Lossless Coding of Point Cloud Geometry using a Deep Generative Model

Dat Thanh Nguyen, Maurice Quach, Giuseppe Valenzise, Pierre Duhamel

Lossless Coding of Point Cloud Geometry using a Deep Generative Model

Dat Thanh Nguyen, Maurice Quach, *Student Member, IEEE,* Giuseppe Valenzise, *Senior Member, IEEE*, Pierre Duhamel, *Life Fellow, IEEE*

***Abstract*—This paper proposes a lossless point cloud (PC) geometry compression method that uses neural networks to estimate the probability distribution of voxel occupancy. First, to take into account the PC sparsity, our method adaptively partitions a point cloud into multiple voxel block sizes. This partitioning is signalled via an octree. Second, we employ a deep auto-regressive generative model to estimate the occupancy probability of each voxel given the previously encoded ones. We then employ the estimated probabilities to code efficiently a block using a context-based arithmetic coder. Our context has variable size and can expand beyond the current block to learn more accurate probabilities. We also consider using data augmentation techniques to increase the generalization capability of the learned probability models, in particular in the presence of noise and lower-density point clouds. Experimental evaluation, performed on a variety of point clouds from four different datasets and with diverse characteristics, demonstrates that our method reduces significantly (by up to 37%) the rate for lossless coding compared to the state-of-the-art MPEG codec.**

***Index Terms*—Point Cloud Coding, Deep Learning, G-PCC, context model, arithmetic coding.**

I. Introduction

POINT clouds (PC) are becoming the most popular data structure for many 3D applications such as augmented, mixed or virtual reality, as they enable six degrees of freedom (6DoF) interaction. Typical PCs contain millions of points, each point being represented by x, y, z coordinates, and attributes (e.g. color, normal, etc.). This entails a high transmission and storage cost. As a result, there is a massive demand for efficient Point Cloud Compression (PCC) methods to enable the practical use of this content.

The Moving Picture Expert Group (MPEG) has studied coding solution for various categories of point clouds, including static point clouds (category 1), dynamic point clouds (category 2), and LiDAR sequences (category 3 – dynamically acquired point clouds). As a result, two PCC standards have been developed [1]–[3]: Video-based PCC (V-PCC) and Geometry-based PCC (G-PCC). V-PCC focuses on dynamic point clouds, and projects the volumetric video onto 2D planes before encoding. The generated 2D videos are then compressed using 2D video coding standards. This approach benefits from efficient 2D video coding solutions which have been optimized over several decades. On the other hand, G-PCC targets static content, and the geometry and attribute information are independently encoded. Color attributes can be encoded using methods based on the Region Adaptive Hierarchical Transform (RAHT) [4], Predicting Transform or Lifting Transform [3]. Coding the PC geometry is particularly important to convey the 3D structure of the PC, but is also challenging, as the non-regular sampling of point clouds makes it difficult to use conventional signal processing and compression tools. In this paper, we focus on *lossless coding* of point cloud geometry.

In particular, we consider the case of *voxelized* point clouds. Voxelization is the process that quantizes the coordinates of a point cloud to integer precision prior to the coding process. This process is common in many coding scenarios, e.g., when dealing with dense point clouds such as those produced by camera arrays. After voxelization, the point cloud geometry can be represented either directly in the voxel domain or using an octree spatial decomposition. PCs are divided into a fixed number of cubes, which defines the resolution (e.g., 10 bit = 1024 cubes per dimension). Each cube is called a voxel. If a voxel contains at least one point, it is called an occupied voxel. Usually, very few voxels are occupied and a large part of the volume is empty. An octree representation can be obtained by recursively splitting the volume into eight sub-cubes until the desired precision is achieved. Then, occupied blocks are marked by bit 1 and empty blocks are marked by bit 0. Consequently, at each level, the generated 8 bits represent the occupancy state of an octree node (octant). Our method operates in both the voxel and octree domain. On the one hand, the octree representation can naturally adapt to the sparsity of the point cloud, as empty octants do not need to be further split; on the other hand, in the voxel domain convolutions can be naturally expressed, and geometric information (i.e., planes, surfaces, etc.) can be explicitly processed by a neural network.

In this work, we propose a deep-learning-based method (named VoxelDNN) for lossless compression of static voxelized point cloud geometry. Our main contributions are:

- We employ for the first time a deep generative model in the voxel domain to estimate the occupancy probabilities sequentially using a masked 3D convolutional network. The conditional distribution is then used to model the context of a context-based arithmetic coder.
- We propose an optimal rate-driven partitioning and context selection algorithm. The partitioning algorithm adapts to the point cloud sparsity by employing a hybrid octree/voxel representation while the context to encode

D. T. Nguyen, M. Quach, G. Valenzise and P. Duhamel are with the Université Paris-Saclay, CNRS, CentraleSupelec, Laboratoire des Signaux et Systèmes (UMR 8506), 91190 Gif-sur-Yvette, France (email: thanh-dat.nguyen@centralesupelec.fr; maurice.quach@l2s.centralesupelec.fr; giuseppe.valenzise@l2s.centralesupelec.fr; pierre.duhamel@l2s.centralesupelec.fr).

each block is expanded to the neighboring blocks and the expansion size is optimally selected.
- We propose specific data augmentation techniques for 3D point clouds coding, to increase its generalization capability.

We demonstrate experimentally that the proposed solution outperforms the state-of-the-art MPEG G-PCC lossless codec in terms of bits per occupied voxel over a set of point clouds with varying density and content type. The rest of the paper is structured as follows: Section II reviews the related work; the proposed method is described in Section III; Section IV presents the experimental results; and finally Section V concludes the paper.

II. Related work

Relevant work related to this paper includes state-of-the-art PC geometry coding and learning-based methods in image and point cloud compression.

A. MPEG G-PCC and Conventional Lossless Codecs

Most existing methods that compress point cloud geometry, including MPEG G-PCC, use octree coding [5]–[12] and local approximations called "triangle soups" (trisoup) [5], [13].

In the G-PCC geometry coder, points are first transformed and voxelized into an axis-aligned bounding box before geometry analysis using trisoup or octree scheme. In the trisoup coder, geometry can be represented by a pruned octree plus a surface model. This model approximates the surface in each leaf of the pruned octree using 1 to 10 triangles. In contrast, the octree coder partitions voxelized blocks until sub-cubes of dimension one are reached. First, the coordinates of isolated points are independently encoded to avoid "polluting" the octree coding (Direct Coding Mode - DCM) [14]. To encode the occupancy pattern of each octree node, G-PCC introduces many methods to exploit local geometry information and obtain an accurate context for arithmetic coding, such as Neighbour-Dependent Entropy Context [15], intra prediction [16], planar/angular coding mode [17], [18], etc. In this paper, we compare our method against G-PCC lossless geometry coding with octree coding which also targets static point clouds.

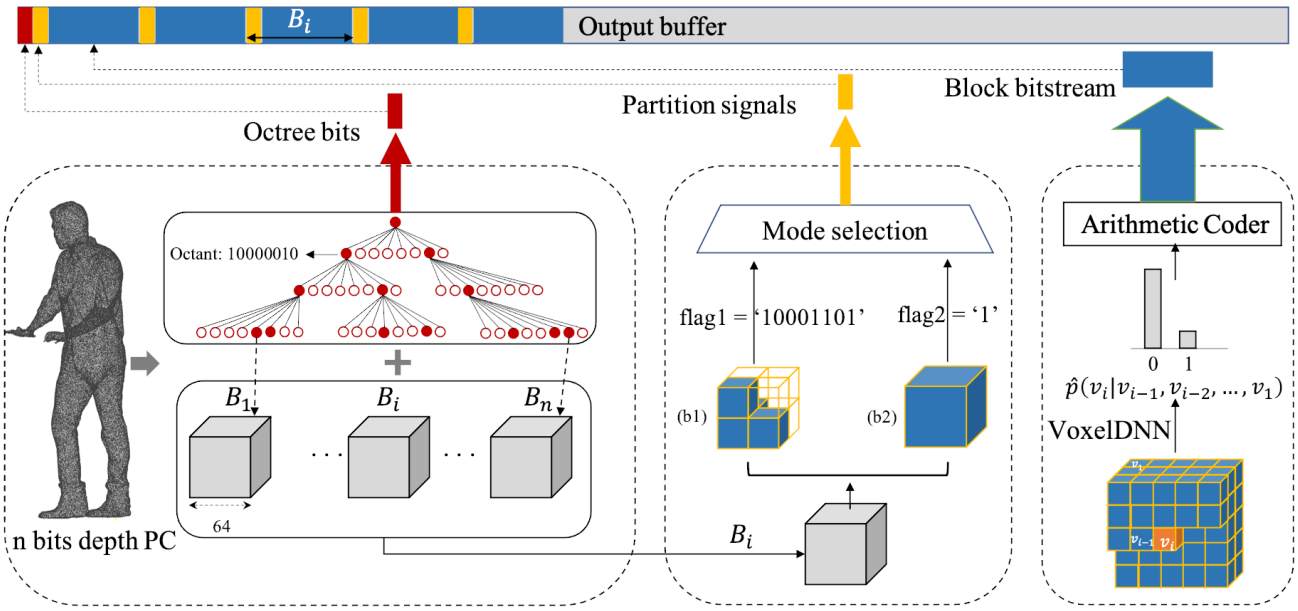
In order to deal with the irregular point space, many octree-based lossless PCC methods have been proposed. In [5], the authors proposed an octree-based method which aims at reducing entropy by employing prediction techniques based on local surface approximations to predict occupancy patterns. Recently, more context modeling based approaches are proposed [8]–[10]. For example, the intra-frame compression method P(PNI) proposed in [10] builds a reference octree by propagating the parent octet to all children nodes, thus providing 255 contexts to encode the current octant. Octree coding allows for a progressive representation of point clouds since each level of the octree is a downsampled version of the point cloud. However, a drawback of octree representation is that, at the first levels of the tree, it produces "blocky" scenes, and geometry information of point clouds (i.e., curve, plane) is lost. The authors of [19] proposed a binary tree based method which analyzes the point cloud geometry using binary tree structure and realizes an intra prediction via the extended Travelling Salesman Problem (TSP) within each leaf node. Instead, in this paper, we employ a hybrid octree/voxel representation to better exploit the geometry information. Besides, the methods in [8]–[10] produce frequency tables which are collected from the coarser level or the previous frame and must be transmitted to the decoder. Our method predicts voxel distributions in a sequential manner at the decoder side, thus avoiding the extra cost of transmitting large frequency tables.

