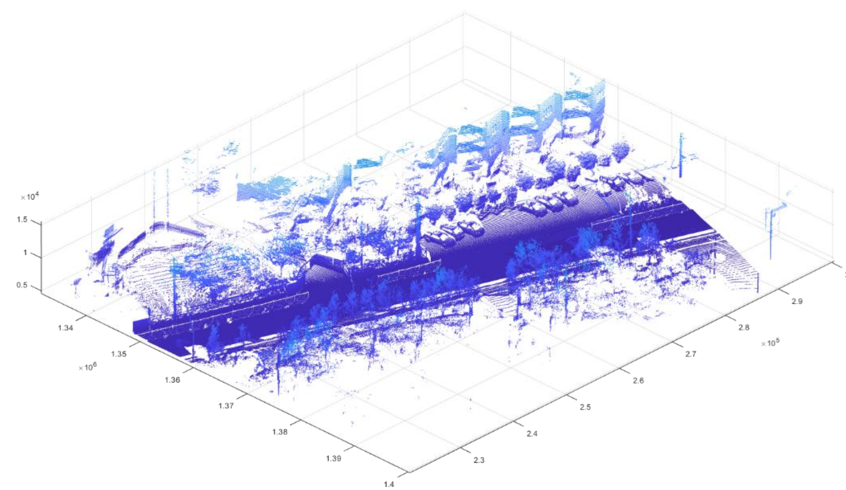
Deep Leanring in
Immersive Media Compression

Zhu Li
Director, UMKC NSF Center for Big Learning
Dept of Computer Science & Electrical Engineering
University of Missouri, Kansas City
Email: zhu.li@ieee.org, lizhu@umkc.edu
Web: http://l.web.umkc.edu/lizhu



Outline

- ❑ Short intro to the MCC Lab
- ❑ Research Motivation and Highlights
- ❑ Sparse Conv Engine based Point Cloud Compression
- ❑ Video based Point Cloud Compression
- ❑ Summary

Short Bio:

Research Interests:

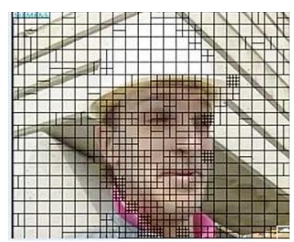
- Immersive Media Communication: light field, point cloud and 360 video capture, coding and low latency communication.
- Data & Image Compression: video, medical volumetric data, DNA sequence, and graph signal compression with deep learning
- Remote Sensing & Vision: vision problem under low resolution, blur, and dark conditions, hyperspectral imaging, sensor fusion
- Edge Computing & Federated Learning: gradient compression, light weight inference time engine, retina features, fast edge cache for video CDN

NSF I/UCRC Center *for* Big Learning
Creating Intelligence

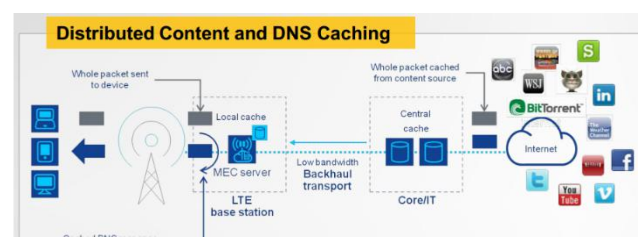
Zhu Li,
Associate Editor-in-Chief,
IEEE Trans on Circuits & System for Video Tech (T-CSVT)
Director, NSF Center for Big Learning at UMKC
Dept of Computer Science & Electrical Engineering
Univ of Missouri, Kansas City (UMKC)

signal processing and learning

image understanding



visual communication



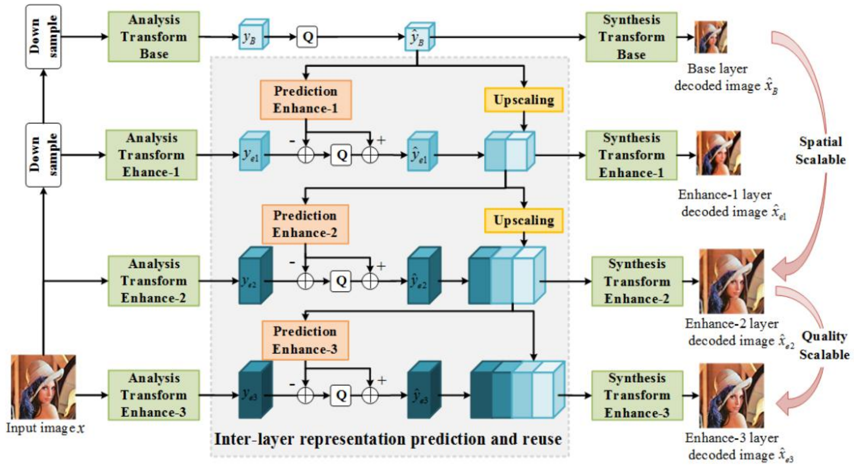
mobile edge computing & communication

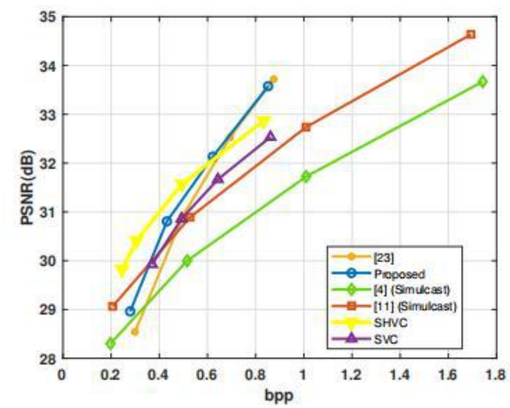
Multimedia Computing & Communication (MC^2) Lab

- MCC Lab@UMKC: FH-261 and FH-262
- Total $3.2m awarded from NSF, AFOSR and ONR, as well as various NSF I/UCRC industry members
- Published 30+ journal papers last 3 years in top notch venus like: T-IP, T-MM, T-CSVT, T-PAMI, IJCV, TOMM, TGRS.
- Currently 9 PhD students, 12 GPU workstation and 10 workstations.
- Recent PhD graduates:
  - 2 joined ByteDance video codec team
  - 1 joined Tuskegee Univ as AP
- Recent Post-docs:
  - 1 post-doc joined USTC as faculty
  - 1 post-doc now AP with NUIST

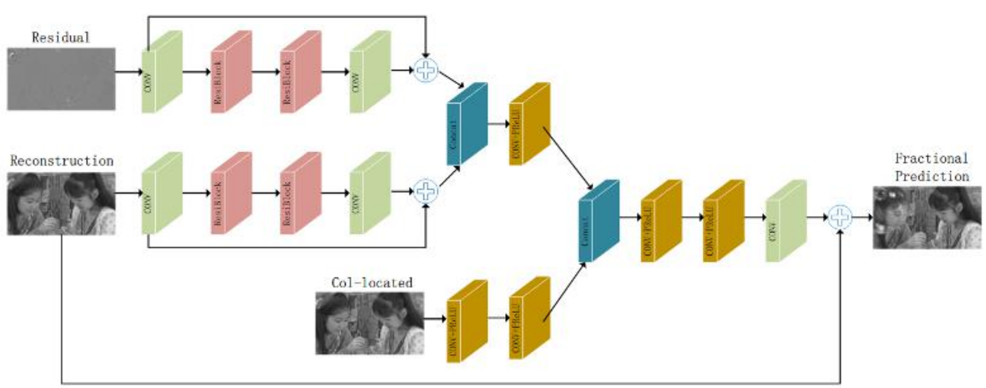
Data & Image Compression Highlights

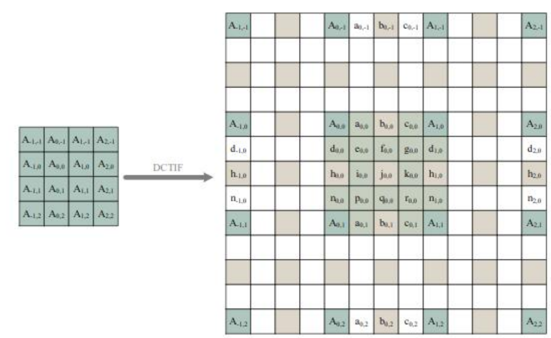
- Y. Mei, L. Li, Z. Li, and F. Li, "Learning-Based Scalable Image Compression with Latent-Feature Reuse and Prediction", *IEEE Trans on Multimedia* (T-MM), 2021.





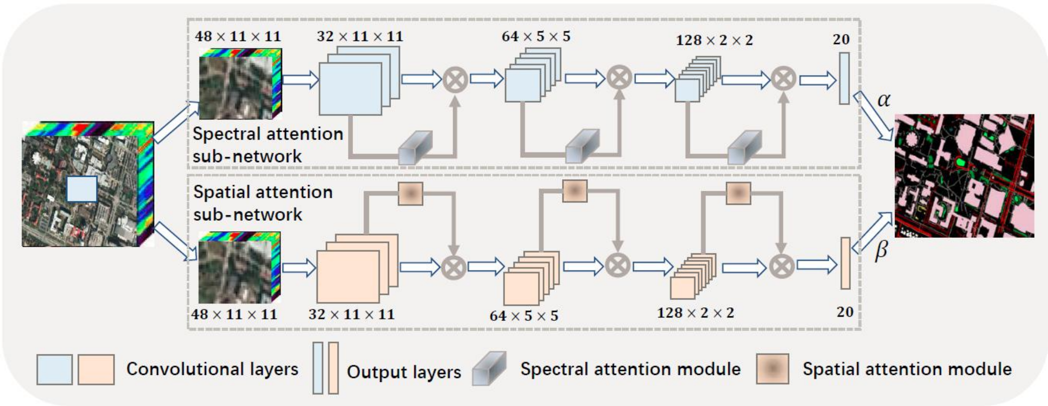
- H. Zhang, L. Song, L. Li, Z. Li, and X.K. Yang"Compression Priors Assisted Convolutional Neural Network for Fractional Interpolation", *IEEE Trans on Circuits and Systems for Video Tech,* 2020





Remote Sensing & Vision Highlights

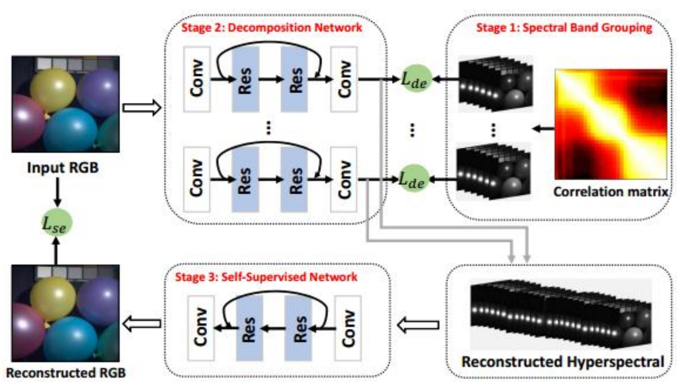
- R. Hang, Z. Li, Q. Liu, P. Ghamisi and S. Bhattacharyya, "**Hyperspectral Image Classification with Attention Aided CNNs**", *IEEE Trans. on Geoscience & Remote Sensing* (T-GRS), 2020. [Highly Cited]



Attention CNN for Hyperspectral Image Classification

- Introducing a dual stream network architecture with separate attention model for spatial and spectral feature maps
- Achieving the SOTA performance.

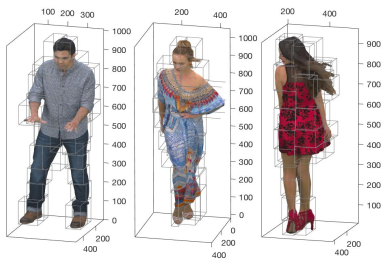
- R. Hang, Q. Liu, and Z. Li, "Spectral Super-Resolution Network Guided by Intrinsic Properties of Hyperspectral Imagery", *IEEE Trans on Image Processing* (T-IP), 2021



PRINET: Spectral Super Resolution

- Super-resolve hyper-spectral info from RGB inputs
- A dual loss network that learn a correlation decomposed HSI images
- Achieving the new SOTA performance.

Immersive Media Coding & Communication (NSF/IUCRC)



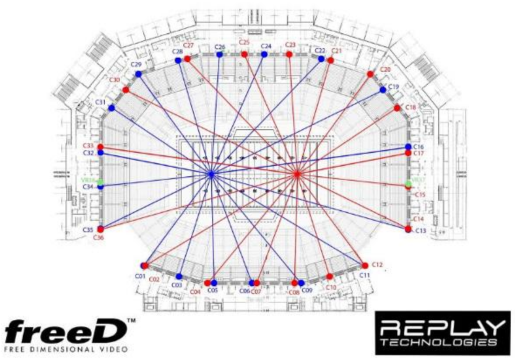

- "PU-Dense: Large Scale Photo-Realistic Point Cloud Upsampling", accepted with *IEEE Trans on Image Processing* (T-IP), 2022.
- "Deep Learning Geometry Compression Artifacts Revomal for Video Based Point Cloud Compression", *Int'l Journal on Computer Vision* (IJCV), 2021.
- "Video-based Point Cloud Compression Artifact Removal", *IEEE Trans on Multimedia* (T-MM), 2021.
- "Efficient Projected Frame Padding for Video-based Point Cloud Compression", *IEEE Trans on Multimedia*(T-MM), 2020.
- "Rate Control for Video-based Point Cloud Compression", *IEEE Transactions on Image Processing* (T-IP), 2020.
- "λ-domain Perceptual Rate Control for 360-degree Video Compression", *IEEE Journal of Selected Topics in Signal Processing* (JSTSP), 2020.
- "Advanced 3D Motion Prediction for Video Based Dynamci Point Cloud Compression", *IEEE Trans on Image Processing*(T-IP), 2019.
- "Quadtree-based Coding Framework for High Density Camera Array based Light Field Image", *IEEE Trans on Circuits and Systems for Video Tech*(T-CSVT), 2019.
- "Advanced Spherical Motion Model and Local Padding for 360 Video Compression", *IEEE Trans on Image Processing* (T-IP) vol. 28, no. 5, pp. 2342-2356, May 2019.
- "Pseudo sequence based 2-D hierarchical coding structure for light-field image compression", *IEEE Journal of Selected Topics in Signal Processing* (JSTSP), Special Issue on Light Field, 2017.

