Enhancing Dynamic Point Clouds in the Wild: A Grand Challenge on Real-World 4D Volumetric Data

Abstract

Dynamic point clouds captured in real-world environments are fundamental to immersive multimedia applications such as volumetric video, XR telepresence, and digital twins. However, point clouds acquired by consumer-grade sensors suffer from severe degradations, including noise, sparsity, missing geometry, temporal instability, and color artifacts, which significantly limit downstream reconstruction, rendering, and compression. Existing point cloud enhancement methods are predominantly evaluated on synthetic benchmarks and static scenes, leaving a critical gap in systematic evaluation for real-world, dynamic (4D), and color point clouds. This Grand Challenge introduces the first in-the-wild benchmark for dynamic point cloud enhancement based on the UVG-CWI-DQPC dataset, which provides paired low-quality consumer-grade captures and high-fidelity multi-sensor ground truth across diverse dynamic human-centric sequences. The challenge targets unified enhancement of denoising, completion, and upsampling, while explicitly accounting for temporal consistency and color fidelity. Participants are evaluated using a comprehensive protocol combining geometric accuracy, texture fidelity, perceptual quality, temporal stability, and computational efficiency, with the top submissions further assessed via controlled subjective studies. This challenge aims to foster realistic algorithm design, fair comparison, and accelerated progress toward practical deployment of dynamic point cloud enhancement in multimedia systems.

ACM Reference Format:

. 2018. Enhancing Dynamic Point Clouds in the Wild: A Grand Challenge on Real-World 4D Volumetric Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Point clouds have emerged as a fundamental data representation in numerous three-dimensional (3D) vision tasks, ranging from object recognition and scene understanding to immersive media and digital twins. However, raw point clouds acquired by commodity sensors often suffer from sparsity, noise, and incompleteness due to the limitations of acquisition hardware and inevitable sensing artifacts. These degradations significantly hinder downstream processing, including reconstruction, rendering, analysis, and compression.

The objective of 3D point cloud enhancement is to resample and refine input point sets to produce higher-quality data that are clean, complete, and dense. Enhancement methods may span a wide spectrum, including:

- **Deep-learning-based approaches**, which learn data-driven priors for denoising, completion, or upsampling [7];
- **Optimization and interpolation methods**, which explicitly exploit geometric consistency [10];
- **Multi-modal approaches**, which leverage 2D images or videos to provide complementary structural or textural priors [14];
- **AI-generated 3D content approaches**, where enhancement is viewed as generative point cloud synthesis.

Overall, point cloud enhancement encompasses three main subtasks: (1) *denoising* to remove noise and outliers; (2) *completion* to recover missing or occluded structures; and (3) *upsampling* to generate dense and uniform point distributions. It is typically performed by independently or sequentially conducting point cloud denoising, completion, and upsampling. Despite the rapid progress of methods across these dimensions, systematic evaluation and benchmarking for real-world point cloud enhancement remain insufficient, especially from the perceptual quality of the point cloud for the immersive media perspective.

2 Significance of the Challenge

Current evaluation protocols for point cloud enhancement rely predominantly on synthetic datasets such as ShapeNet [2] and ModelNet [13], where ground truth is available but data statistics deviate significantly from real-world captures. Even robustness benchmarks that build on these synthetic sets by simulating corruptions demonstrate the challenge of bridging synthetic training to real noise patterns, but still do not fully represent genuine real sensor data distributions [8]. In contrast, comparisons on real-world point clouds are frequently limited to qualitative assessments, which lack both reproducibility and quantitative rigor required for standardized benchmarking. As a result, there is a pressing need for new benchmarks that provide high-quality ground truth under real-world conditions and support reproducible, quantitative evaluation. Unlike existing benchmarks that focus on small-scale scenes or single tasks, this grand challenge will enable holistic evaluation across multiple real-world domains[1], drive the development of realistic solutions for large, heterogeneous point clouds, and catalyze the adoption of standardized perceptual quality metrics that reflect human perception beyond the objective metrics.

Another gap lies in the focus of existing methods: most enhancement algorithms are designed for single-frame point clouds. However, with the increasing availability of dynamic/time-varying (4D) point clouds, exploiting temporal information for spatio-temporal enhancement is an open and underexplored challenge. Moreover, most current datasets only provide geometric information, whereas real-world applications often require both geometry and color fidelity.