B. Generative Models and Learning-based Compression

Estimating the data distribution from a training dataset is the main objective of generative models, and is a central problem in unsupervised learning. It has a number of applications, from image generation [20]–[23], to image compression [24]–[26] and denoising [27]. Among the several types of generative models proposed in the literature [28], auto-regressive models such as PixelCNN [22], [23] are particularly relevant for our purpose as they allow to compute the exact likelihood of the data and to generate realistic images, although with a high computational cost. Specifically, PixelCNN factorizes the likelihood of a picture by modeling the conditional distribution of a given pixel's color given all previously generated pixels. These conditional distributions only depend on the possible pixel values with respect to the scanned context, which imposes a *causality* constraint. PixelCNN models the distribution using a neural network and the causality constraint is enforced using masked filters in each convolutional layer. Recently, this approach has also been employed in image compression to yield accurate and learnable entropy models [26]. Our paper explores the potential of this approach for point cloud geometry compression by adopting and extending conditional image modeling and masking filters into the 3D voxel domain.

Inspired by the success in learning-based image compression, deep learning has been widely adopted in point cloud coding both in the octree domain [11], [12], voxel domain [29]–[34] and point domain [35]–[37]. Recently, the authors of [11] proposed an octree-based entropy model that models the probability distributions of the octree symbols based on the contextual information from octree structure. This method only targets static LiDAR point cloud compression. The extension version for intensity-valued LiDAR streaming data using spatio-temporal relations is proposed in [12]. However, these methods target dynamically acquired point clouds, while in this paper we mainly focus on dense static point clouds.

Working in the voxel domain enables to easily extend most 2D tools, such as convolutions, to the 3D space. Many recent 3D convolution based autoencoder approaches for lossy coding [31]–[34] compress 3D voxelized blocks into latent representations and cast the reconstruction as a binary classification problem. The authors of [35] proposed a pointnet-based auto-encoder method which directly takes points as input rather than voxelized point cloud. To handle sparse point clouds, recent methods leverage advances in sparse convolution [38], [39] to allow point-based approaches [36], [37]. For example, the proposed lossy compression method



(a) High level partitioning **(b) Multi-resolution encoder** **(c) Context model – VoxelDNN**

Fig. 1: Overview of the proposed method. (a): a n bit depth point cloud is partitioned down to the $n-6$ octree level, yielding occupied blocks of size $64 \times 64 \times 64$. (b): We encode each block of 64^3 voxels as a single block (b1), or divide it into 8 children blocks (b2), depending on the total number of bits of each solution (partitioning level = 2). This procedure is repeated recursively for increasing partitioning levels up to 5. (c): For each occupied block of size d, the context model estimates the distribution of each voxel given the previously encoded voxels.

in [37] progressively downscale the point cloud into multiple scales using sparse convolutional transforms. Then, at the bottleneck, the geometry of scaled point cloud is encoded using an octree codec and the attributes are compressed using a learning-based context model. In contrast, in this paper, we focus on dense voxelized point clouds and losslessly encode each voxel using the learned distribution from its 3D context. In addition, we apply this approach in a block-based fashion, which has been successfully employed in traditional image and video coding.

III. Proposed Method

A. System overview

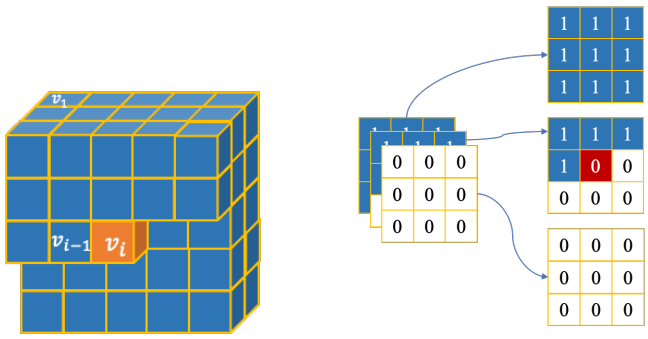
In this work, we propose a learning-based method for lossless compression of point cloud geometry. We aim at minimizing the encoded rate measured by the number of bits per occupied voxel (bpov) by exploiting the spatial redundancies within point cloud. The general scheme of our method is shown in Figure 1. A point cloud voxelized over a $2^n \times 2^n \times 2^n$ grid is known as an n-bit depth PC, which can be represented by an n level octree. In this work, we represent point cloud geometry in a hybrid manner, by combining the octree and voxel domains. We coarsely partition an n-depth point cloud up to level $n-6$. This allows to coarsely remove most of the empty space in the point cloud. As a result, we obtain a $n-6$ level octree and a number of non-empty binary blocks v of size $2^6 \times 2^6 \times 2^6$ voxels, which we refer to as resolution $d = 64$ or simply block 64 (Figure 1(a)). Blocks 64 can be further partitioned at resolution $d = \{64, 32, 16, 8, 4\}$ corresponding to maximum partitioning level $maxLv = \{1, 2, 3, 4, 5\}$ as detailed in Section III-C. Figure 1(b) shows the multi-resolution encoder with $maxLv = 2$. A block of size d can be encoded as a single block (b2) or partitioned into 8 sub-cubes (b1). We then encode each non-empty block (blocks in blue in the figure) using the proposed method in the voxel domain (Section III-B) and select the partitioning mode resulting in the smallest bpov. The overview of a single block encoder is shown in Figure 1(c). Our context model predicts the distribution of each voxel given all encoded voxels and pass it to an arithmetic coder to generate the final bitstream. The context is chosen adaptively following a rate optimization algorithm (Section III-C). The high-level octree, partitioning signal, selected context as well as the depth of each block are converted to bytes and signaled to the decoder as side information. We first define a 3D raster scan order that scan voxel by voxel in depth, height and width order. For ease of notation, we index all voxels in block v at resolution d from 1 to d^3 in raster scan order with:

$$v_i = \begin{cases} 1, & \text{if } i^{th} \text{ voxel is occupied} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

B. VoxelDNN

Our method losslessly encodes the voxelized point cloud using context-adaptive binary arithmetic coding. Specifically, we focus on estimating accurately a probability model $p(v)$ for the occupancy of a block v composed by $d \times d \times d$ voxels. We factorize the joint distribution $p(v)$ as a product of conditional distributions $p(v_i|v_{i-1}, \ldots, v_1)$ over the voxel volume:

$$p(v) = \prod_{i=1}^{d^3} p(v_i|v_{i-1}, v_{i-2}, \ldots, v_1). \quad (2)$$



(a) 3D voxel context (b) 3D type A mask

Fig. 2: (a): Example 3D context in a $5\times5\times5$ block. Previously scanned elements are in blue. (b): $3\times3\times3$ 3D type A mask. Type B mask is obtained by changing center position (marked red) to 1.

Each term $p(v_i|v_{i-1},\ldots,v_1)$ above is the probability of the voxel v_i being occupied given the occupancy of all previous voxels, referred to as a context. Figure 2(a) illustrates such a 3D context. We estimate $p(v_i|v_{i-1},\ldots,v_1)$ using a neural network which we dub **VoxelDNN**.

The conditional distributions in (2) depend on previously decoded voxels. This requires a *causality* constraint on the VoxelDNN network. To enforce causality, we extend to 3D the idea of masked convolutional filters, initially proposed in PixelCNN [22]. Specifically, two kinds of masks (A or B) are employed. Type A mask is filled by zeros from the center position to the last position in raster scan order as shown in Figure 2(b). Type B mask differs from type A in that the value in the center location is 1 (colored in red). Type A masks are used in the first convolutional filter to remove the connections between all future voxels and the voxel currently being predicted. From the second layer, the value of the current voxel is not used in its spatial position and is replaced by the result of the convolution over previous voxels. As a result, from the second convolutional layer, type B masks are applied which relaxes the restrictions of mask A by allowing the connection from the current spatial location to itself.

