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

1 Overview

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;
- **Optimization and interpolation methods**, which explicitly exploit geometric consistency;
- **Multi-modal approaches**, which leverage 2D images or videos to provide complementary structural or textural priors;
- **AI-generated 3D content approaches**, where enhancement is viewed as generative point cloud synthesis.

Overall, point cloud enhancement encompasses three main sub-tasks: (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 Rules for participation

Participants from all over the world can participate in the challenge. The participants must be of legal age to participate. The maximum team size is 5. Each Team is required to appoint a team leader. This team leader will be the main contact point between the Organizers and the Team.

1. Participants must register at register here.

2. The participants ranking in the top 3 upon centralizing results for all submissions must present a summary of their technology during the dedicated ACM MM session.

3. The coordinators will make available a training set with the corresponding ground truth (high quality point cloud) as well as a validation set without ground truth. The test set and the corresponding ground truth will not be available to the competitors during the challenge. The contributors will send the code (Python preferably or Matlab) using a Docker container. The final evaluation will be based only on the test set.

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 2 classical 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 are invited to submit a challenge paper to ACM MM 2026.

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:** 9 sequences (full paired high-end and consumer-grade data).
  Specifically: BlueSpeech, BlueVolley, BouncingBlue, FitFluencer, GoodVision, Mannequin, OrangeKettlebell, PinkNoir, TicTacToe.
- **Validation set:** 3 sequences (paired, with ground truth provided).
  Specifically: TrumanShow, VictoryHeart, VirtualLife.
- **Test set:** 3 sequences (withheld; participants solution ran by the organisers).

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 [3, 6, 7, 4] 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) [5].

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 [1, 2].

Perceptual Quality Metrics.

- **PCQM.**
- **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, 6 views around the

bounding box, including the left, right, front, back, top and down), 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.

- **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.

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: .

References

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

[2] ISO/IEC JTC1/SC29/WG11 MPEG. Common test conditions for point cloud compression. MPEG Technical Report N18668, ISO/IEC JTC1/SC29/WG11, Marrakesh, Morocco, 2019.

[3] Guillaume Meynet, Yana Nehmé, Julien Digne, and Guillaume Lavoué. Pcqm: A full-reference quality metric for colored 3d point clouds. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2020.

[4] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d reconstruction with transformers? 2019. We cite this work for the standard F-score definition used in multi-view 3D reconstruction evaluation.

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

[6] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun. Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration. *IEEE Transactions on Multimedia*, 23:3877–3891, 2021.

[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.