What is Point Cloud

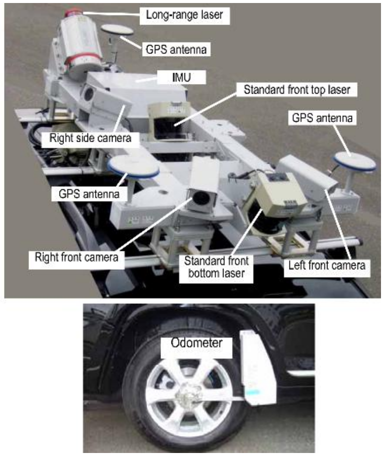
- A collection of Un-ordered points with
  - Geometry: expressed as [x, y, z]
  - Color Attributes: [r g b], or [y u v]
  - Additional info: normal, timestamp, …etc.
- Key difference from mesh: no order or local topology info

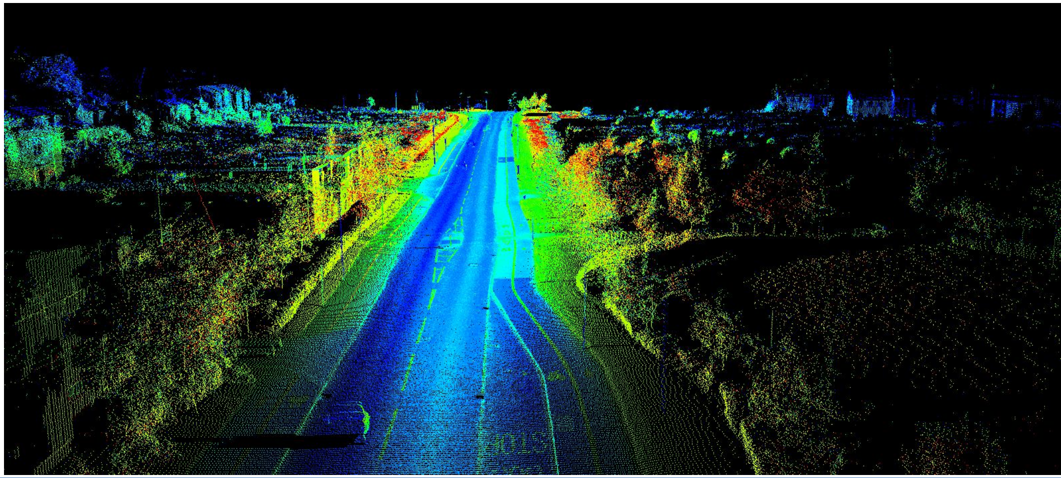
Point Cloud Capture

- Passive: Camera array stereo depth senso



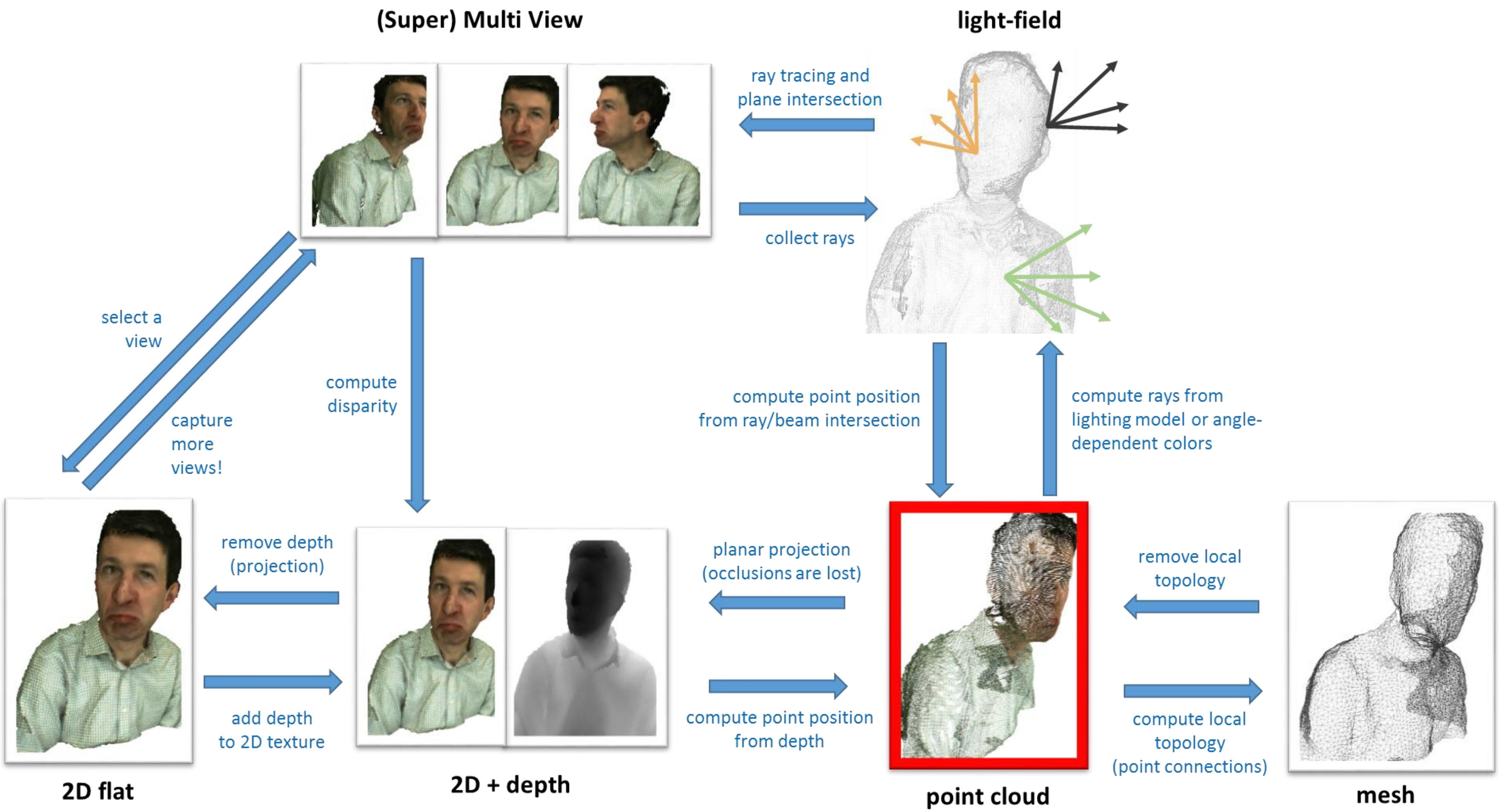
- Active: LiDAR, mmWave, TOF sensors





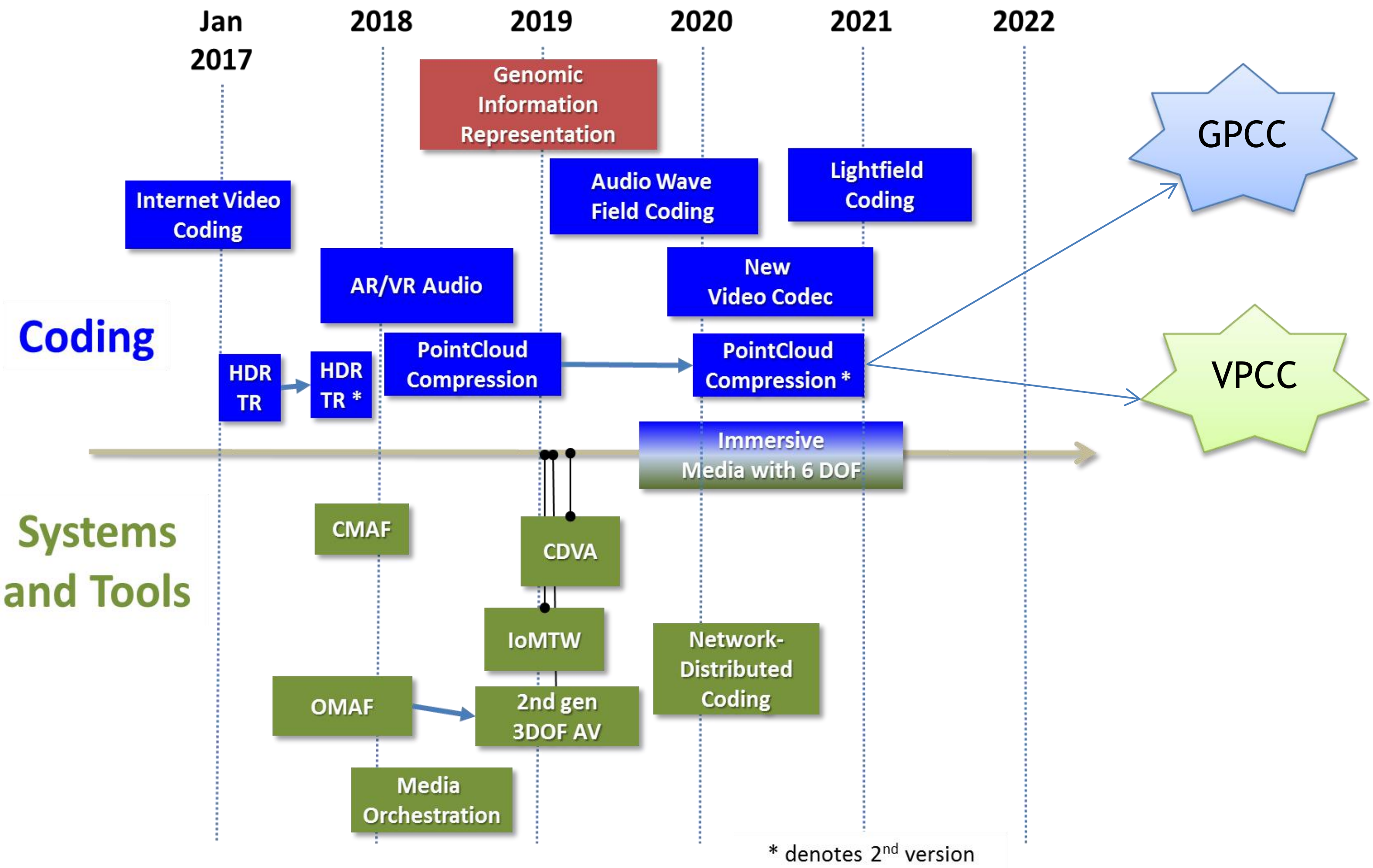
Point Cloud Inter-Operability with Other Formats