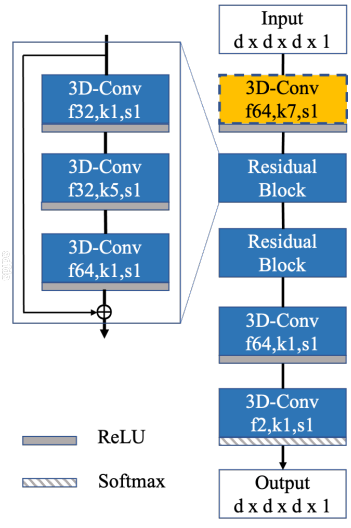


Fig. 3: VoxelDNN architecture, d is the dimension of the input block, masked layers are colored in yellow and blue. A type A mask is applied to the first layer (dashed borders) and type B masks afterwards. 'f64,k7,s1' stands for 64 filters, kernel size 7 and stride 1. Only probabilities of voxels being occupied are kept after the last Softmax layer.

Algorithm 1: Block partitioning selection

Input: block: B, current level: $curLv$, max level: $maxLv$
Output: partitioning flags: fl, output bitstream: $bits$

Function `partitioner`$(B, curLv, maxLv)$**:**
  $fl2 \leftarrow 2$; `// encode as 8 child blocks`
  for *block b in child blocks of* B **do**
    if *b is empty* **then**
      $child\_flag \leftarrow 0$;
      $child\_bit \leftarrow empty$;
    else
      if $curLv == maxLv$ **then**
        $child\_flag \leftarrow 1$;
        $child\_bit \leftarrow$ singleBlockCoder(b);
      else
        $child\_flag, child\_bit \leftarrow$ partitioner(b, $curLv+1$,$maxLv$);
      end
    end
    $fl2 \leftarrow [fl2, child\_flag]$;
    $bit2 \leftarrow [bit2, child\_bit]$;
  end
  $total\_bit2 = sizeOf(bit2) + len(fl2) \times 2$;
  $fl1 \leftarrow 1$; `// encode as a single block`
  $bit1 \leftarrow$ singleBlockCoder(B);
  $total\_bit1 = sizeOf(bit1) + len(fl1) \times 2$;
  `/* partitioning selection */`
  if $total\_bit2 \geq total\_bit1$ **then**
    return $fl1, bit1$;
  else
    return $fl2, bit2$;
  end

In order to learn good estimates $\hat{p}(v_i|v_{i-1},\ldots,v_1)$ of the underlying voxel occupancy distribution $p(v_i|v_{i-1},\ldots,v_1)$, and thus minimize the coding bitrate, we train VoxelDNN using cross-entropy loss. That is, for a block v of resolution d, we minimize :

$$H(p,\hat{p}) = \mathbb{E}_{v\sim p(v)}\left[\sum_{i=1}^{d^3} -\log\hat{p}(v_i)\right]. \quad (3)$$

It is well known that cross-entropy represents the bitrate cost to be paid when the approximate distribution $\hat{p}$ is used instead of the true distribution p. More precisely, $H(p,\hat{p}) = H(p) + D_{KL}(p\|\hat{p})$, where D_{KL} denotes the Kullback-Leibler divergence and $H(p)$ is Shannon entropy. Hence, by minimizing (3), we indirectly minimize the distance between the estimated conditional distributions and the real data distribution, yielding accurate contexts for arithmetic coding. Note that this is different from what is typically done in learning-based *lossy* PC geometry compression, where the focal loss is used [31], [32]. In this lossy context, the motivation behind using focal loss is to cope with the high spatial unbalance between occupied and non-occupied voxels. The reconstructed PC is then obtained by hard thresholding $\hat{p}(v)$, and the target is thus the final classification accuracy. Conversely, here we

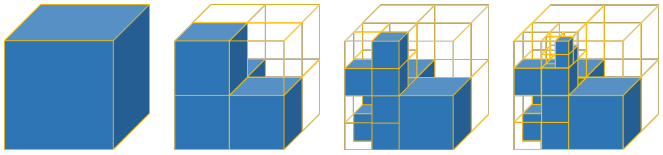
aim at estimating accurate soft probabilities to be fed into an arithmetic coder.

Figure 3 shows our VoxelDNN network architecture for a block of dimension d. Given the $d \times d \times d$ input block, VoxelDNN outputs the predicted occupancy probabilities of all input voxels. Our first 3D convolutional layer uses $7 \times 7 \times 7$ kernels with a type A mask. Type B masks are used in the subsequent layers. To avoid vanishing gradients and speed up the convergence, we implement two residual blocks [40] with $5 \times 5 \times 5$ kernels. Since type A masks are applied at the first layer, identity skip connection of residual block does not violate the causality constraint. Throughout VoxelDNN, the ReLu activation function is applied after each convolutional layer, except in the last layer where we use softmax activation. Using more filters generally increases the performance of VoxelDNN, at the expense of an increase in the number of parameters and computational complexity. After experimenting with various number of filters, we concluded that for input voxel block ($d \times d \times d \times 1$) which only has a single feature, 64 convolutional filters give a good trade-off between complexity and model performance.

C. Multi-resolution encoder and adaptive partitioning

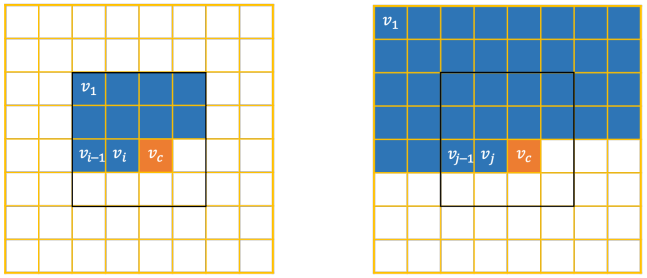
We use an arithmetic coder to encode the voxels sequentially from the first voxel to the last voxel of each block in a generative manner. Specifically, every time a voxel is encoded, it is fed back into VoxelDNN to predict the probability of the next voxel. Note that at this prediction step, all future voxels are filled with zeros. Then, we pass the probability to the arithmetic coder to encode the next symbol.

However, applying this coding process at a fixed resolution d (in particular, on larger blocks) can be inefficient when blocks are sparse, i.e., they contain only a few occupied voxels. This is due to the fact that in this case, there is little or no information available in the receptive fields of the convolutional filters. To overcome this problem, we propose to optimize the block size based on a rate-optimized multi-resolution splitting algorithm as follows. We partition a block into 8 sub-blocks recursively and signal the occupancy of sub-blocks as well as the partitioning decision (0: empty, 1: encode as a single block, 2: further partition). The partitioning decision depends on the bit rate after arithmetic coding. If the total bitstream of partitioning flags and occupied sub-blocks is larger than encoding the parent block as a single block, we do not perform partitioning. The details of this process are shown in Algorithm 1. The maximum partitioning level or the maximum number of block sizes is controlled by $maxLv$ and partitioning is performed up to $maxLv = 5$ corresponding to a smallest block size of 4. Depending on the output bits of each partitioning solution, a block of size 64 can contain a combination of blocks with different sizes. Figure 4 shows 4 partitioning examples for an encoder with $maxLv = 4$. Note that VoxelDNN learns to predict the distribution of the current voxel based on previously encoded voxels. As a result, we can use a bigger model size to predict the probabilities for smaller input block size.



(a) (b) (c) (d)

Fig. 4: Partitioning a block of size 64 into: (a) a single block of size 64, (b): blocks of size 32, (c): 32 and 16, (d): 32, 16 and 8. Non-empty blocks are indicated by blue cubes.



(a) (b)

Fig. 5: 2D illustration of context extension from block 4×4 to block 8×8. (a): Before extension, (b): after extension. Blue squares are active voxels in the context, voxels in the white area are ignored by masks or from the bigger block.

D. Context extension

We have discussed our multi-resolution encoder with multiple block sizes to adapt to the point cloud structure. However, with smaller block sizes, an implicit context model (using the content of the block) will be less efficient because the context may be too small. Therefore, we extend the context of each block to the encoded voxels that are above and on the left of the current voxel (causality constraint). Figure 5 illustrates the context before and after extension. Before extending the context, to encode voxel v_c, only voxels from v_1 to v_{i-1} in Figure 5(a) are considered as contexts. After extending the context to the bigger block, the context is now composed of all voxels in the blue area in Figure 5(b). The white area represent inactive voxels, i.e., not used in Eq. (2). Extending the context does not change the partitioning algorithm discussed above, although it might change the optimal selected partitions. Also, the causality is still enforced as long as we use masked filters in our network.

However, extending to a larger context is not always efficient when the extension area is sparse or contains noise, therefore we employ a rate-optimized block extension decision. To limit the computational complexity, we only allow certain combinations of block sizes and extension sizes, as shown in Table I. To encode a block with context extension, in Algorithm 1, we encode a block with all the possible extension sizes and select the best one in terms of bpov. In total, we build 5 models for 5 input sizes which are $\{128, 64, 32, 16, 8\}$ in the context extension mode.