To address these challenges, we introduce the **UVG-CWI-DQPC dataset** [3], a dual-quality dynamic point cloud benchmark designed explicitly for enhancement, compression, and quality assessment.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

[1] https://urban3dchallenge.github.io/?utm_source=chatgpt.com#intro

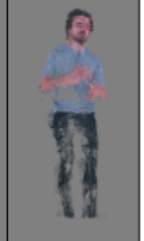
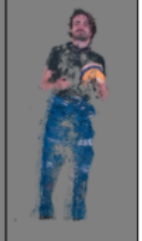
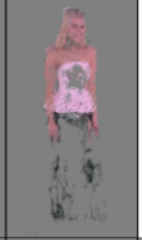
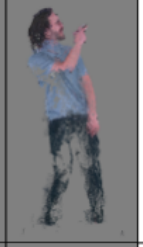
Snapshot		Name and description	Snapshot		Name and description
High end	Cons. grade		High end	Cons. grade	
		**Name:** BlueSpeech **Length:** 169 frames **Description:** A person delivers a speech while using hand gestures. **Specific features:** Moderate movement; Human; Simple textures			**Name:** BlueVolley **Length:** 171 frames **Description:** A person plays with a volleyball, passing the ball behind the back and between the legs. **Specific features:** Fast movement; Human; Object; Interactions; Complex texture; High occlusion
		**Name:** BouncingBlue **Length:** 157 frames **Description:** A person sits and bounces on a gym ball. **Specific features:** Global movement; Human; Object; Interactions; Simple textures			**Name:** FitFluencer **Length:** 201 frames **Description:** A person stretches their back sideways from one side to the other. **Specific features:** Global movement; Human; Simple textures
		**Name:** GoodVision **Length:** 168 frames **Description:** A person conducts an eye exam, presenting letters on a chart one at a time. **Specific features:** Little movement; Human; Objects; Text; Simple textures			**Name:** Mannequin **Length:** 188 frames **Description:** A mannequin wearing a head-mounted display (HMD) and a T-shirt with various logos stands still. The T-shirt shifts slightly. **Specific features:** Static sequence; Objects; Text; Complex textures
		**Name:** OrangeKettlebell **Length:** 170 frames **Description:** A person performs multiple repetitions of the kettlebell swing exercise. **Specific features:** Global movement; Human; Object; Simple textures			**Name:** PinkNoir **Length:** 201 frames **Description:** A person adopts various poses facing the camera. **Specific features:** Little movement; Human; Simple textures
		**Name:** TicTacToe **Length:** 165 frames **Description:** Two persons play Tic Tac Toe using plastic building bricks. **Specific features:** Fast movement; Humans; Objects; Interactions			**Name:** TrumanShow **Length:** 171 frames **Description:** A person greets cameras all around with smiles and gestures. **Specific features:** Fast movement; Human; Simple textures
		**Name:** VictoryHeart **Length:** 197 frames **Description:** A person greets the camera with friendly hand gestures and forms a heart shape. **Specific features:** Moderate movement; Human; Simple textures			**Name:** VirtualLife **Length:** 196 frames **Description:** A person wearing a head-mounted display (HMD) engages in virtual reality gameplay, involving arm movements and body rotations. **Specific features:** Fast movement; Human; Object; Simple textures

Figure 1: Characteristics of the Dual-Quality Point Cloud Sequences in the UVG-CWI-DQPC Dataset

The dataset comprises 12 dynamic sequences, as shown in Figure1, captured simultaneously by: (1) a high-end multi-sensor system producing high-fidelity point clouds after extensive processing; and (2) a consumer-grade RGB-D setup with lightweight processing and open-source tools. Each sequence provides paired high-end ground-truth point clouds, raw RGB-D footage, calibration data, and tools for point cloud generation. This unique dual-quality design enables direct benchmarking of enhancement algorithms across densification, registration, and perceptual quality tasks.

Beyond the dataset, our proposed Grand Challenge targets several important research gaps:

- **Unified enhancement frameworks**: Real-world point clouds are simultaneously affected by multiple degradations (noise, sparsity, occlusion). Current methods often address these factors independently. Developing unified approaches, like a sequecial of denoising, completion, and upsampling is critical for robust 3D processing.
- **Dynamic and color point clouds**: Our benchmark evaluates algorithms not only on geometry but also on temporal consistency and color fidelity, extending beyond the scope of existing datasets.
- **Comprehensive evaluation metrics**: In addition to widely used geometry-based distances (e.g., Chamfer Distance), we will assess texture fidelity using image-based perceptual metrics for color (e.g., PSNR and SSIM) as well as computational efficiency (e.g., average runtime per point cloud sequence). For the top three submissions, we will further conduct a controlled subjective quality assessment with human participants. The final evaluation will therefore combine both quantitative and qualitative measures. Final rankings will be based on a weighted combination of these metrics, reflecting both reconstruction quality and practical usability.