- Provide true 6-DoF Content capacity



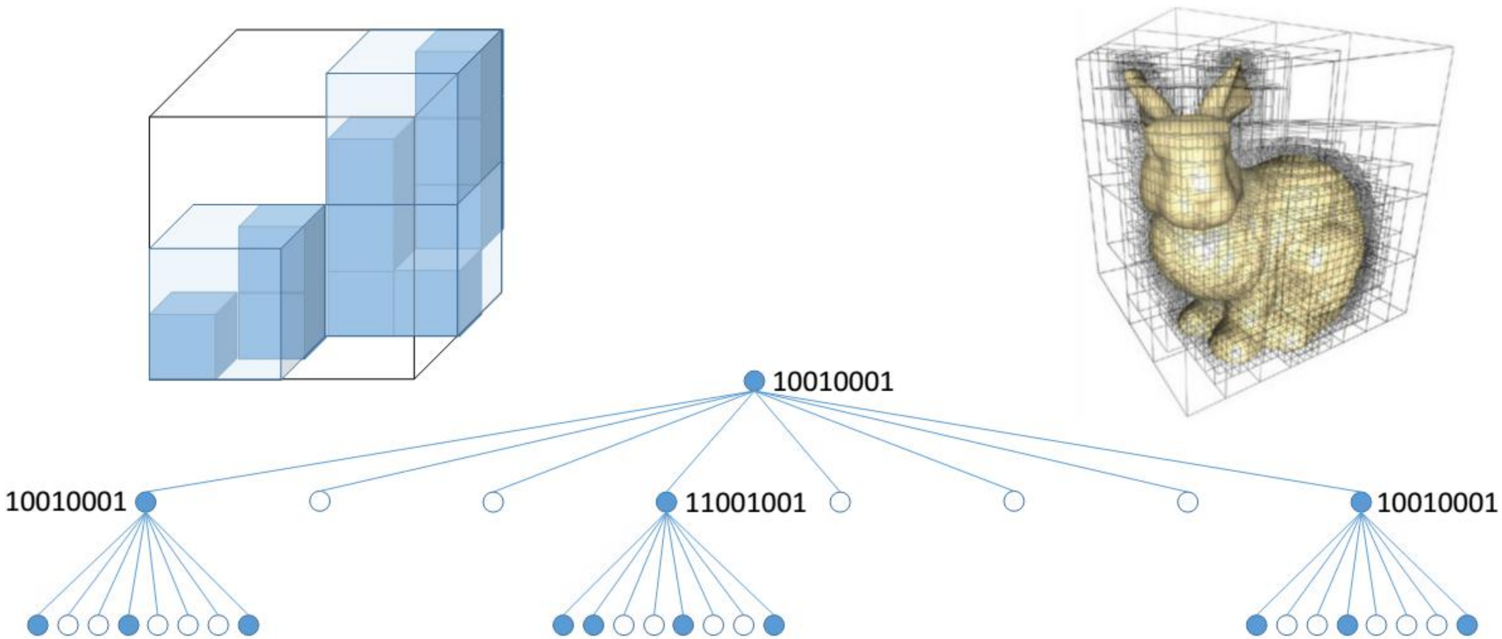
PCC in MPEG

- Part of the MPEG-Immersive grand vision



Octree Based Point Cloud Compression

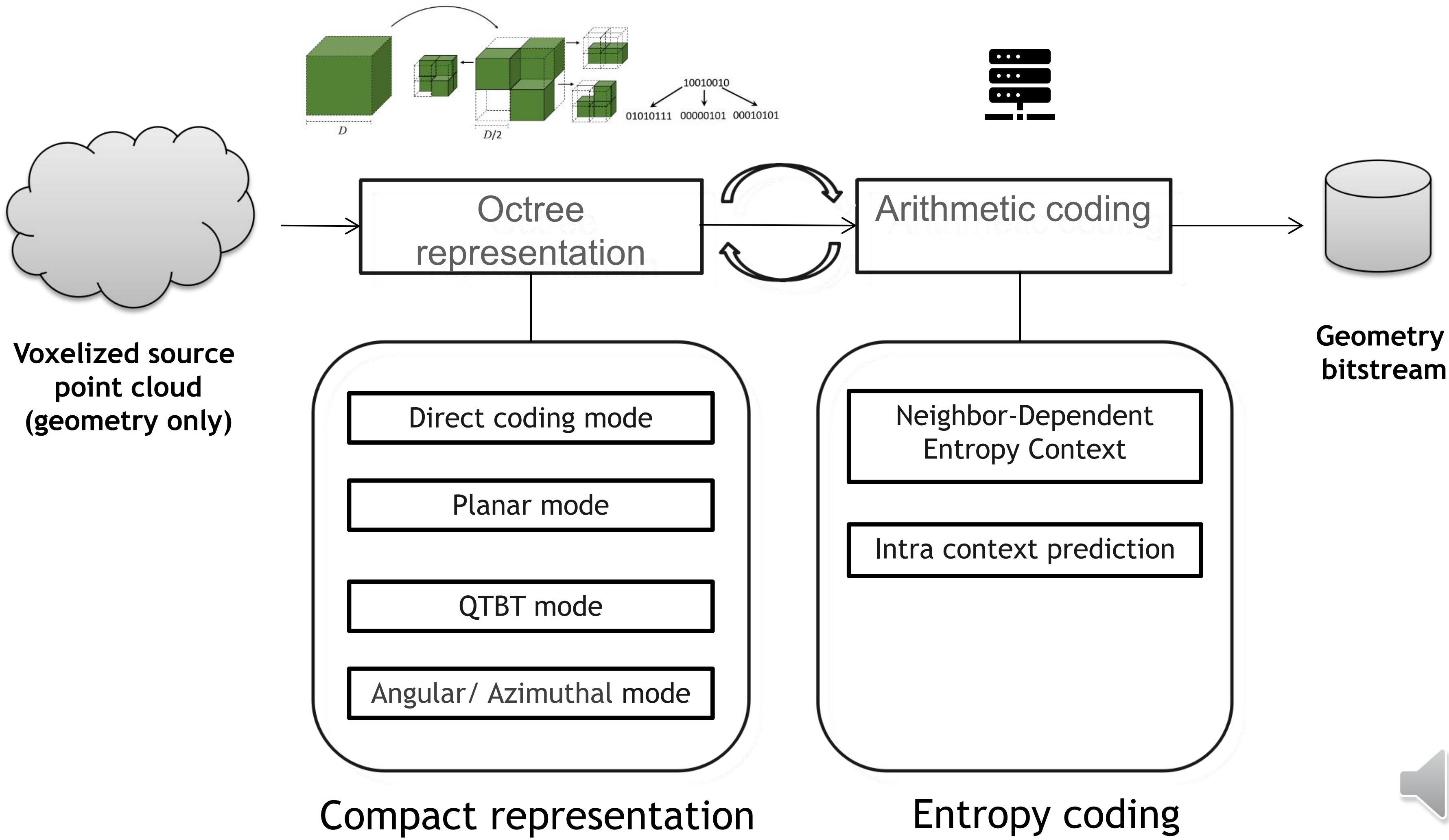
- Octree is a space partition solution
  - Iteratively partition the space into sub-blocks.
  - Encoding: 0 if empty, 1 if contains data points
  - Level of the tree controls the quantization error



Credit: Phil Chou, PacketVideo 2016

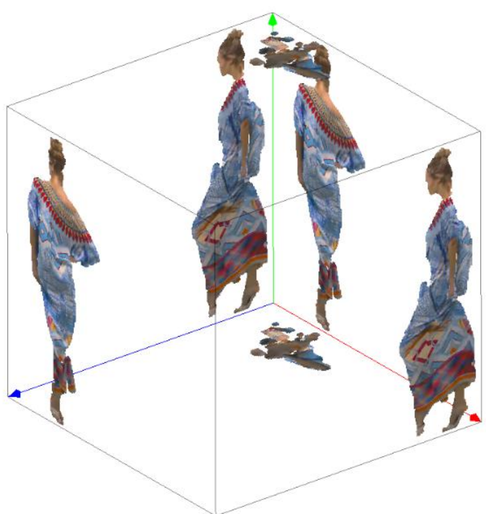
Geometry: Octree coding

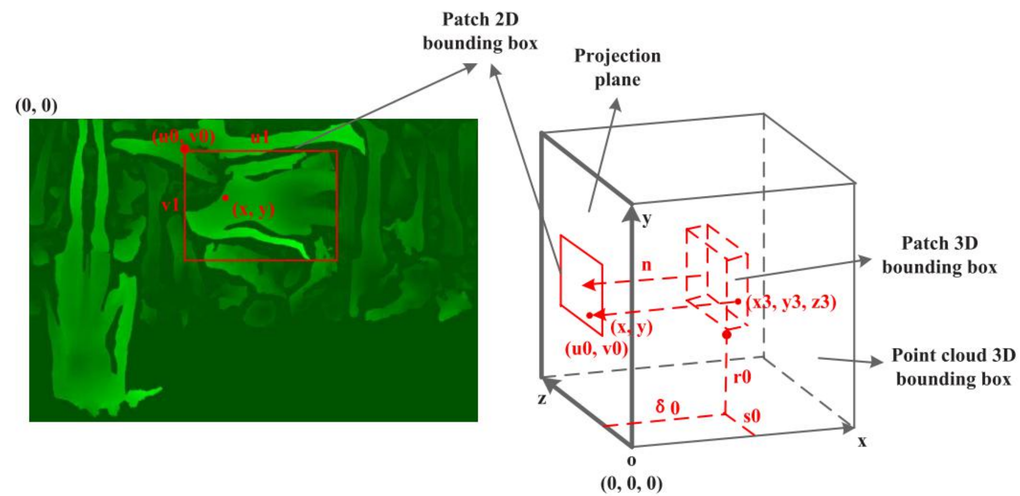
- Octree Context Model is the key for AC efficiency:



Video-based point cloud compression

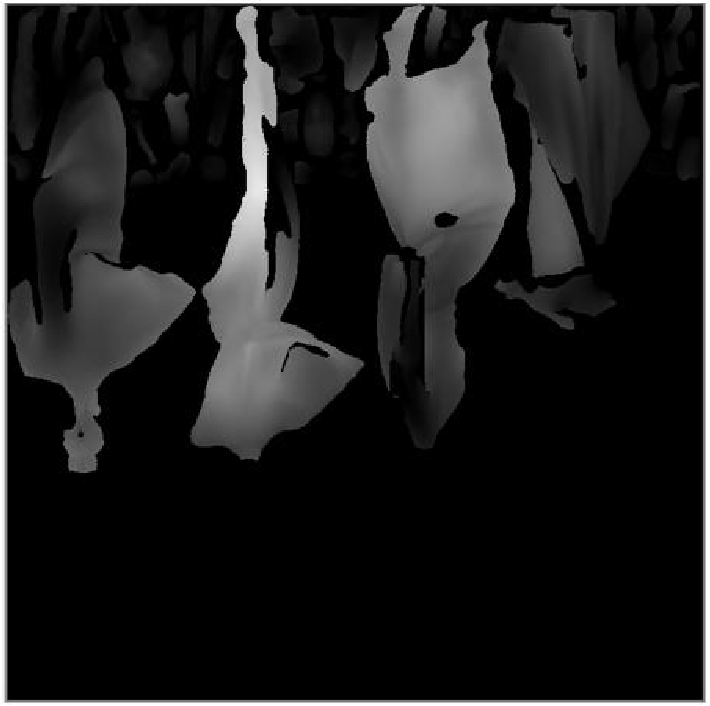
- Basic steps
  - Normal-based projection, frame packing, and frame padding
- Normal-based projection
  - Organize the points with similar normal into a patch
  - Project each patch to the 3D point cloud bounding box





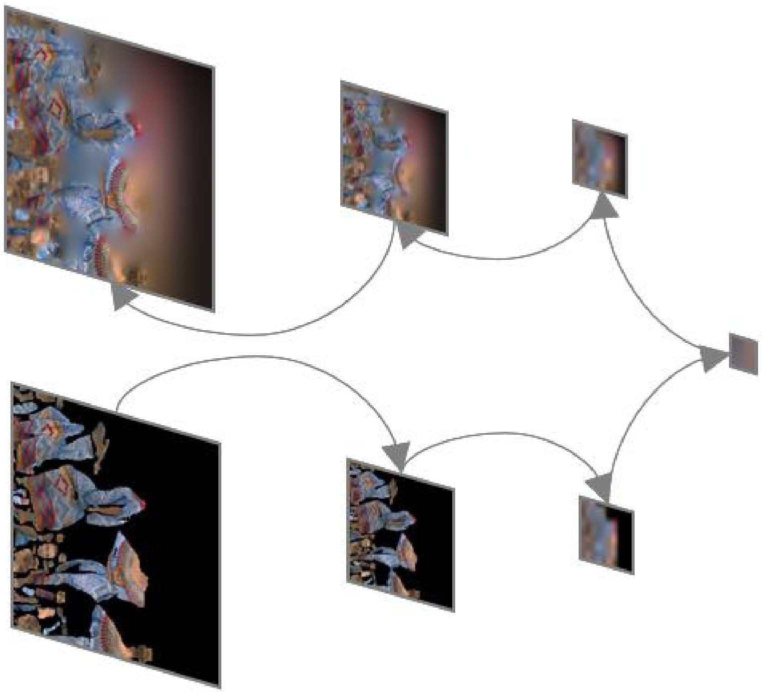
Video-based point cloud compression

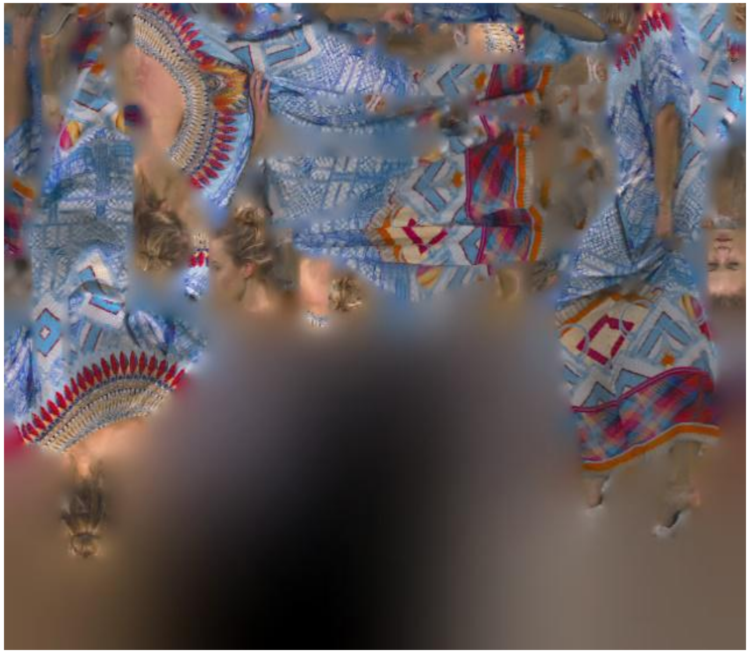
- Frame packing: pack the patches into frames
  - Exhaustive search empty space for the current patch
  - Patch rotation is supported
  - Introduced a lot of sharp edges