E. Data augmentation

In order to train more robust probability estimation models and to increase the generalization capabilities of our model,

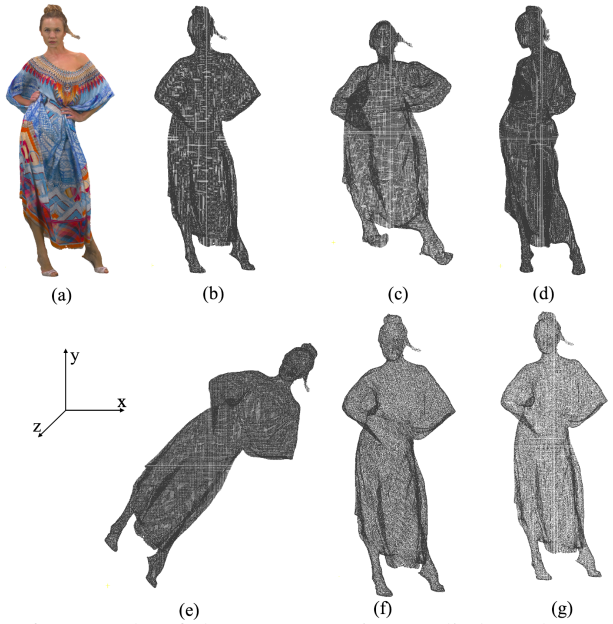


Fig. 6: Example of data augmentation applied on the Longdress point cloud. (a) Original; (b) After removing color attributes; (c),(d),(e) Rotation with $\theta = 45°$ on x, y and z axis; (f),(g) Sampling rate $f_s = 0.7$ and $f_s = 0.4$

we employ data augmentation techniques specifically suited for PCC. In particular, we observed that methods based on convolutional neural networks are especially sensitive to changes in PC density and acquisition noise. Therefore, in addition to typical rotation and shifting data augmentation used for other PC analysis tasks [37], [41], we also consider here alternative techniques, such as downsampling. Note that even though our VoxelDNN operates on voxel domain, to reduce the complexity, all input pipelines process point clouds in the form of x, y, z coordinates before converting into dense block in the final step. Specifically, for each generated block from the training datasets, we rotate them by an angle θ around each x, y, z axis. In addition, to adapt to varying density levels of the test point clouds, we randomly remove points from the original block as well as rotated blocks by controlling the probability of an occupied voxel being kept f_s ($f_s \in [0, 1]$). Figure 6 shows our data augmentation methods applying on Longdress point cloud from MPEG.

IV. Experimental Results

A. Experimental Setup

1) Training dataset: We consider point clouds from different and varied datasets, including ModelNet40 [42] which contains 12,311 models from 40 categories and three smaller datasets: MVUB [43], MPEG CAT1 [44] and 8i [45], [46].

TABLE I: Extending block size

Block size	Extending block size
64	128,64
32	64,32
16	64,32,16
8	64,32,16,8

TABLE II: Training and Testing Point Clouds

Training Set			Test Set	
Point Cloud	# Fr	ρ	Point Cloud	ρ
MVUB, 10 bits			*MVUB, 10 bits, dynamic upper body*	
Andrews	6	1.70	Phil	1.64
David	5	1.65	Ricardo	1.77
Sarah	6	1.72		
8i, 10 bits			*8i, 10 bits, dynamic full body*	
Soldier	9	1.51	Redandblack	1.49
Longdress	9	1.52	Loot	1.43
CAT1, 10 bits			Thaidancer	1.68
Facade	1	1.20	Boxer	1.56
Egyptian mask	1	0.12	*CAT1, 10 bits, static cultural heritage*	
Statue klimt	1	0.89	Frog	1.13
Head	1	1.43	Arco Valentino	0.45
House w/o roof	1	1.21	Shiva	0.88
ModelNet40, 9 bits			*USP, 10 bits, static cultural heritage*	
200 largest PCs	200	1.53	BumbaMeuBoi	0.18
			RomanOilLight	0.94

TABLE III: Number of blocks in the training sets of each model.

	MVUB	**8i**	**CAT1**	**ModelNet40**	**Total**
Model 128	1516	1101	677	2860	6154
Model 64	5777	4797	2777	11147	24498
Model 32	22082	20436	15243	50611	108372
Model 16	87578	86106	45626	224951	444261
Model 8	354617	349760	180037	986253	1870667

We uniformly sample points from the mesh models from ModelNet40 and then scale them to voxelized point clouds with 9 bit precision. To enforce the fairness between the smaller datasets in which we select point clouds for testing, point clouds from MPEG CAT1 are sampled to 10 bit precision as in MVUB and 8i. In addition, we measure the *local density* ρ of a point cloud, computed as the average portion of occupied voxels in the blocks of size 64, that is:

$$\rho = \frac{1}{N_{\mathcal{B}}} \times \sum_{\mathcal{B}_i \in \mathcal{B}} \frac{100 \times \text{number of points in } \mathcal{B}_i}{64^3} \quad (\%) \qquad (4)$$

where $\mathcal{B}$ is the set of occupied blocks of size 64, and $N_{\mathcal{B}}$ is the cardinality of $\mathcal{B}$. The higher the value of ρ is, the denser the point cloud. The selected point clouds, number of frames as well as ρ of the training data are shown in Table II.

To train a VoxelDNN model of size d we divide all selected PCs into occupied blocks of size $d \times d \times d$. Table III reports the number of blocks from each dataset for training, with the majority coming from the ModelNet40 dataset. For the models trained with data augmentation, we apply rotation with $\theta = 45°$ on x, y, z axis and then sampling from all blocks with sampling rate $f_s = [0.7; 0.4]$. In total, we augment from each block to 12 variations in terms of density and rotation which significantly increase the volume and diversity of our training set.

2) Test data: We evaluate the performance of our approach on a diverse set of point clouds in terms of spatial density and content type. All selected point clouds are either used in MPEG Common Test Condition or JPEG Pleno Common Test Condition to evaluate point cloud compression methods. As shown in Table II the test PCs can be categorized into four sets:

- **MVUB**: Microsoft Voxelized Upper Bodies [43] - a dynamic voxelized point cloud dataset containing five subjects. For testing, we randomly select 2 frames from

Phil (frame number 10) and *Ricardo* (76) sequences which are both very dense (high ρ) with smooth surfaces.

- **8i**: Dense point clouds from 8i Labs. They are also dynamic voxelized point clouds but each sequence contains the full body of a human subject. In the test set, *loot* (1000) and *redandblack* (1510) are from 8i Voxelized Full Bodies (8iVFB v2) [45] while *boxer* and *thaidancer* are selected and downsampled to 10 bits from 8i Voxelized Surface Light Field (8iVSLF) dataset [46].
- **CAT1**: static point clouds for cultural heritage and other related 3D photography applications [44]. We select *Arco_Valentino_Dense_vox12, Frog_00067_vox12, and Shiva_00035_vox12* from this dataset and downsample to 10 bits to validate the performance of our method. PCs from this dataset are less dense compared to the previous two datasets. *Frog_00067* has smoother surfaces compared to the other two PCs which contain rough surfaces.
- **USP**: an inanimate dataset from the University of São Paulo, Brazil, related to cultural heritage with 10 bits geometry precision [47]. *BumbaMeuBoi* and *RomanOilLight* are two selected point clouds from this dataset. PCs from USP dataset have simple shape with smooth surfaces. *BumbaMeuBoi* is the sparsest PC in our test set with the smallest ρ.

Figure 7 illustrates the test point clouds.

3) Training procedure: We train 5 models for 5 input block sizes, i.e., 128, 64, 32, 16, 8. The mini-batch sizes are 1, 8, 64, 128, 128, respectively. Our models are implemented in TensorFlow and trained with Adam [48] optimizer, a learning rate of 0.001 for 80 epochs on a GeForce RTX 2080 GPU.[1]

B. Performance evaluation and ablation studies

In the following, we evaluate the performance of the proposed approach as well as the impact of its various components. We start with models without data augmentation nor context extension in order to study the optimal maximal partitioning depth for our method and establish a baseline for the evaluation. Next, on top of the best encoder in this experiment (**Baseline**), we separately add data augmentation (**Baseline + DA**) and context extension (**Baseline + CE**). Finally, **Baseline + DA + CE** incorporates both data augmentation and context extension. We compare our method against the state-of-the-art point cloud compression method G-PCC v12 from MPEG [3] which has a dedicated lossless geometry mode for static point clouds. We report the number of bits per occupied voxel (bpov) for each test point cloud and the average per dataset.

In all experiments, the high-level octree plus partitioning signal are directly converted to bytes without any compression. For the encoders with context extension, we signal the selected size using two bits (maximum 4 options on block 8). This information is also directly converted to bytes in the bitstream. On average, signaling bits account for 2.44% of the bitstream.

[1]The source code is available at https://github.com/Weafre/VoxelDNN_v2.

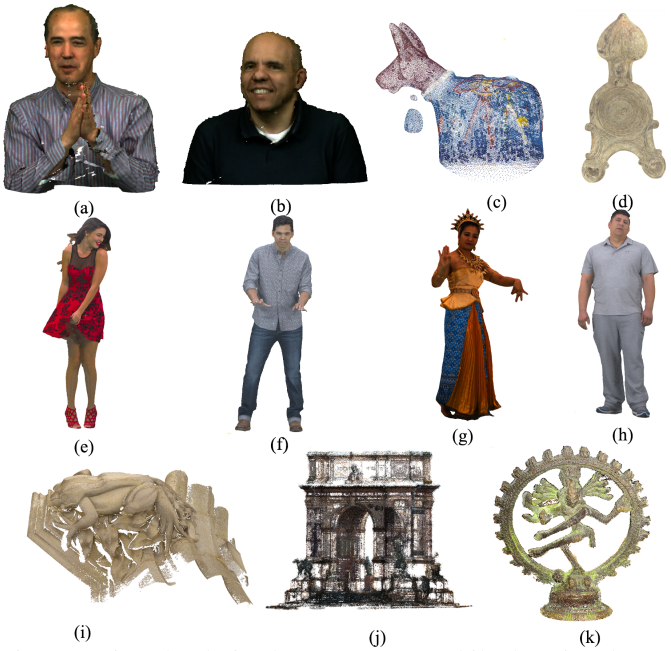


Fig. 7: Point clouds in the test set. (a) Phil, (b) Ricardo (c) BumbaMeuBoi (d) RomanOilLight, (e) Redandblack, (f) Loot, (g) Thaidancer (h) Boxer, (i) Frog, (j) Arco Valentino, (k) Shiva.