Through this Grand Challenge, we aim to establish the first systematic benchmark for real-world, dynamic, and color point cloud enhancement. This will stimulate new research directions, foster fair comparisons, and accelerate progress toward practical and robust 3D vision solutions.

3 Task Definition and Participation Guidelines

3.1 Overview and Tracks

The goal of the Grand Challenge is to advance algorithms that **enhance real-world, dynamic (4D) color point clouds** by producing accurate, temporally consistent, and visually faithful reconstructions from consumer-grade captures. Participants may submit methods that address the following question:

Unified Enhancement: a single method that jointly addresses denoising, completion, and upsampling for dynamic color point clouds. And it can be composed of the following methods:

(1) **Denoising & Refinement**: remove noise and outliers while preserving fine geometric details.
(2) **Completion & Inpainting**: recover missing geometry and occluded regions in a sequence.
(3) **Upsampling & Densification**: increase sampling density and improve point distribution uniformity.

We provide the benchmark results (low-quality point cloud vs. high-quality point cloud) as the baseline, along with the performance of five state-of-the-art methods (reconstructed point cloud vs. high-quality point cloud) on the leaderboard. All submissions will be ranked on the overall leaderboard. The top three submissions will undergo a standard subjective pairwise comparison [1].

3.2 Dataset and Splits

The challenge is based on the UVG-CWI-DQPC dataset, which contains 12 dynamic sequences captured simultaneously with (1) a high-end multi-sensor system producing high-fidelity processed point clouds (serving as ground truth), and (2) a consumer-grade RGB-D capture pipeline producing raw footage and derived point clouds.

We propose the following split:

- **Training set:** 8 sequences (full paired high-end and consumer-grade data).
- **Validation set:** 2 sequences (paired, with ground truth provided).
- **Test set:** 2 sequences (high-end ground truth withheld; participants run their method on consumer-grade inputs and submit outputs).

For dynamic sequences, participants should process each frame in temporal order. Ground-truth geometry and texture for the test set will remain private; evaluation scripts will be run by organizers.

3.3 Input / Output Format

Input: For each sequence, point clouds captured by consumer-grade cameras alongside the raw RGB and depth images will be provided.

Output: Participants must submit enhanced point clouds in a standard format (PLY) for every frame in the test sequences, with per-point color (RGB). Each submission must include:

- A compressed archive containing per-frame point cloud files (one file per frame).
- A JSON manifest describing sequence names, frame indices, coordinate system, and any post-processing applied.
- A runtime log reporting per-frame processing time and hardware used.
- A short README (max 2 pages) describing the method and external data used for training (if any).

Point Cloud Format

- pred $\in \mathbb{R}^{N\times 6}$: Enhanced point cloud with N points
- gt $\in \mathbb{R}^{M\times 6}$: Ground truth point cloud with M points

3.4 Evaluation Metrics and Scoring

We evaluate submissions using a combination of geometric, color, perceptual, temporal, and efficiency metrics [6, 9, 15, 16] to provide a comprehensive benchmark. The final score is computed as a weighted combination of the individual metrics described below.

Geometrical Metrics.

- **Chamfer Distance.** We report the symmetric nearest-neighbor distance between the predicted point set and the ground truth, capturing overall geometric deviation.

- **F-score** $\in [0, 1]$. We compute precision and recall under a fixed distance threshold τ in 3D space, then take their harmonic mean (F1). Intuitively, it rewards reconstructions that are both *accurate* (high precision) and *complete* (high recall) [11].

Color Fidelity.
PSNR. We compute PSNR directly on the point cloud colors after establishing point correspondences between the enhanced result and the ground truth. Following common practice in visual quality assessment, PSNR is computed in YUV space (on the luminance channel) as a simple baseline for color/texture distortion: higher PSNR indicates smaller mean-squared error in color signals [4, 5].

Perceptual Quality Metrics.