VPCC - Texture Padding

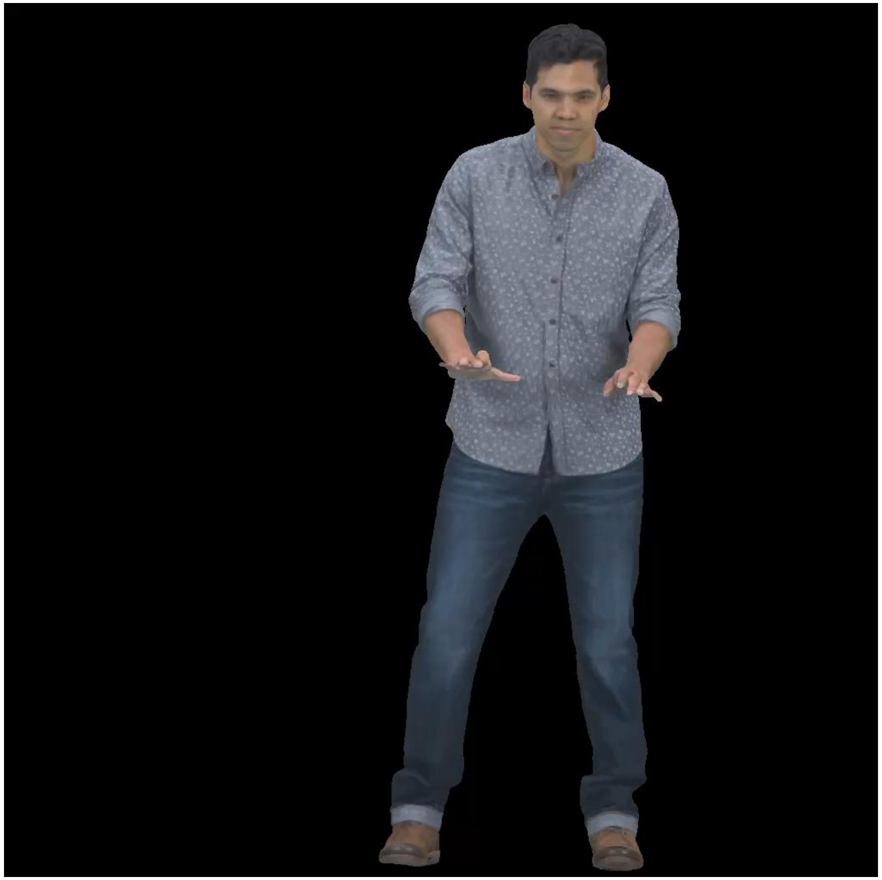
- Texture padding: a number of methods are proposed to minimize the bitrate of the unoccupied pixels
- Using push-pull algorithm as an example, like dilation



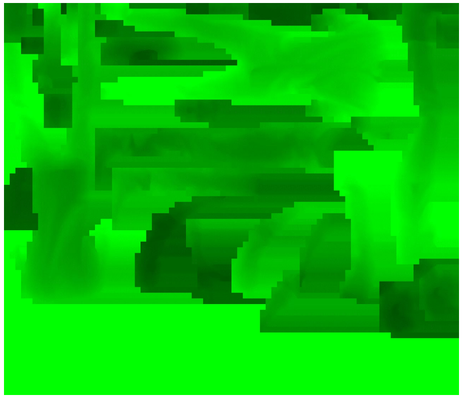


Video-based point cloud compression

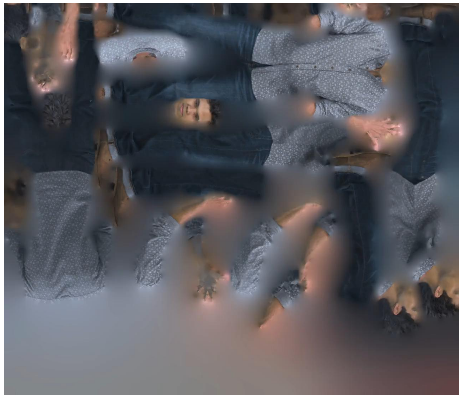
- Basic idea: project a point cloud to a 2-D video for an efficient compression



Geometry



Attribute



Outline

- Short Self Intro
- Research Motivation and Highlights
- Sparse Conv Engine Based PCC
  - PU-Dense: point cloud upscaling work (T-IP)
  - Compression artifacts removal (T-MM)
  - Point Cloud Interpolation
- Video based PCC
  - Advanced 3D motion for VPCC (T-IP)
  - Depth Field Denoising (IJCV)
- Summary

Point Cloud Upsampling

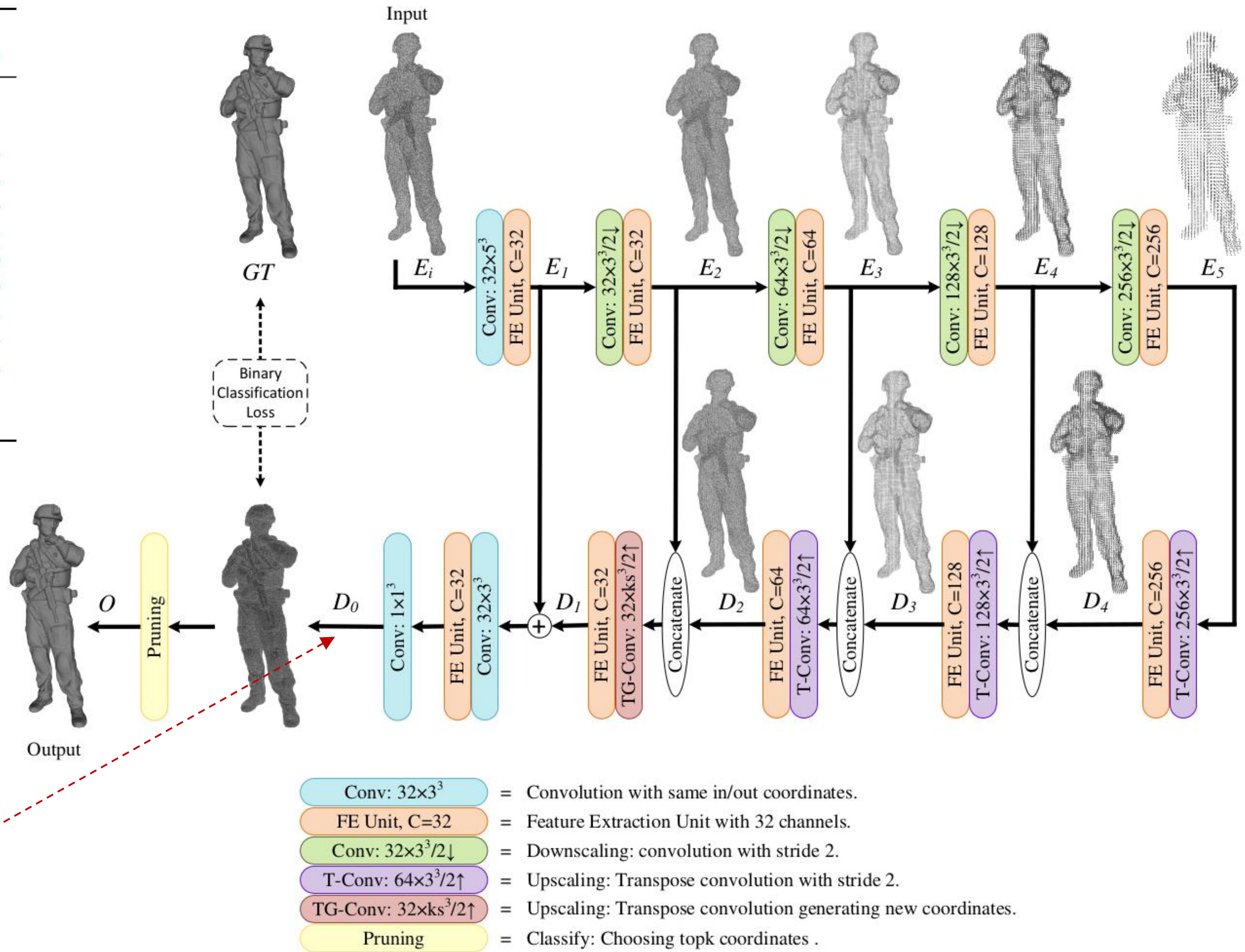
- A very relevant problem:
  - lack of scan line density from LiDAR
  - undersampling of the mesh
  - zoom in for more details
  - as a coding tool for prediction across scale
- Main Challenges:
  - Backbone network limitations (PointNet based): patch based, computationally expensive, cannot support deep and large network due to memory in-efficiency
  - Performance and robustness still lagging: cannot handle large real word data like 8i with > 1M points,
  - Overfitting with PointNet, not generalizeable.

PU-Dense: Point Cloud Upsampling

- Sparse Convolution Back Bone (Minkowski Engine):
  - A fully convolutional geometry upsampling network that is translation invariant and has a variable input size.
  - Novel Feature Embedding (FE) with Inception-Residual Block (IRB) and a 3D Dilated Pyramid Block (3D-DPB)
  - Much larger network with more trainable network weights
- New Loss Function:
  - Employs memory efficient binary voxel classification / cross-entropy loss instead of CD
- Memory efficiency:
  - allows processing of millions of points per inference time.
- Robustness:
  - Can generalize to different datasets. It doesn't just work on synthetic point clouds but can also work for real-world scanned LiDAR based datasets as well as dense photo-realistic point clouds.
  - Robust against noise. Faster inference time.

PU-Dense Architecture

Tensor	Size of coordinates (C)	Size of features (F)
GT	830,397	1
E_i	207,599	1
E_1	207,599	32
E_2	139,244	32
E_3	52,612	64
E_4	14,440	128
E_5	3,623	256
D_4	14,440	256
D_3	52,612	128
D_2	139,244	64
D_1	4,379,676	32
D_0	4,379,676	1
O	830,397	1



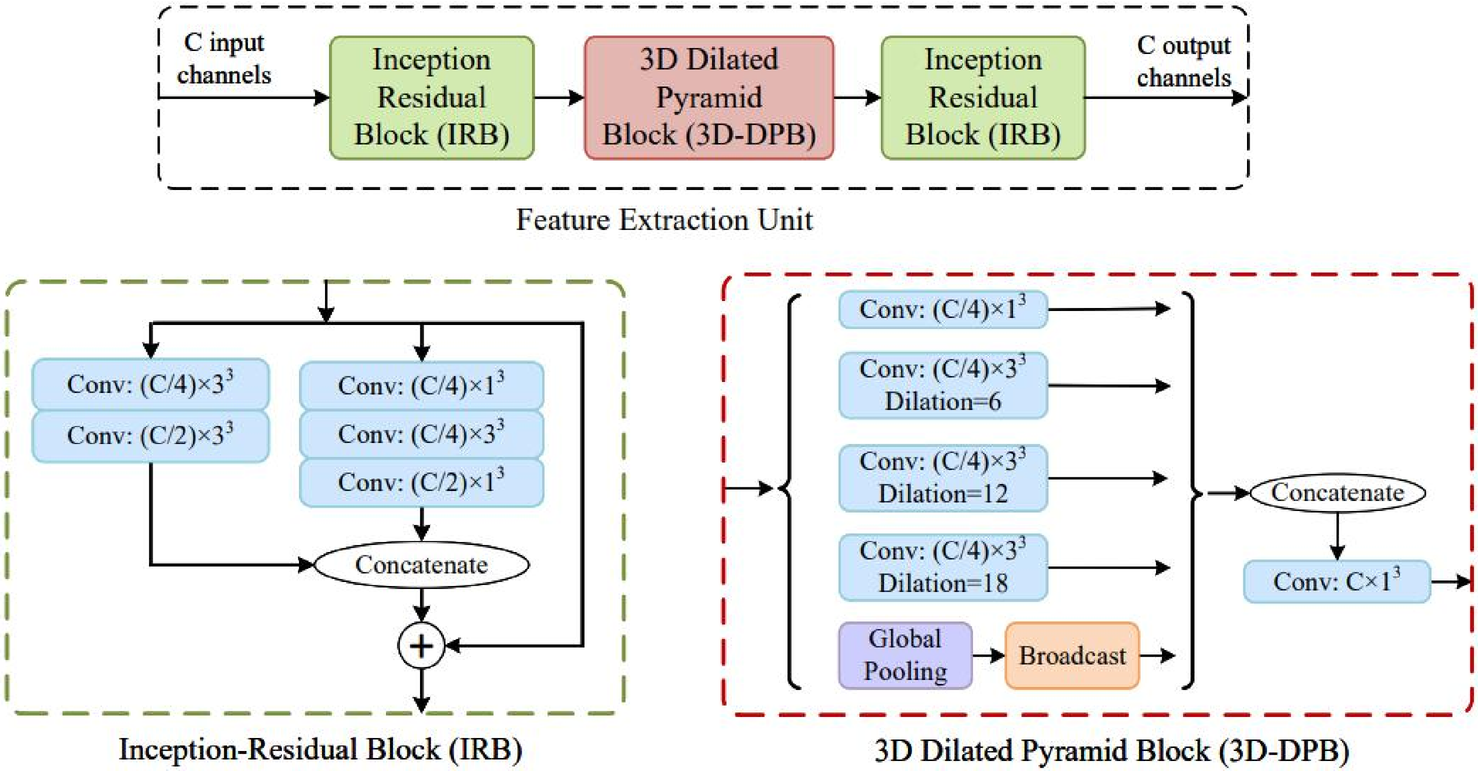
D_0: voxel occupancy prob

PU-Dense Architecture

- Overall a U-Net like structure
  - Voxilized point cloud representation : limited variations.
  - 3 Downscaling with stride
  - increasing feature dimension which encodes occupancy for 2^3, 4^3 ,8^3 sized cubes.
  - Novel Feature Embedding (more details later)
  - Decoding into an occupancy prob function for each voxel location via TG-Conv (Transpose Generative Conv) layer
- Loss function:
  - instead of Chamfer Distance (CD) loss and other similar distance based, we use occupancy prob loss
  - Binary Cross Entropy (BCE) : this is the key, CD usually not working well.