1) Optimal maximum partition depth: To evaluate the effectiveness of the partitioning scheme, we increase the maximum partitioning level from 1 to 5, corresponding to a minimum block size of 64, 32, 16, 8, and 4. As 3D convolution is not able to efficiently exploit voxel relations on a very small receptive field, we do not train a separate model for block 4. Instead, we use the model trained on blocks of size 8 to predict its probabilities.

Table IV shows the average bpov of our encoder on the 4 test datasets at 4 partitioning levels. The results with 5 partitioning levels are identical to 4 partitioning levels and are not shown in the table. We observe that, as partitioning levels increases, the corresponding gain over single-level also increases. However, adding the 3^{rd} and 4^{th} level yields only a slight improvement compared to adding the 2^{nd} level. This can be explained with Figure 8 showing the percentages of occupied voxels in each partition size. We observe that most voxels are encoded using blocks 64 and 32, while very few voxels are encoded using blocks of smaller size. Adding more partitioning levels enables to better adapt to point cloud geometry, however, this is not often compensated by a bitrate reduction of the sub-blocks, since in the smaller partitions the encoder has access to limited contexts, resulting in less accurate probability estimation. However, there is an increase in the portion of block 32 and 16 on CAT1 and USP compared to MVUB and 8i. This reflects the density characteristics of each dataset: on sparser datasets (CAT1 and USP), the algorithm tends to partition point cloud into smaller blocks to eliminate as much empty space as possible. Based on these observations, we use a maximum of 4 partitioning levels for our baseline codec in later experiments.

TABLE IV: Average rate in bpov per dataset at different partitioning levels and the gain over the encoder with 1 partitioning level.

		1 level	**2 levels**		**3 levels**		**4 levels**	
Dataset	Point Cloud	bpov	bpov	Gain	bpov	Gain	bpov	Gain
MVUB	Phil	0.8943	0.8295	-7.25%	0.8206	-8.24%	0.8205	-8.25%
	Ricardo	0.8109	0.7511	-7.37%	0.7440	-8.25%	0.7440	-8.25%
	Average	**0.8256**	**0.7903**	**-7.31%**	**0.7823**	**-8.25%**	**0.7823**	**-8.25%**
8i	Redandblack	0.7920	0.7269	-8.22%	0.7191	-9.20%	0.7190	-9.22%
	Loot	0.7017	0.6347	-9.56%	0.6271	-10.63%	0.6271	-10.63%
	Thaidancer	0.7941	0.7360	-7.32%	0.7298	-8.10%	0.7297	-8.11%
	Boxer	0.6462	0.5960	-7.77%	0.5901	-8.68%	0.5900	-8.70%
	Average	**0.7335**	**0.6734**	**-8.22%**	**0.6665**	**-9.15%**	**0.6665**	**-9.16%**
CAT1	Frog	1.9497	1.8406	-5.60%	1.8216	-6.57%	1.8214	-6.58%
	Arco Valentino	5.4984	5.2947	-4.52%	5.2051	-5.33%	5.2050	-5.34%
	Shiva	3.7964	3.6632	-3.51%	3.6400	-4.01%	3.6403	-4.11%
	Average	**3.7482**	**3.5845**	**-4.54%**	**3.5569**	**-5.31%**	**3.5556**	**-5.34%**
USP	BumbaMeuBoi	6.3618	5.8235	-8.46%	5.7305	-9.92%	5.7305	-9.92%
	RomanOilLight	1.8708	1.7157	-5.14%	1.7030	-5.84%	1.7030	-5.84%
	Average	**4.0853**	**3.7696**	**-6.80%**	**3.7168**	**-7.88%**	**3.7168**	**-7.88%**

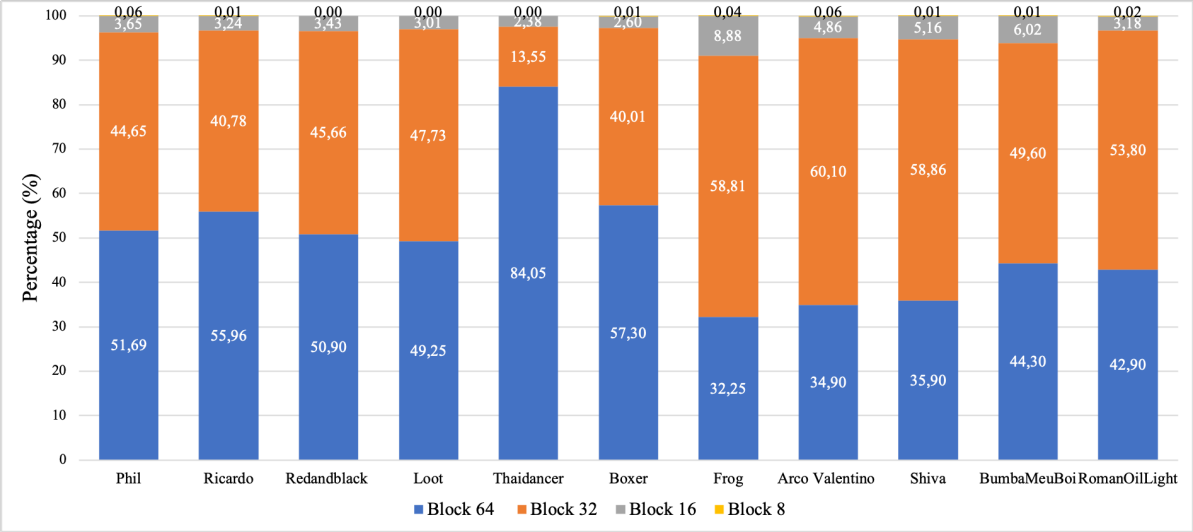


Fig. 8: Percentage of occupied voxels encoded in each partition size. From top to bottom: block 8, 16, 32, 64. Most of occupied voxel are encoded in block 64 and block 32.

2) Comparison with G-PCC: In table V, we report the output bitrate of our methods to compare with MPEG G-PCC. Both our method and G-PCC perform better on dense PCs while having higher rates on sparser PCs. Compared to G-PCC, the **Baseline** encoder obtains a significant gain – over 29% bitrate reduction on dense point clouds from MVUB and 8i dataset. On CAT1 and USP datasets, our method achieves a comparable rate with G-PCC. In particular, for Arco Valentino and BumbaMeuBoi, the two point clouds having the lowest ρ, our baseline codec yields a rate higher than G-PCC ($+7.25\%$ and $+5.99\%$, respectively). For point clouds with high local density levels, our VoxelDNN could efficiently leverage the relations between voxels and predict more accurate probability. In contrast, probability prediction is less accurate on sparser point clouds.

This can be partially solved by adding data augmentation during training. Indeed, by random subsampling the point clouds in the training set, VoxelDNN learns to predict more accurate probabilities when the point cloud is less dense. **Baseline + DA** yields higher gains over G-PCC on CAT1 and USP compared to **Baseline**, with average bitrate reductions of about 1.68% and 3.67%, respectively. On the other hand, we observe a small degradation of the performance compared to **Baseline** for denser datasets, such as MVUB and 8i dataset. This is somehow expected, as data augmentation increases the generalization capability of VoxelDNN, which instead is more adapted to denser PCs in the baseline mode.

The encoder with context extension, **Baseline + CE**, obtains a better rate on all test point clouds compared to the **Baseline** encoder, regardless of the density, with an average further bitrate reduction of 4.8% over G-PCC. The cost to be paid for this performance improvement is a higher computational complexity in the encoding process.

The last two columns of Table V show the experiment results for the encoder incorporating both data augmentation and context extension, **Baseline + DA + CE**. On average, we have a higher gain than **Baseline** and **Baseline + DA** because of the Context Extension. As expected, comparing with **Baseline + CE**, **Baseline + DA + CE** has increasing gains on CAT1 and USP datasets while obtaining a lower gain on MVUB and 8i datasets. Despite the different performance trends for different densities of the input point clouds, we obtain, on average, a bitrate reduction of 20.66% compared to

TABLE V: Average rate in bpov of proposed method and percentage gains compared with MPEG G-PCC v12 (negative percentages mean bitrate reduction).