- **PCQM.** PCQM is a full-reference objective point cloud quality metric designed for *colored* point clouds. It combines geometry- and color-related features into an optimally weighted score calibrated against human subjective judgments, making it more perceptually meaningful than purely geometric distances for rendered content [6].
- **Projection-based SSIM.** Following the common *3D-to-2D projection* paradigm for point cloud quality assessment, we render multiple views of both the enhanced and ground-truth point clouds (using consistent camera viewpoints and rendering settings), and compute SSIM on the resulting 2D images. This captures view-dependent structural degradations that align better with human perception during visual inspection than point-to-point distances alone [12, 15].
- **LPIPS.** LPIPS measures perceptual similarity between images using deep feature distances (instead of pixel-wise errors). We apply it on the same set of projected views as above to quantify perceptual differences in a way that correlates well with human judgments, especially for complex texture/appearance changes [16].

Temporal Consistency.
Pooling Method. For dynamic sequences, per-frame quality scores are aggregated into a single sequence-level score using temporal pooling. By default, we apply average pooling over frames. Participants may optionally adopt alternative pooling strategies (e.g., weighted or semantics-aware pooling) that better reflect the temporal characteristics of their enhancement methods. This allows fair evaluation of sequence-level enhancement performance while accommodating methods with different temporal modeling assumptions.

Efficiency.
Runtime per Sequence. We report average processing time per frame and total runtime per sequence under a specified hardware setting. This encourages practical methods suitable for real-world pipelines, where throughput and latency constraints matter.

All metric terms will be normalized to a common scale prior to weighting; for distance-based metrics lower is better, while for similarity metrics, higher is better. Organizers will publish the exact normalization and scoring code.

4 Challenge organizers

Xuemei Zhou is a Ph.D. candidate at TU Delft and CWI (Centrum Wiskunde & Informatica). She holds a B.S. degree in Industrial Engineering with a minor in Accounting from Northeast Forestry University, and an M.S. degree in Computer Technology from the Shenzhen Institute of Advanced Technology, University of Chinese Academy of Sciences. Her research interests lie in multimedia systems, image and 3D visual processing, and human–computer interaction.

Jack Jansen is a researcher at Centrum Wiskunde & Informatica. His research focus is on empowering people to put available technology to a use they themselves envision including activities ranging from systems (Amoeba), languages (Python), synchronized networked multimedia (SMIL, Ambulant, Ta2, Vconect, 2-Immerse) and IoT to his current interest in social VR. Jansen received his M.Sc. degree from Vrije Universiteit.

Pablo Cesar is a senior researcher at CWI (Centrum Wiskunde & Informatica, the Dutch National Research Institute for Mathematics and Computer Science), where he leads the Distributed and Interactive Systems group. He is also a Professor of Human-Centered Multimedia Computing at Delft University of Technology. He is an ACM Distinguished Member, an IEEE Senior Member, and the recipient of the 2020 Netherlands Prize for ICT.

Guillaume Gautier is a postdoctoral researcher with the Ultra Video Group (UVG) at Tampere University (TAU), Finland. In 2020, he received his Ph.D. in Signal, Image, and Vision from INSA Rennes, France, where his thesis focused on the secure implementation of encryption algorithms on embedded platforms. His current research interests include volumetric video generation, compression, and transmission. He also contributes to open-source projects, including the award-winning point cloud encoder, uvgVPCCenc.

Alexandre Mercat received the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the Institut National des Sciences Appliquées (INSA) of Rennes, France, in 2015 and 2018, respectively. Since 2024, he has been an Assistant Professor (tenure track) in Computing Sciences at Tampere University (TAU), Tampere, Finland. His research focuses on video coding, real-time implementations of next-generation video coding standards, complexity-aware and energy-aware video coding, and video coding for machines, AI, and immersive media applications (VR/AR/MR/XR). He has authored over 40 peer-reviewed publications and received the Best Paper Award at VCIP 2024 and the Best Open Dataset and Software Paper Awards at ACM MMSys 2020 and ACM MMSys 2025. He is a member of the IEEE Visual Signal Processing and Communications Technical Committee (VSPC-TC) and has served on the technical program committees of international conferences including SAMOS, SiPS, ICIP, and ISCAS. In 2021, he co-founded and co-edits the *Insights from Negative Results* track in JSPS.

Jarno Vanne received the M.Sc. degree in Information Technology and the Ph.D. degree in Computing and Electrical Engineering from Tampere University of Technology (TUT), Tampere, Finland, in 2002 and 2011, respectively. He is currently a Professor with the Unit of Computing Sciences, Tampere University (TAU), Tampere, Finland. He is also the Founder and Leader of the Ultra Video Group, the leading academic video research group in Finland. He has served as a project manager for 25 international and national research projects and is the author of over 100 peer-reviewed publications. His research interests include real-time video coding and streaming, volumetric video communication, hybrid human and machine vision, vision-based remote operation in smart manufacturing, vision-based driver assistance systems, and virtual modeling for smart mobility.