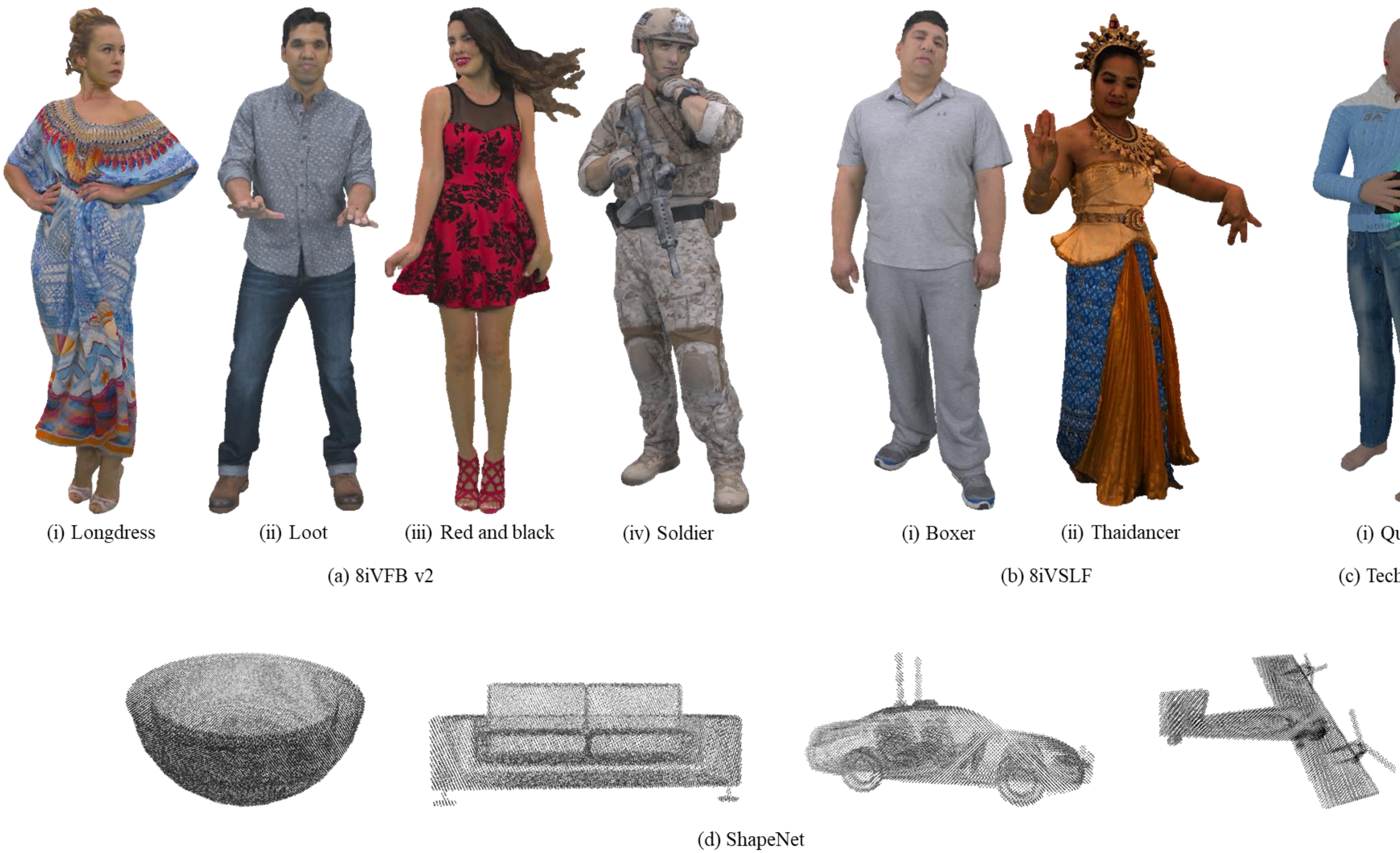
Feature Embedding (FE) Unit

- FE Units.
  - IRB: similar to the inception in image domain, variable kernel size
  - 3D DPB: use dilation in kernel to improve receptive field size



Data Set

- Training from ShapeNet, testing on 8i and Technicolor



Performance

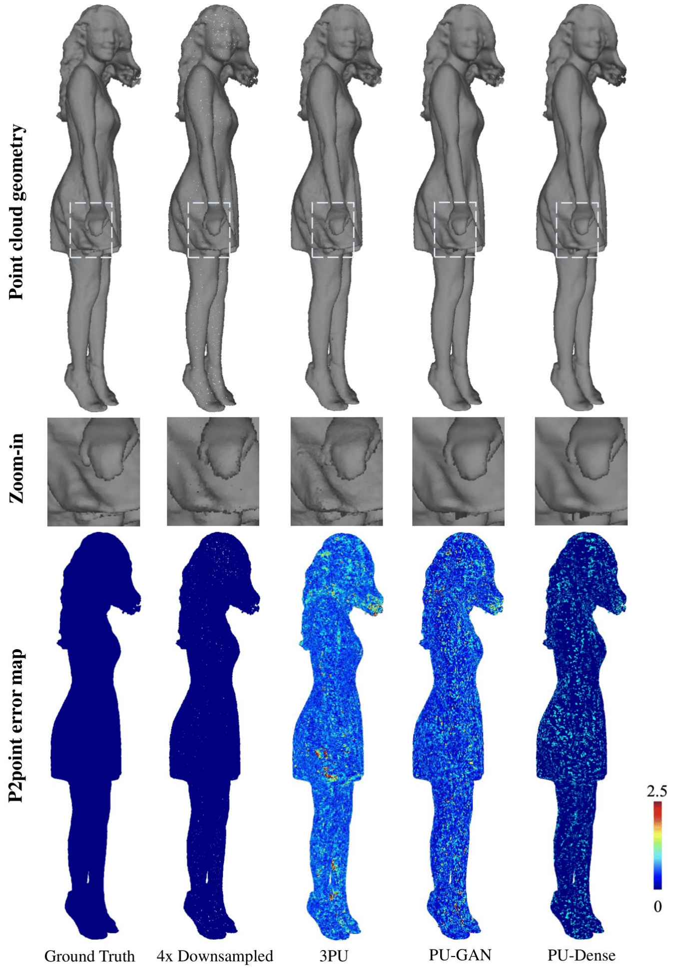
- Point CloudUpsampling
  - very significant performance gains (5~9dB for 4x and 8x) over previous SOTA

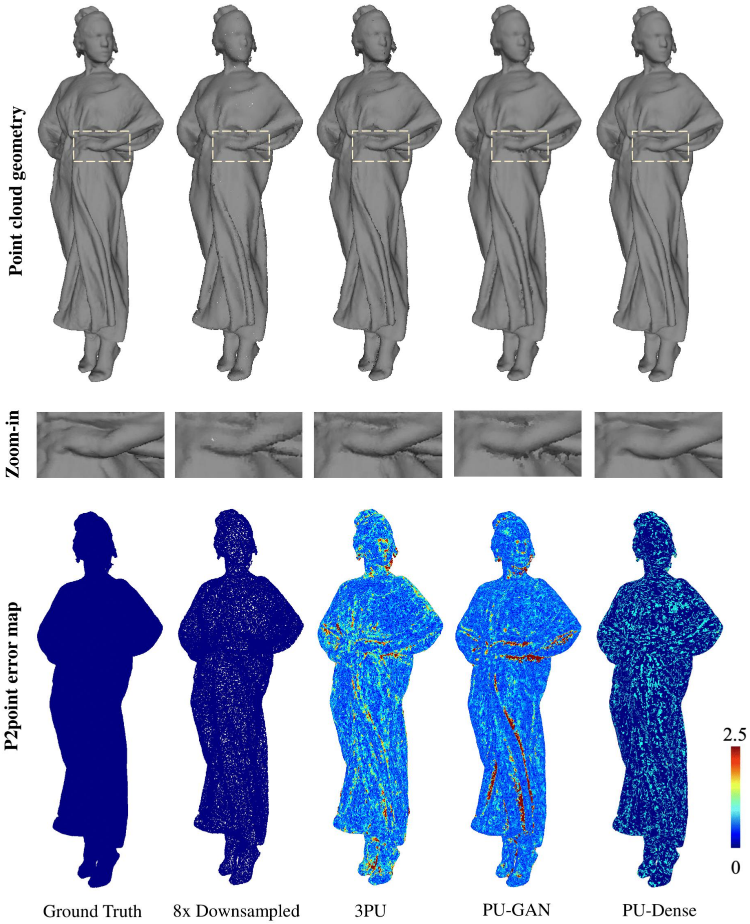
TABLE II
EXTENDED COMPARATIVE RESULTS (CD (10^{-2}) AND PSNR).

Dataset	Upsampling Method	4x		8x	
		CD (10^{-2}) ↓	MSE PSNR (dB) ↑	CD (10^{-2}) ↓	MSE PSNR (dB) ↑
ShapeNet	Downsampled PC	108.18	64.63	199.94	61.96
	3PU	76.36	68.65	149.20	65.37
	PU-GAN	49.41	70.64	174.58	64.88
	PU-Dense	**18.82**	**75.24**	**30.52**	**73.11**
8iVFB	Downsampled PC	114.63	64.38	222.91	61.49
	3PU	67.04	69.41	105.43	66.83
	PU-GAN	45.60	70.92	117.66	66.19
	PU-Dense	**19.38**	**75.05**	**33.18**	**72.57**
8iVSLF	Downsampled PC	286.67	73.17	600.34	70.00
	3PU	204.92	76.98	368.63	74.78
	PU-GAN	156.94	77.18	231.39	75.34
	PU-Dense	**135.41**	**78.92**	**202.82**	**76.79**
Queen	Downsampled PC	106.69	64.69	196.46	62.04
	3PU	57.13	70.19	90.90	67.55
	PU-GAN	41.67	71.43	110.42	66.36
	PU-Dense	**15.76**	**75.93**	**25.45**	**73.76**

