.-Y. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video Coding for Machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.
- [19] 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. IEEE ICME*, 2020, pp. 1–6.
- [20] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," *IEEE Trans. Image Process.*, vol. 31, pp. 2739–2754, 2022.
- [21] A. F. R. Guarda, N. M. M. Rodrigues, and F. Pereira, "Point cloud geometry scalable coding with a single end-to-end deep learning model," in *Proc. IEEE ICIP*, 2020, pp. 3354–3358.
- [22] Y. Foroutan, A. Harell, A. de Andrade, and I. V. Bajić, "Base layer efficiency in scalable human-machine coding," in *Proc. IEEE ICIP*, 2023, pp. 3299–3303.
- [23] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE/CVF CVPR*, 2015, pp. 1912–1920.
- [24] J. Bégin, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: a PyTorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.
- [25] M. Ulihaq and F. Racapé, "CompressAI Trainer," 2022. [Online]. Available: <https://github.com/InterDigitalInc/CompressAI-Trainer>
- [26] H. Blum, "A transformation for extracting new descriptors of shape," in *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, Ed. Cambridge, MA: MIT Press, 1967, pp. 362–380.