Irene Viola is a senior (tenured) researcher at Centrum Wiskunde en Informatica (CWI) in Amsterdam, The Netherlands. She received her M.Sc. in Computer Engineering from the Polytechnic University of Turin, Italy, in 2015, and her Ph.D. in Electrical Engineering from the Ecole Polytechnique Federale de Lausanne, Switzerland, in 2019. Her research interests include compression, delivery, and QoE for immersive media systems. She has served as a Qualinet chair for the task force on Immersive Media Experiences since 2017 and is actively involved in standardization bodies, including MPEG and ITU. She has served as Technical Program Committee (TPC) chair for the ACM Multimedia Systems conference (MMSys) workshop Immersive Mixed And Virtual Environment Systems (MMVE) in 2021, for MMSys in 2022, for Quality of Multimedia Experiences (QoMEX) in 2023, and ACM International Conference on Interactive Media Experiences (IMX) in 2024, and has organised three editions of the Spring School in Social XR (2023-2025).

5 Commitment

If our proposal is accepted, we commit to publishing and maintaining a dedicated website for the Grand Challenge, providing up-to-date information, datasets, and task descriptions for at least the next three years. For any questions regarding the challenge, please contact Xuemei Zhou (xuemei.zhou@cwi.nl) and Irene Viola (irene.viola@cwi.nl).

References

[1] Evangelos Alexiou, Irene Viola, Tomás M Borges, Tiago A Fonseca, Ricardo L De Queiroz, and Touradj Ebrahimi. 2019. A comprehensive study of the rate-distortion performance in MPEG point cloud compression. *APSIPA Transactions on Signal and Information Processing* 8 (2019), e27.

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[3] Guillaume Gautier, Xuemei Zhou, Thong Nguyen, Jack Jansen, Louis Fréneau, Marko Viitanen, Uyen Phan, Jani Käpylä, Irene Viola, Alexandre Mercat, et al. 2025. UVG-CWI-DQPC: Dual-Quality Point Cloud Dataset for Volumetric Video Applications. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 13112–13118.

[4] International Telecommunication Union. 2024. ITU-T Recommendation P.910: Subjective video quality assessment methods for multimedia applications. https://www.itu.int/rec/t-rec-p.910.

[5] ISO/IEC JTC1/SC29/WG11 MPEG. 2019. *Common Test Conditions for Point Cloud Compression.* MPEG Technical Report N18668. ISO/IEC JTC1/SC29/WG11, Marrakesh, Morocco.

[6] Guillaume Meynet, Yana Nehmé, Julien Digne, and Guillaume Lavoué. 2020. PCQM: A Full-Reference Quality Metric for Colored 3D Point Clouds. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. doi:10.1109/QoMEX48832.2020.9123147

[7] Siwen Quan, Junhao Yu, Ziming Nie, Muze Wang, Sijia Feng, Pei An, and Jiaqi Yang. 2024. Deep learning for 3d point cloud enhancement: A survey. *arXiv preprint arXiv:2411.00857* (2024).

[8] Jiawei Ren, Liang Pan, and Ziwei Liu. 2022. Benchmarking and Analyzing Point Cloud Classification under Corruptions. In *International Conference on Machine Learning (ICML)*.

[9] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2019. Multi-View 3D Reconstruction with Transformers?. arXiv:1905.03678. We cite this work for the standard F-score definition used in multi-view 3D reconstruction evaluation..

[10] Ignacio Vizzo, Benedikt Mersch, Rodrigo Marcuzzi, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. 2022. Make it dense: Self-supervised geometric scan completion of sparse 3d lidar scans in large outdoor environments. *IEEE Robotics and Automation Letters* 7, 3 (2022), 8534–8541.

[11] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. 2021. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5722–5731.

[12] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861

[13] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.

[14] ChengKai Xia, Fan Lu, Bin Li, Guo Yu, Alois Knoll, and Guang Chen. 2025. Points, Images and Texts: Boosting Point Cloud Completion with Multi-Modal Features.

In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12759–12765.

[15] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun. 2021. Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration. *IEEE Transactions on Multimedia* 23 (2021), 3877–3891. doi:10.1109/TMM.2020.3033117

[16] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595. doi:10.1109/CVPR.2018.00068