Subjective Results

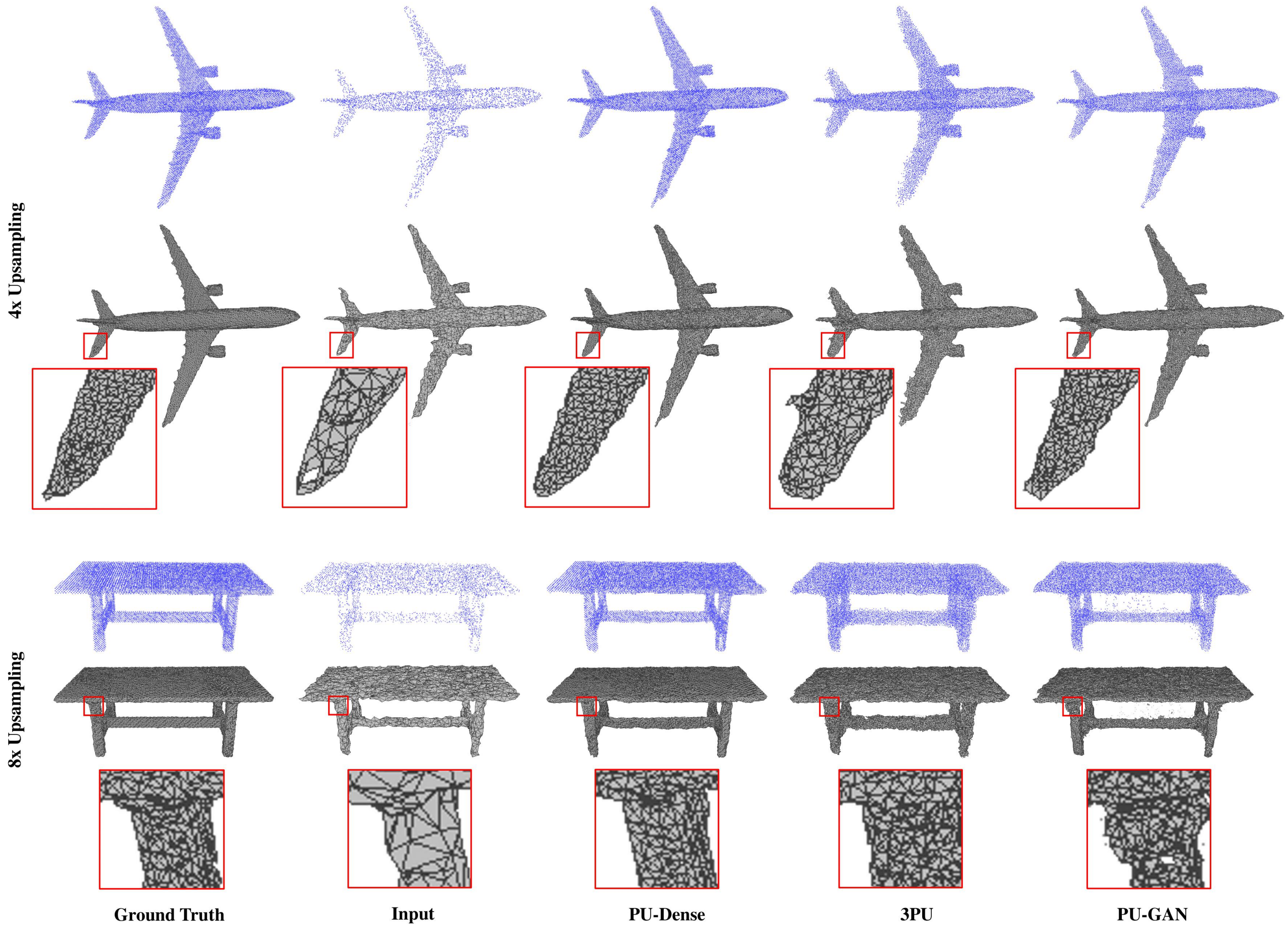
- 8i Sequences 4X upsacling





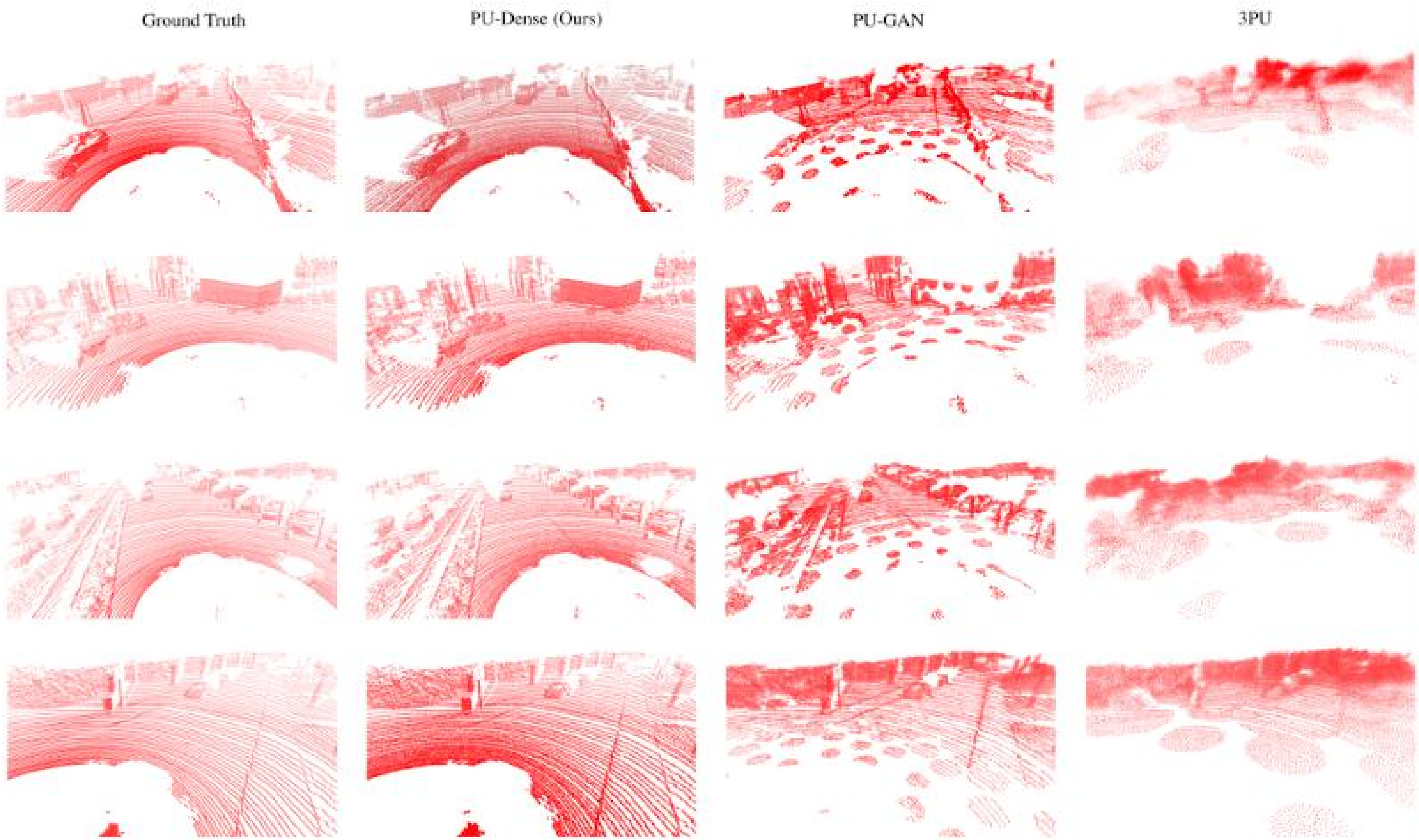
Mesh Quality- Synthetic Objects

- Mesh quality from re-genration: 4x and 8x upscaled



LiDAR data results

- KITTI: 4x upscaling



Complexity

- Much more network parameters
- But faster inferences

TABLE IV
QUANTITATIVE COMPARISON: AVERAGE EVALUATION TIME PER POINT CLOUD FOR 4X UPSAMPLING.

Dataset	Upsampling Method	Computation time (min)
8iVFB	3PU	27.49
	PU-GAN	24.78
	PU-Dense	**00.79**

TABLE V
QUANTITATIVE COMPARISON: NUMBER OF TRAINABLE PARAMETERS.

Upsampling Method	Trainable parameters
3PU	152,054
PU-GAN	541,601
PU-Dense	**13,172,441**

PU-Dense Summary

- Loss function: Point Cloud Upsampling is about voxel occupancy prediction, switching from distance based loss to occupancy prob loss is the main break through
- Network backbone: PointNet and variations are limited in efficiency and performance, sparse conv network backbones like Minkowski Engine allows for much larger data set and deeper network, lead to significantly better performance
- New SOTA: This sparse conv backbone + occupancy prob loss framework gives us new performance in a variety of problems, including, upsampling, denoising (next topic), and inter-prediction coding*.

Outline

- Short Self Intro
- Research Motivation and Highlights
- Sparse Conv Engine Based PCC
  - PU-Dense: point cloud upscaling work (T-IP)
  - Compression artifacts removal (T-MM)
  - Point Cloud Interpolation
- Video based PCC
  - Advanced 3D motion for VPCC (T-IP)
  - VPCC Depth Field Denoising (IJCV)
- Summary

VPCC Motion Model

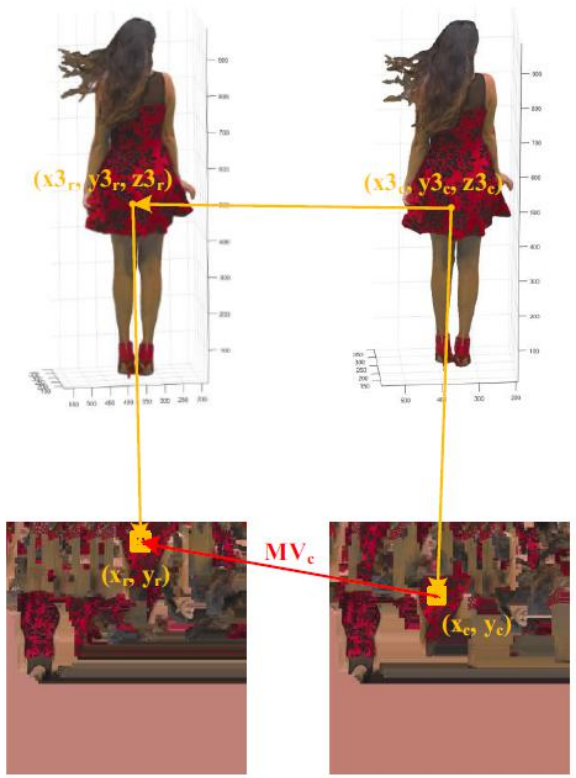
- The corresponding patches may be put in totally different positions in various frames (Green squares)
  - The current video codec may be unable to find a good motion vector for each block in this case
  - The geometry is encoded before the attribute, we can use the geometry to derive a better motion vector for attribute

General 3D to 2D motion model

- Given the 3D motion and the 3D to 2D correspondence, we can derive the 2D motion
  - g(), f(): 3D to 2D projection in reference and current frames

$$\boldsymbol{MV_c} = \boldsymbol{g}(x3_r, y3_r, z3_r) - \boldsymbol{f}(x3_c, y3_c, z3_c)$$

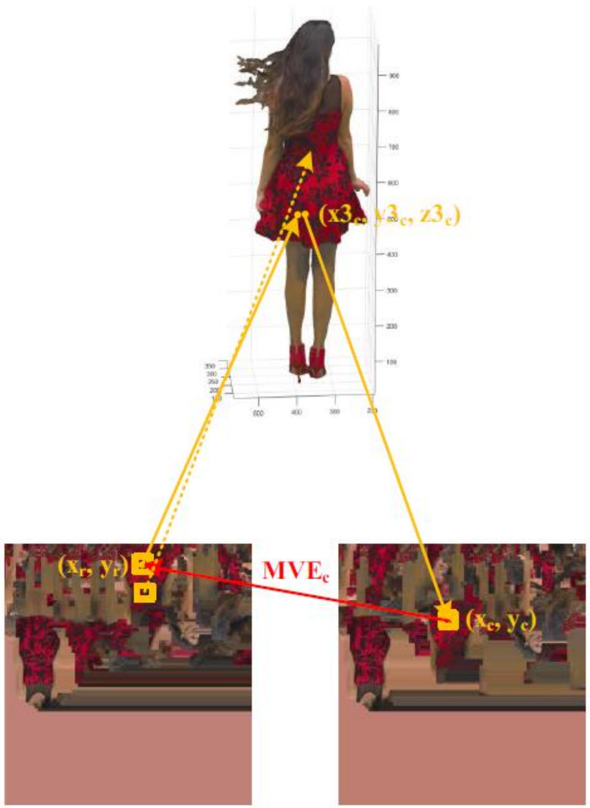


3D Geometry-based motion prediction

- In the V-PCC, we know the 3D-to-2D correspondence but do not know the 3D motion
- We assume the current frame and the reference frame will not change dramatically

$$\boldsymbol{MVE_c} = \boldsymbol{g}(x3_c, y3_c, z3_c) - \boldsymbol{f}(x3_c, y3_c, z3_c)$$

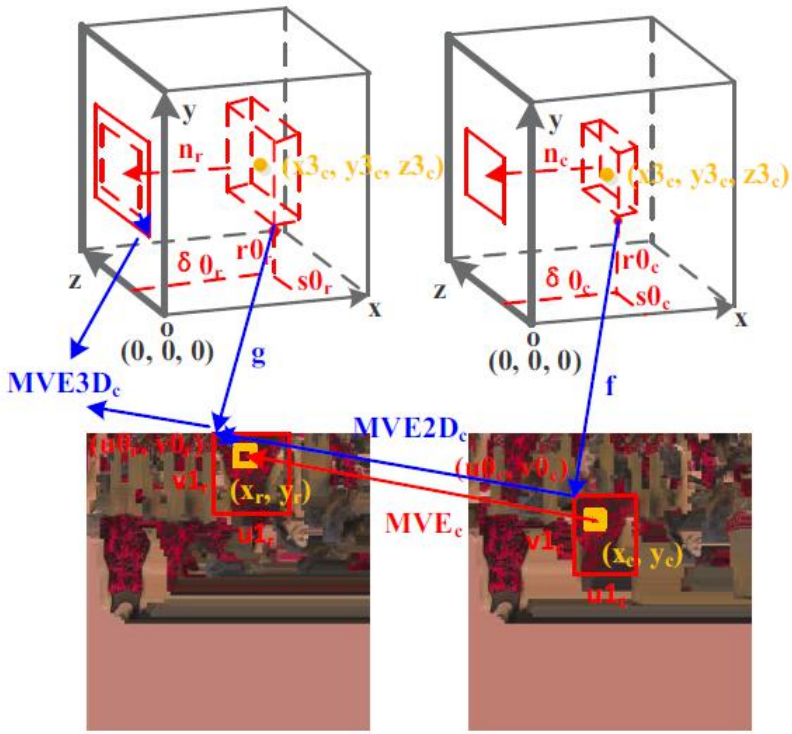
- The problem is that (x3c,y3c,z3c) may not have a corresponding 2D point in the reference frame
  - We perform motion estimation which will increase the encoder and decoder complexity



Auxiliary information based motion prediction

- The previous method has the following two disadvantages
  - The high encoder and decoder complexity
  - It can only apply to the attribute

- The auxiliary information based motion prediction
  - The auxiliary information basically provides the coarse geometry
  - We use the 3D offset plus the 2D offset



Experiments setup

- The proposed algorithm is implemented in the V-PCC reference software and the corresponding HEVC reference software
- We test the all the dynamic point clouds defined in the common test condition including loot, redandblack, soldier, queen, longdress
- For the geometry, both point-to-point is point-to-plane are used
- For the attribute, the qualities of the luma, Cb, and Cr are considered