		G-PCC	**Baseline**		**Baseline + DA**		**Baseline + CE**		**Baseline + DA + CE**	
Dataset	Point Cloud	bpov	bpov	Gain over G-PCC	bpov	Gain over G-PCC	bpov	Gain over G-PCC	bpov	Gain over G-PCC
MVUB	Phil	1.1599	0.8205	-29.26%	0.8954	-22.80%	0.7601	-34.47%	0.8252	-28.86%
	Ricardo	1.0673	0.7440	-30.29%	0.8235	-22.84%	0.6874	-35.59%	0.7572	-29.05%
	Average	**1.1136**	**0.7823**	**-29.78%**	**0.8595**	**-22.82%**	**0.7238**	**-35.01%**	**0.7912**	**-28.95%**
8i	Redandblack	1.0893	0.7190	-33.90%	0.7772	-28.65%	0.6645	-39.00%	0.7003	-35.71%
	Loot	0.9524	0.6271	-34.16%	0.6282	-34.04%	0.5766	-39.46%	0.6084	-36.12%
	Thaidancer	0.9990	0.7297	-26.96%	0.7253	-27.40%	0.6769	-32.23%	0.6627	-33.66%
	Boxer	0.9492	0.5900	-37.81%	0.6573	-30.75%	0.5503	-41.94%	0.5906	-37.78%
	Average	**0.9975**	**0.6665**	**-33.24%**	**0.6870**	**-30.21%**	**0.6171**	**-38.13%**	**0.6405**	**-35.79%**
CAT1	Frog	1.8990	1.8214	-4.09%	1.7662	-6.99%	1.6971	-10.63%	1.7071	-10.11%
	Arco Valentino	4.8531	5.2050	+7.25%	5.0639	+4.34%	4.9923	+2.87%	4,9900	+2.82%
	Shiva	3.6716	3.6403	-0.85%	3.5838	-2.39%	3.4619	-5.71%	3.5135	-4.31%
	Average	**3.4746**	**3.5556**	**+0.77%**	**3.7413**	**-1.68%**	**3.3838**	**-2.61%**	**3.4035**	**-3.86%**
USP	BumbaMeuBoi	5.4068	5.7305	+5.99%	5.3831	-0.44%	5.3580	-0.90%	5.066	-6.29%
	RomanOiLight	1.8604	1.7030	-8.46%	1.7319	-6.91%	1.6130	-13.30%	1.6231	-12.76%
	Average	**3.6336**	**3.7168**	**-1.24%**	**3.5575**	**-3.67%**	**3.4855**	**-7.10%**	**3.4855**	**-9.52%**

G-PCC. Note that, in practice, if the characteristics of point cloud to be coded are known in advance, our approach is flexible, in that we could deploy different models targeting a specific application (cultural heritage, tele-immersive conferencing, etc.) and content type to obtain the best compression rate.

3) Effect of PC content and density on coding performance: In order to better understand the performance of our codec for different types of content, we plot in Figure 9 the average bpov as a function of the percentage of occupied voxels for each block 64 of *Phil, Loot, Arco Valentino* and *BumbaMeuBoi* with the **Baseline + DA + CE** encoder. Notice that each block 64 can be split up to different partition levels, indicated by the size of the dots in the figure. The distribution of the density of blocks 64 is shown in the top panel.

From this figure, we can draw some observations. First, most of blocks are partitioned into 3 levels (smallest dots) and the majority of the remaining blocks are partitioned into 2 or 4 levels. Second, in each point cloud, denser blocks are easier to compress, as mentioned before, due to the better capabilities of convolution to capture spatial relations. On the other hand, our approach becomes inefficient when the blocks are less dense, and the bitrate associated to the very sparse blocks rapidly grows by an order of magnitude compared to the rest. This phenomenon is true for all kinds of contents, although it has a stronger effect when the block density distribution is skewed to the left, such as for *Arco Valentino* or *BumbaMeuBoi*, which have the highest bitrates in our experiments.

We can also observe a content-dependence trend in the figure, which appears like a vertical offset for different PCs. *Arco Valentino* and *RomanOilLight* overall have higher bpov compared to *Phil* and *Loot* with the same number of occupied voxels. This suggests that local density alone is not the only factor affecting the performance of our approach, but that somehow higher-order statistics enter into play. We will speculate more about this behaviour when discussing the bitrate allocation in Figure 11. Further analysis of this trend, as well as how to take better into account the PC characteristics to improve coding performance, are left to future work.

4) Selection of context extension and impact on the partitioning: Figure 10 shows how many times an extended block size is selected in the **Baseline + DA + CE** experiments. First, it can be seen that in most cases our encoder choose to extend the context to encode the current block, and mostly the immediate larger size is selected. By extending context to exploit geometry information from the neighboring voxels, VoxelDNN can leverage a larger amount of information and predict a better probability. In most cases where the encoder does not extend the context, the blocks are on the border of the volume, corresponding to a mostly empty extending area.

By summing the quantities in each column, we obtain the number of blocks which are encoded using each block size and we observe that large parts of the point cloud are partitioned into block 32 or 16. This is in contrast with the previous observation on baseline experiments where the most frequent partitions are 64 and 32 (Figure 8). This has an intuitive explanation: without context extension, small block sizes of 32 or 16 were insufficient to provide a representative enough context for VoxelDNN in most of the cases, even if they would better adapt to areas with low point density. Conversely, the context extension allows to compensate for the small block dimension and renders these modes competitive. As a result, context extension significantly affects the optimal partitioning and enables VoxelDNN to adapt better to local sparsity while still providing enough contextual information to predict accurate probabilities.

5) Using multiple models for the context: For the multi-resolution encoder, instead of using a separate model for each block size, VoxelDNN can use only a single neural network to predict the distribution. Specifically, we place small blocks (8, 16, 32) into a block of size 64 and then use the network for block 64 to predict and extract the corresponding distributions. This method of computing the occupancy distribution is different from Context Extension in that the surrounding voxels are always set to 0. In Table VI, we compare the performance of using a single model with **Baseline**, which is a multi-models encoder. In this experiment, both encoders have 4 maximum partitioning levels and use the same model 64. On average, by having a separate model for each block size,

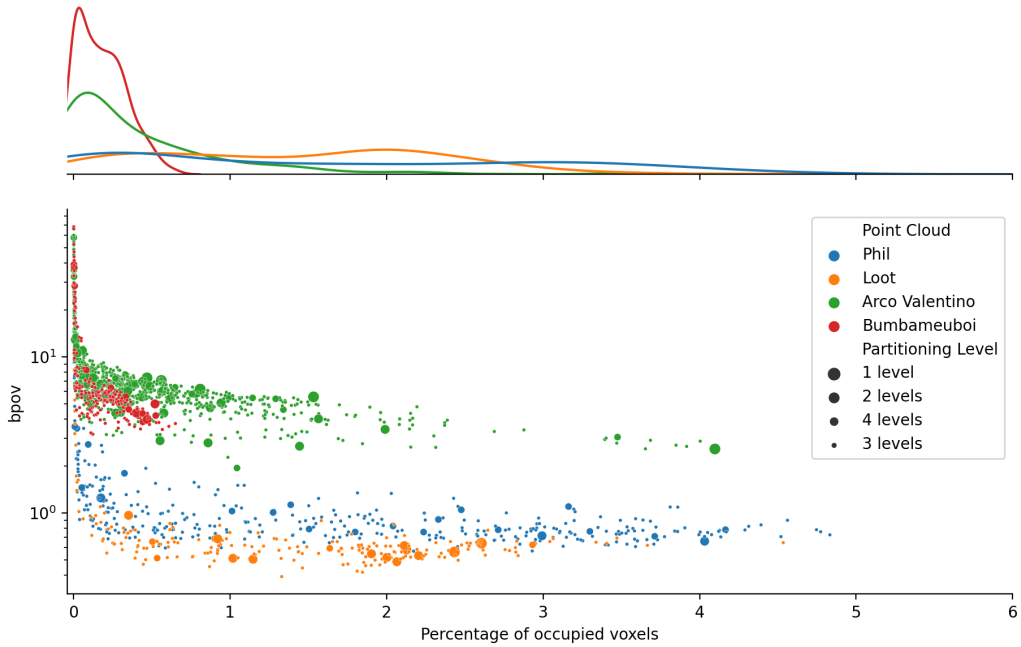


Fig. 9: Performance on block 64 on four test point clouds. Each point corresponds to a block 64 with percentage of occupied voxels (ρ) and bpov (log scale) performance of that block. The size of each point indicates the partitioning level and each partitioning level was sized according to its frequency. Higher points indicate that VoxelDNN is performing worse. The marginal distributions of occupied voxels for each point cloud are on the top of the scatter plot.

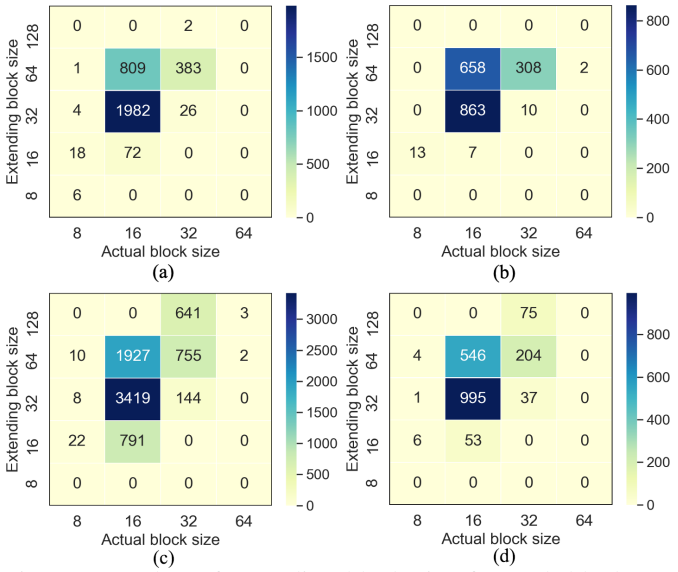


Fig. 10: Number of extending block size for each block. (a) Phil, (b) Loot, (c) Arco Valentino, (d) BumbaMeuBoi. Most of the time, the encoder extend the context to neighboring voxels instead of independently encoding a block.

a multi-model encoder obtains about 1% additional gain over G-PCC compared to the single model encoder. This amount of gain indicates that the bigger VoxelDNN model can predict the conditional distribution on smaller blocks as efficiently as using a separate model for each block size. However, model 64 is trained on blocks of size 64 only, and learns features at that scale. In general, a model trained on small blocks could better capture the context from small input blocks and thus provides a higher gain in some circumstances.