Experimental results on the overall scheme

- Overall scheme results: 3D motion vs SEI messaging

TABLE III
PERFORMANCE OF THE GEOMETRY-BASED MOTION PREDICTION COMPARED WITH THE V-PCC ANCHOR

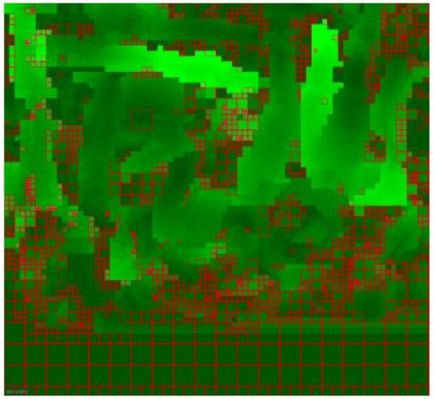
Test	Geom.BD-GeomRate		Attr.BD-AttrRate			Geom.BD-TotalRate		Attr.BD-TotalRate		
point cloud	D1	D2	Luma	Cb	Cr	D1	D2	Luma	Cb	Cr
Loot	0.0%	0.0%	−18.1%	−31.4%	−30.4%	−3.4%	−6.1%	−8.4%	−17.7%	−16.9%
RedAndBlack	0.0%	0.0%	−16.3%	−25.0%	−15.9%	−4.6%	−4.6%	−8.8%	−15.4%	−8.4%
Solider	0.0%	0.0%	−33.4%	−42.5%	−43.2%	−8.2%	−8.2%	−17.2%	−26.3%	−27.0%
Queen	0.0%	0.0%	−13.7%	−20.5%	−19.2%	−3.5%	−3.6%	−7.8%	−12.7%	−11.6%
LongDress	0.0%	0.0%	−9.8%	−13.5%	−12.3%	−3.7%	−3.7%	−6.4%	−9.5%	−8.4%
Avg.	0.0%	0.0%	−18.2%	−26.6%	−24.2%	−4.7%	−4.7%	−9.7%	−16.3%	−14.5%
Enc. time self	97%									
Dec. time self	98%									
Enc. time child	486%									
Dec. time child	337%									

TABLE IV
PERFORMANCE OF THE AUXILIARY-INFORMATION-BASED MOTION PREDICTION COMPARED WITH THE V-PCC ANCHOR UNDER THE NORMATIVE SOLUTION

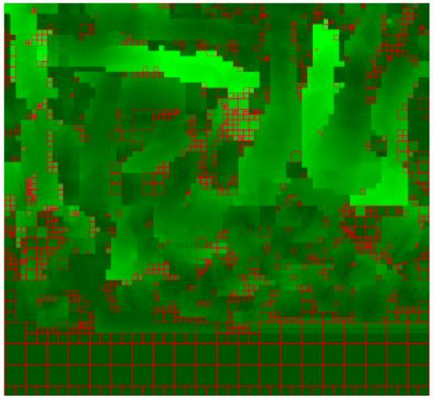
Test	Geom.BD-GeomRate		Attr.BD-AttrRate			Geom.BD-TotalRate		Attr.BD-TotalRate		
point cloud	D1	D2	Luma	Cb	Cr	D1	D2	Luma	Cb	Cr
Loot	−4.0%	−3.9%	−16.3%	−26.4%	−28.5%	−6.3%	−6.2%	−9.6%	−16.7%	−17.9%
RedAndBlack	−1.0%	−1.1%	−12.2%	−18.9%	−10.9%	−4.0%	−4.1%	−7.2%	−12.1%	−6.2%
Solider	−8.0%	−7.9%	−31.3%	−41.4%	−40.4%	−13.6%	−13.4%	−19.8%	−28.7%	−28.1%
Queen	−5.9%	−5.9%	−11.8%	−17.0%	−15.7%	−7.3%	−7.3%	−9.1%	−12.9%	−11.8%
LongDress	−1.1%	−1.1%	−8.3%	−11.2%	−10.2%	−3.8%	−3.6%	−5.7%	−8.2%	−7.3%
Avg.	−4.0%	−4.0%	−16.0%	−23.0%	−21.1%	−7.0%	−6.9%	−10.3%	−15.7%	−14.3%
Enc. time self	100%									
Dec. time self	100%									
Enc. time child	98%									
Dec. time child	99%									

Performance Analysis

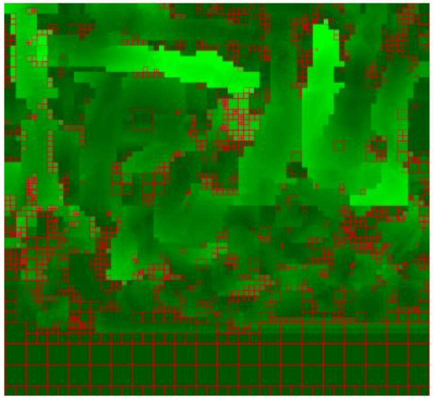
- Intra blocks (in red) reduce significantly, resulting in taking adv of inter coding efficiency



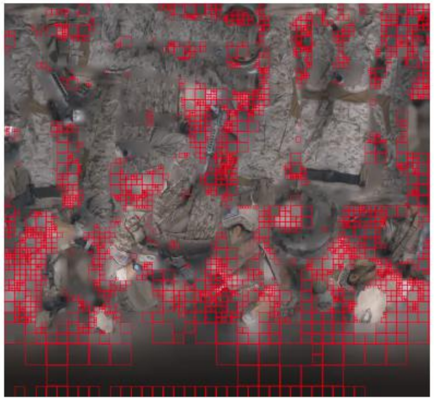
(a) Soldier Geometry Anchor



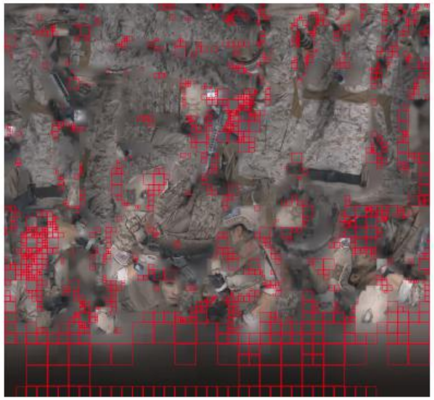
(b) Soldier Geometry Normative



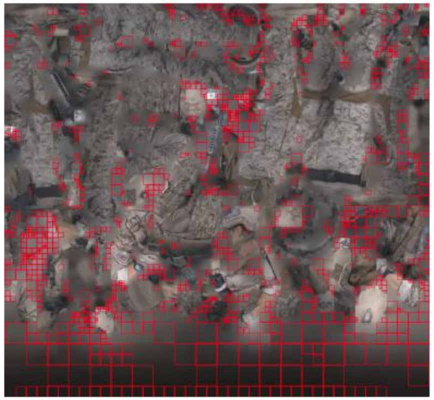
(c) Soldier Geometry Non-normative



(d) Soldier Geometry Anchor

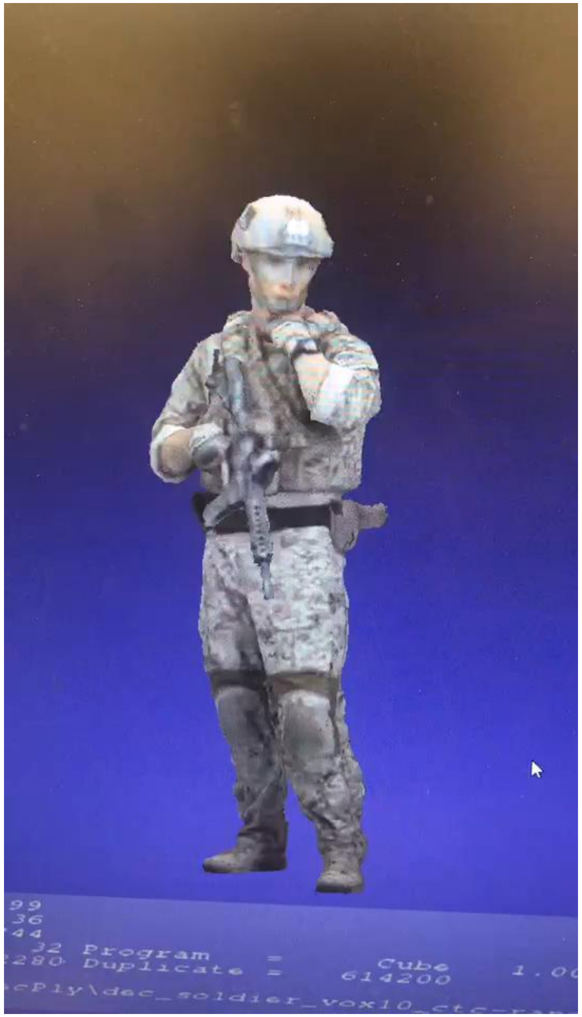


(e) Soldier Geometry Normative

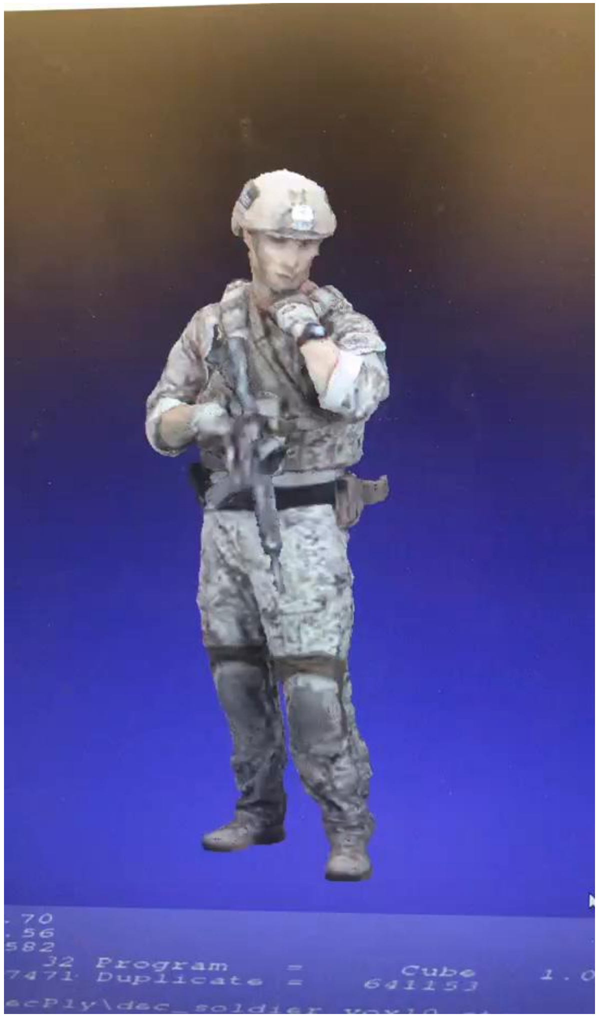


(f) Soldier Geometry Non-normative

Subjective quality



Anchor



Proposed

Adv 3D Motion for VPCC Summary

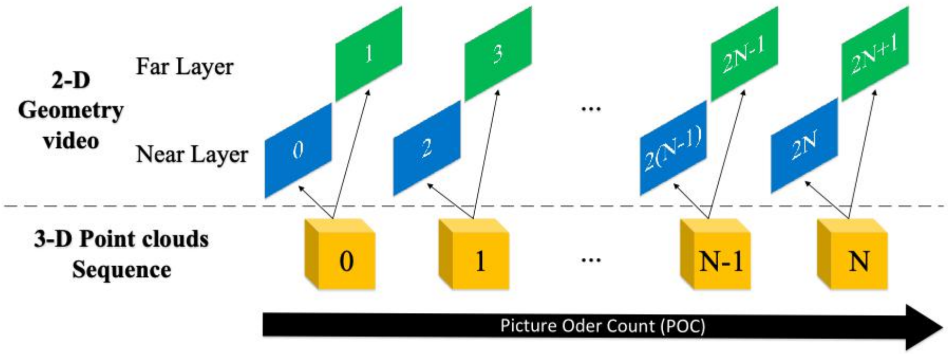
- Pain point: motion coherence is destroyed in the VPCC projection process, leads to poor motion compensation performance
- Key contribution: recover motion coherence in 3D domain, and generate a predictor for 2D motion estimation and compensation in HEVC codec.
- Significance: adopted in the VPCC test model.