TABLE VI: Single model and multi-models comparison.

		G-PCC	Single model		Multi-models	
	Point Cloud	bpov	bpov	Gain over G-PCC	bpov	Gain over G-PCC
MVUB	Phil	1.1599	0.8312	-28.33%	0.8205	-29.26%
	Ricardo	1.0673	0.7541	-29.34%	0.7440	-30.29%
	Average	**1.1136**	**0.7927**	**-28.81%**	**0.7823**	**-29.78%**
8i	Redandblack	1.0893	0.7320	-32.80%	0.7190	-33.99%
	Loot	0.9524	0.6403	-32.77%	0.6271	-34.16%
	Thaidancer	0.9990	0.7305	-26.87%	0.7297	-26.96%
	Boxer	0.9492	0.6008	-36.70%	0.5900	-37.84%
	Average	**0.9975**	**0.6759**	**-32.24%**	**0.6665**	**-33.24%**
CAT1	Frog	1.8990	1.8433	-2.93%	1.8214	-4.09%
	Arco Valentino	4.8531	5.2173	+7.50%	5.2050	+7.25%
	Shiva	3.6716	3.6595	-0.32%	3.6403	-0.85%
	Average	**3.4746**	**3.5734**	**+2.84%**	**3.5556**	**+0.77%**
USP	BumbaMeuBoi	5.4068	5.7501	+6.34%	5.7305	+5.10%
	RomanOiLight	1.8604	1.7094	-8.11%	1.7030	-8.46%
	Average	**3.6336**	**3.7298**	**-2.64%**	**3.7168**	**-1.24%**

6) Visualization of the bitrate allocation on coded PCs: The bpov heatmaps of 4 point cloud are shown in Figure 11. The blocks in the figures reflect the optimal partitioning obtained by the algorithm. First, we visually confirm what found in Figure 9, i.e., VoxelDNN performs better, i.e., achieves a small bitrate, in the smooth and dense areas of the point cloud. Conversely, in the noisy areas (*Phil*'s hand, *Loot*'s hand), sudden holes (*Arco Valentino*) or very sparse regions (edges in *Arco Valentino*, the bottom part of *BumbaMeuBoi*), which are indicated in red, the performance is worse. We can argue that the density of a point cloud, together with the smoothness and noise characteristics of the content, are among the main factors that influence the performance of VoxelDNN. On the other hand, we can argue that noisy and very sparse areas are intrinsically difficult to code in general, and indeed also the MPEG G-PCC codec requires a large number of bits to encode point clouds such as *BumbaMeuBoi* and *Arco Valentino*.

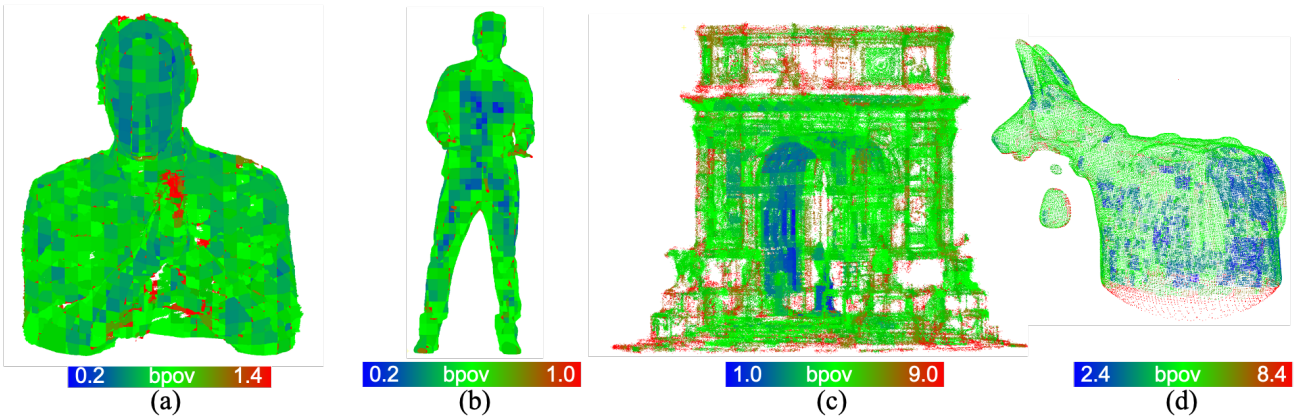


Fig. 11: Output geometry bitrate in bpov per block. (a) Phil, (b) Loot, (c) Arco Valentino, (d) BumbaMeuBoi. The heatmap bar below each subfigure shows the minimum and maximum bpov and the corresponding color.

TABLE VII: Average runtime (in seconds) of different encoders comparing with G-PCC v12

	G-PCC	Baseline	Baseline + CE
(Enc)	2.90	298	885
(Dec)	2.79	672	640

C. Computational complexity analysis

A well-known drawback of auto-regressive generative models such as PixelCNN and VoxelDNN is the sequential generation of the symbol probabilities. This requires to run the network for each voxel, which has a complexity that increases linearly with the number of voxels. Therefore, VoxelDNN has a computational complexity which is 3 orders of magnitures bigger than G-PCC.

Table VII reports the encoding and decoding time of our **Baseline** and **Baseline + CE**. Tests are benchmarked on an Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz machine with an Nvidia GeForce GTX 2080 GPU and 16 GB of RAM, running Ubuntu 16.04. Our encoding time is highly dependent on the number of blocks and the number of voxels within each block. Besides, the number of modes in the partitioning algorithm and context extension also influence the complexity. The **Baseline + CE** encoder tries all the extending modes and selects the best one, thus its average encoding time is higher than **Baseline** – an increase of about 196%. The reason why the encoding time for the **Baseline** codec is lower than the decoding time is purely implementative: at the encoder it is possible to predict the whole block probabilities in a single batch on a GPU, while in a realistic scenario, at the decoder side the voxels need to be individually decoded. When context extension is enabled, point clouds are partitioned into even smaller blocks, corresponding to a smaller complexity at the decoder, as a smaller number of voxels need to be decoded. On the other hand, the total parameters of each VoxelDNN model corresponds only to about 3.5 MB which is a small-size network in practice. Notice that the bottleneck in our system comes from the adoption of an auto-regressive model, which has the advantage of providing, in principle, an exact likelihood estimation of the data, though at a high computational cost. We are currently investigating the use of alternative generative approaches that avoid sequential probability estimation.

V. Conclusions and Future Work

This paper presents a lossless compression method for point cloud geometry. We extend a well-known auto-regressive generative model initially proposed for 2D images to the 3D voxel space, and we incorporate 3D data augmentation to efficiently exploit the redundancies between points. This approach enables to build accurate probability models for the arithmetic coder. As a result, when using an adaptive partitioning scheme and context extension, our solution outperforms MPEG G-PCC over a diverse set of point clouds.

Our analyses on the performance of the proposed method indicate at least two major avenues for improvement. On one hand, handling low-density point clouds would require to rethink the network architecture to handle sparse input data. On the other hand, a major drawback of VoxelDNN is the high computational cost of sequential probability generation, which we plan to replace in the future by a more efficient generative model.

References

[1] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuca, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko, "Emerging MPEG Standards for Point Cloud Compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, pp. 1–1, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8571288/

[2] E. S. Jang, M. Preda, K. Mammou, A. M. Tourapis, J. Kim, D. B. Graziosi, S. Rhyu, and M. Budagavi, "Video-Based Point-Cloud-Compression Standard in MPEG: From Evidence Collection to Committee Draft [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 118–123, May 2019.

[3] 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 Transactions on Signal and Information Processing*, vol. 9, 2020.

[4] R. L. De Queiroz and P. A. Chou, "Compression of 3d point clouds using a region-adaptive hierarchical transform," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3947–3956, 2016.

[5] R. Schnabel and R. Klein, "Octree-based point-cloud compression." *Spbg*, vol. 6, pp. 111–120, 2006.

[6] R. Mekuria, K. Blom, and P. Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2017.

[7] J. Kammerl, N. Blodow, R. B. Rusu, S. Gedikli, M. Beetz, and E. Steinbach, "Real-time compression of point cloud streams," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 778–785.

[8] D. C. Garcia and R. L. de Queiroz, "Context-based octree coding for point-cloud video," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 1412–1416, iSSN: 2381-8549.

[9] D. C. Garcia and R. L. d. Queiroz, "Intra-Frame Context-Based Octree Coding for Point-Cloud Geometry," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, pp. 1807–1811.

[10] D. C. Garcia, T. A. Fonseca, R. U. Ferreira, and R. L. de Queiroz, "Geometry Coding for Dynamic Voxelized Point Clouds Using Octrees and Multiple Contexts," *IEEE Transactions on Image Processing*, vol. 29, pp. 313–322, 2019.