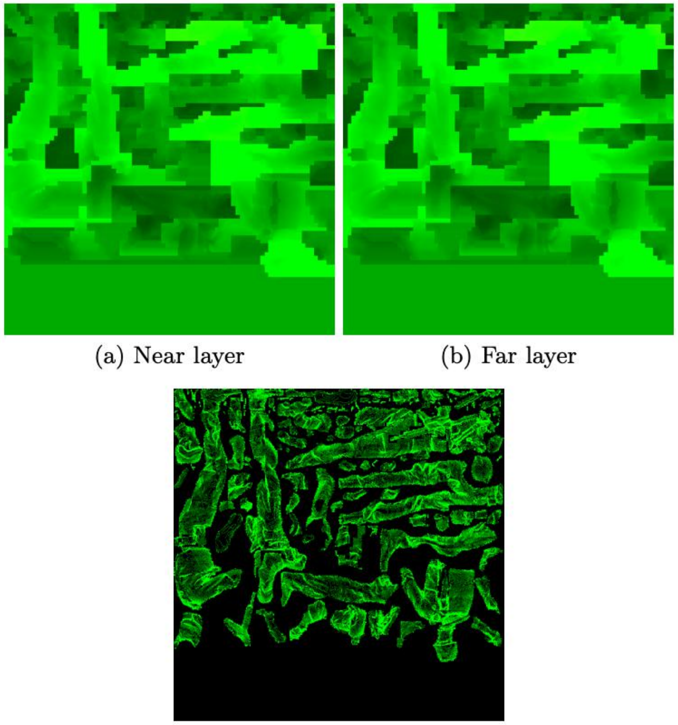
Outline

- Short Self Intro
- Research Motivation and Highlights
- Sparse Conv Engine Based PCC
  - PU-Dense: point cloud upscaling work (T-IP)
  - Compression artifacts removal (T-MM)
  - Point Cloud Interpolation
- Video based PCC
  - Advanced 3D motion for VPCC (T-IP)
  - VPCC Depth Field Denoising (IJCV)
- Summary

The Geometry in VPCC

- Geometry is projected to 2 layers of depth in VPCC
  - mainly used to handle occulation
  - coding wise is treated as interleaved sequences





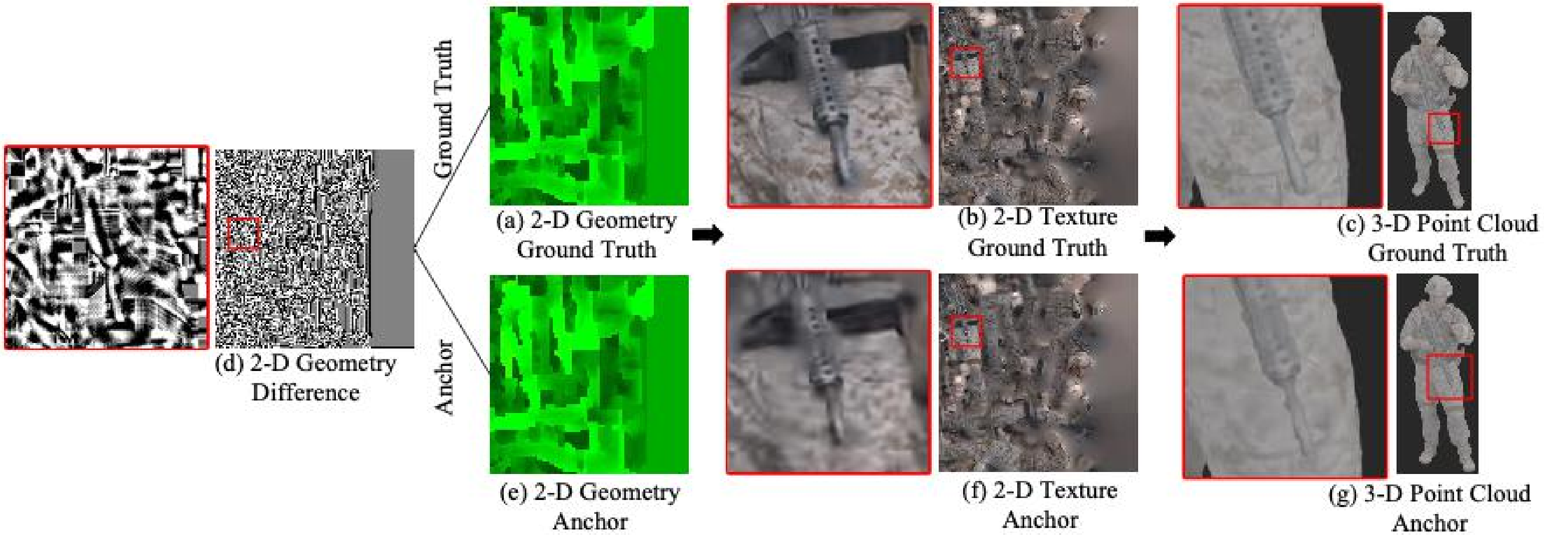
(a) Near layer (b) Far layer

(c) The difference of near and far layer

Near (a) and far (b) layer frames in 2-D geometry video

Geometry Recovery from Depth

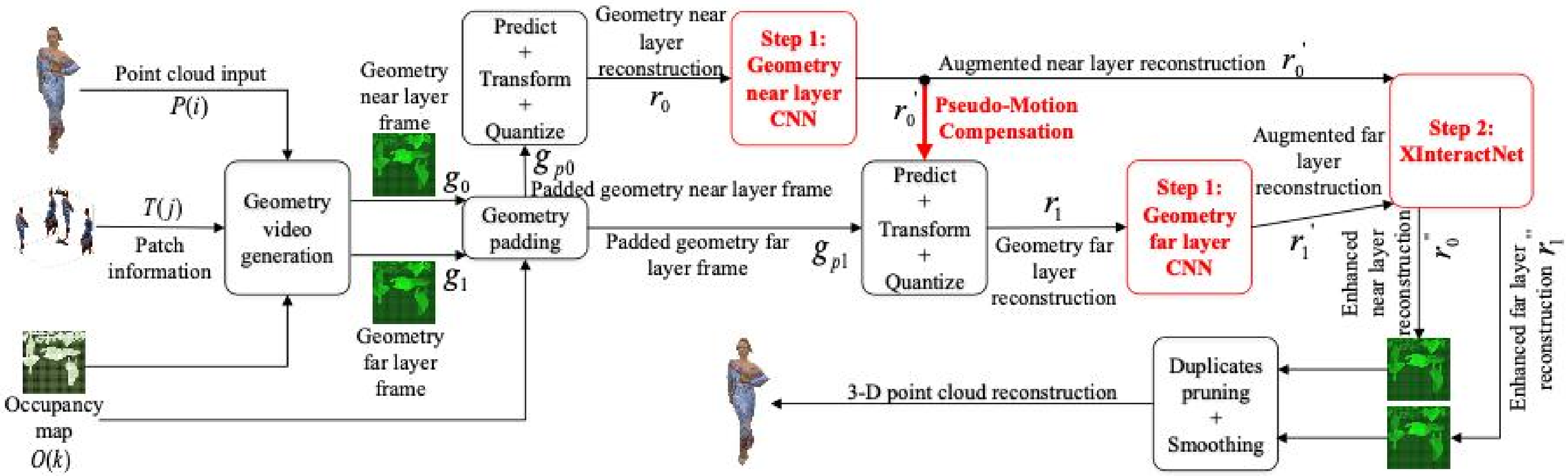
- VPCC use occupancy map to signal patches, error comes from:
  - quantization loss in depth (near and far field)
  - as well as the occupancy resolution loss in [Jia21TMM]* (4x4 by default)



*[Jia21TMM] W. Jia, L. Li, A. Akhtar, Z. Li and S. Liu, "Convolutional Neural Network-based Occupancy Map Accuracy Improvement for Video-based Point Cloud Compression," in IEEE Transactions on Multimedia, doi: 10.1109/TMM.2021.3079698.

Joint Near and Far Depth Field Denoising

- A dual path network taking in near and far depth field
- Training strategy:
  - near field denoising network is trained first
  - far field benefit from denoised near field input in a pseudo-motion compensation scheme
  - some regularization in loss of near and far depth field interaction



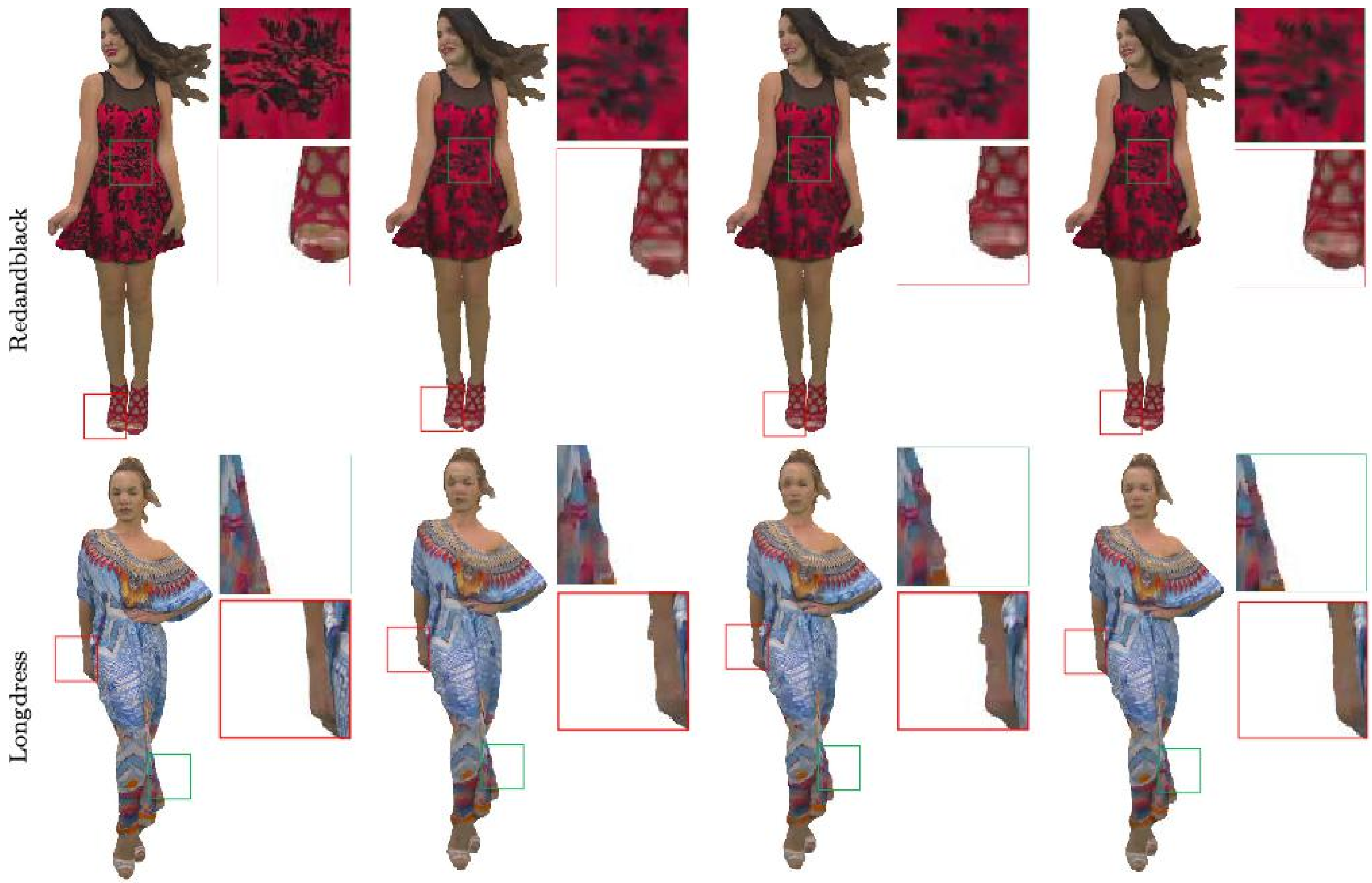
Results

- BD rate for all intra case:
  - significant coding gains :)
  - while very limited decoding complexity cost

Table 3 Proposed two-step method and geometry padding method [18] against the geometry smoothing method [32] respectively on BD-rate and time complexity within the first 32 frames of sequences under all intra

Class	Sequence	V-PCC with geometry padding method (SOTA) [18]					Proposed two-step method				
		Geom.BD-Totalrate		Attr.BD-Totalrate			Geom.BD-Totalrate		Attr.BD-Totalrate		
		D1 ↓	D2 ↓	Luma ↓	Cb ↓	Cr ↓	D1 ↓	D2 ↓	Luma ↓	Cb ↓	Cr ↓
A	Loot	−4.4%	−9.9%	4.4%	4.8%	5.4%	**-20.8%**	**-18.2%**	−1.2%	−1.4%	−0.5%
	Redandblack	−0.8%	−7.5%	4.4%	5.9%	5.0%	**-10.1%**	**-11.3%**	−1.1%	−0.9%	−2.4%
	Soldier	−1.5%	−7.8%	4.4%	7.3%	6.4%	**-13.8%**	**-12.8%**	−1.9%	0.1%	−0.8%
B	Longdress	−1.3%	−8.2%	2.3%	3.5%	3.2%	**-12.8%**	**-13.7%**	−2.9%	−1.4%	−2.3%
	Class a	−2.2%	−8.4%	4.4%	6.0%	5.6%	**-14.9%**	**-14.1%**	−1.4%	−0.7%	−1.2%
	class b	−1.3%	−8.2%	2.3%	3.5%	3.2%	**-12.8%**	**-13.7%**	−2.9%	−1.4%	−2.3%
Avg.	All	−2.0%	−8.3%	3.9%	5.4%	5.0%	**-14.4%**	**-14.0%**	−1.8%	−0.9%	−1.5%
	Enc.Self			102%					103%		
	Enc.Child			101%					101%		
	Dec.Self			102%					121%		
	Dec.Child			101%					212%		

Subjective Results



(a) Ground truth (b) Geometry smoothing method (c) Geometry padding method (d) Proposed two-step method

Summary

- Deep Learning is a powerful tool in signal recovery, interpolation and prediction.
- Many applications in immersive media (point cloud, light field, and 360) compression, sampled a small part of our work in these areas.
- Looking forward to new revolution in compression efficiency with deep learning framework.
- Q&A

Thanks