[11] L. Huang, S. Wang, K. Wong, J. Liu, and R. Urtasun, "OctSqueeze: Octree-Structured Entropy Model for LiDAR Compression," *arXiv:2005.07178 [cs, eess]*, May 2020. [Online]. Available: http://arxiv.org/abs/2005.07178

[12] S. Biswas, J. Liu, K. Wong, S. Wang, and R. Urtasun, "Muscle: Multi sweep compression of lidar using deep entropy models," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] A. Dricot and J. Ascenso, "Adaptive multi-level triangle soup for geometry-based point cloud coding," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–6.

[14] "Inference of a mode using point location direct coding," in *TMC3, ISO/IEC JTC1/SC29/WG11 input document m42239, Gwangju, Korea, January 2018.*

[15] "Neighbour-dependent entropy coding of occupancy patterns," in *TMC3, ISO/IEC JTC1/SC29/WG11 input document m42238, Gwangju, Korea, January 2018.*

[16] "Intra mode for geometry coding," in *TMC3, ISO/IEC JTC1/SC29/WG11 input document m43600, Ljubljana, Slovenia, July 2018.*

[17] "Planar mode in octree-based geometry coding," in *TISO/IEC JTC1/SC29/WG11 input document m48906, Gothenburg, Sweden, July 2019.*

[18] "An improvement of the planar coding mode," in *ISO/IEC JTC1/SC29/WG11 input document m50642, Geneva, CH, Oct 2019.*

[19] W. Zhu, Y. Xu, L. Li, and Z. Li, "Lossless point cloud geometry compression via binary tree partition and intra prediction," in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.

[20] L. Theis and M. Bethge, "Generative image modeling using spatial lstms," in *Advances in Neural Information Processing Systems*, 2015, pp. 1927–1935.

[21] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[22] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel Recurrent Neural Networks," *arXiv:1601.06759 [cs]*, Aug. 2016. [Online]. Available: http://arxiv.org/abs/1601.06759

[23] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications," *arXiv:1701.05517 [cs, stat]*, Jan. 2017. [Online]. Available: http://arxiv.org/abs/1701.05517

[24] M. J. Weinberger, J. J. Rissanen, and R. B. Arps, "Applications of universal context modeling to lossless compression of gray-scale images," *IEEE Transactions on Image Processing*, vol. 5, no. 4, pp. 575–586, 1996.

[25] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," in *2017 5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: http://arxiv.org/abs/1611.01704

[26] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool, "Conditional Probability Models for Deep Image Compression," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 4394–4402. [Online]. Available: https://ieeexplore.ieee.org/document/8578560/

[27] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3155–3164.

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[29] A. F. R. Guarda, N. M. M. Rodrigues, and F. Pereira, "Point cloud coding: Adopting a deep learning-based approach," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5.

[30] ——, "Point cloud geometry scalable coding with a single end-to-end deep learning model," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3354–3358.

[31] M. Quach, G. Valenzise, and F. Dufaux, "Learning Convolutional Transforms for Lossy Point Cloud Geometry Compression," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 4320–4324, iSSN: 1522-4880.

[32] ——, "Improved Deep Point Cloud Geometry Compression," in *arXiv:2006.09043 [cs, eess, stat]*, Jun. 2020. [Online]. Available: http://arxiv.org/abs/2006.09043

[33] J. Wang, H. Zhu, Z. Ma, T. Chen, H. Liu, and Q. Shen, "Learned point cloud geometry compression," *arXiv preprint arXiv:1909.12037*, 2019.

[34] A. F. Guarda, N. M. Rodrigues, and F. Pereira, "Point cloud geometry scalable coding with a single end-to-end deep learning model," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3354–3358.

[35] e. Yan, S. Liu, T. H. Li, Z. Li, G. Li *et al.*, "Deep autoencoder-based lossy geometry compression for point clouds," *arXiv preprint arXiv:1905.03691*, 2019.

[36] T. Huang and Y. Liu, "3d point cloud geometry compression on deep learning," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 890–898.

[37] J. Wang, D. Ding, Z. Li, and Z. Ma, "Multiscale point cloud geometry compression," *arXiv preprint arXiv:2011.03799*, 2020.

[38] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[39] B. Graham and L. van der Maaten, "Submanifold sparse convolutional networks," *arXiv preprint arXiv:1706.01307*, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] I. Alonso, L. Riazuelo, L. Montesano, and A. C. Murillo, "3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation," 2020.

[42] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1912–1920.

[43] C. Loop, Q. Cai, S. O. Escolano, and P. A. Chou, "Microsoft voxelized upper bodies - a voxelized point cloud dataset," in *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document m38673/M72012*, May 2016.

[44] "Common test conditions for PCC," in *ISO/IEC JTC1/SC29/WG11 MPEG output document N19324*.

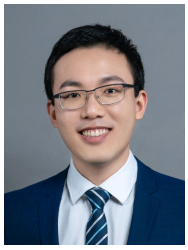
[45] E. d'Eon, B. Harrison, T. Myers, and P. A. Chou, "8i Voxelized Full Bodies - A Voxelized Point Cloud Dataset," in *ISO/IEC JTC1/SC29 Joint WG11/WG1 (MPEG/JPEG) input document WG11M40059/WG1M74006*, Geneva, Jan. 2017.

[46] "Common test conditions for PCC," in *ISO/IEC JTC1/SC29/WG11 MPEG output document N19324*.

[47] M. Zuffo, "University of sao paulo point cloud dataset," (accessed Dec 19, 2020). [Online]. Available: http://uspaulopc.di.ubi.pt

[48] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *2015 3rd International Conference on Learning Representations*, Dec. 2014, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

Dat Thanh Nguyen received the Engineer degree in Electronics and Telecommunications from Hanoi University of Science and Technology, Vietnam and the M.S. degree in Electronics from the Polytechnic Institute of Paris, France. He has been working as an intern, first, and as research engineer, at L2S from May 2020 to May 2021. He is currently a PhD candidate at at Chair of Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Germany.



Maurice Quach received the Computer Science Engineer Diploma from University of Technology of Compiègne in 2018. He is currently studying for a PhD on Deep Learning-based Point Cloud Compression under the supervision of Frederic Dufaux and Giuseppe Valenzise at L2S, Centrale Supélec, Université Paris-Saclay, France.

Giuseppe Valenzise is a CNRS researcher (chargé de recherches) at Université Paris-Saclay, CNRS, CentraleSupélec, in the Laboratoire des Signaux et Systèmes (L2S). He completed a master degree and a Ph.D. in Information Technology at the Politecnico di Milano, Italy, in 2007 and 2011, respectively. In 2012, he joined the French Centre National de la Recherche Scientifique (CNRS) as a permanent researcher, first at the Laboratoire Traitement et Communication de l'Information (LTCI) Telecom Paristech, and from 2016 at L2S. He got the French « Habilitation à diriger des recherches » (HDR) from Université Paris-Sud in 2019. His research interests span different fields of image and video processing, including traditional and learning-based image and video compression, light fields and point cloud coding, image/video quality assessment, high dynamic range imaging and applications of machine learning to image and video analysis. He is co-author of more than 70 research publications and of several award-winning papers. He is the recipient of the EURASIP Early Career Award 2018.

Giuseppe serves/has served as Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, Elsevier Signal Processing: Image communication. He is an elected member of the MMSP and IVMSP technical committees of the IEEE Signal Processing Society, as well as a member of the Technical Area Committee on Visual Information Processing of EURASIP.

Pierre Duhamel received the Eng. Degree in Electrical Engineering from the National Institute for Applied Sciences (INSA) Rennes, France in 1975, and the Dr. Eng. and the D.Sc degrees from Orsay University, Orsay, France in 1978 and 1986, respectively.

From 1975 to 1980, he was with Thomson-CSF, Paris, France, where his research interests included circuit theory and signal processing. In 1980, he joined the National Research Center in Telecommunications (CNET), Issy les Moulineaux, France, where his research activities were first concerned with the design of recursive CCD filters. Later, he worked on fast algorithms for computing various signal processing functions (FFT's, convolutions, adaptive filtering, and wavelets. From 1993 to Sept. 2000, he has been professor with ENST (National School of Engineering in Telecommunications), Paris with research activities focused on Signal processing for Communications. He was the head of the Signal and Image processing Department from 1997 to 2000. He is currently with CNRS/LSS (Laboratoire de Signaux et Systemes, Gif sur Yvette, France), where he developed studies in Signal processing for communications and signal/image processing for multimedia applications, including source/protocol/channel coding/decoding. He is also investigating the connections between communication theory and networking as well as information theory and AI. He has been "directeur de recherches émérite" since March 2019.

He has published more than 100 articles in international journals, more than 300 papers in international conferences, and holds 29 patents. He is a co-author of the book 'Joint Source and Channel Decoding: A cross layer perspective with applications in video broadcasting" which appeared in 2009, Academic Press. He successfully advised or co-advised more than 60 PhD students, and two of them are now fellows of the IEEE.

Dr. Duhamel is a fellow of EURASIP in 2008. He was awarded the "grand prix France Telecom" by the French Science Academy in 2000. He was a Distinguished lecturer, IEEE, in 1999, and was co-technical chair of ICASSP 06, Toulouse, France and WCNC 2012, Paris, France. He also held the "Jacques Beaulieu Excellence Chair" of Institut National de la Recherche Scientifique, Montreal, in 2015. A paper on subspace-based methods for blind equalization, which he co-authored, received the "Best paper award" from the IEEE transactions on Signal Processing in 1998.