WebGSBench: A Comprehensive Benchmarking System for Web-Based Gaussian Splatting Deployment

ANONYMOUS SUBMISSION, Anonymous for Review, Anonymous

While 3D Gaussian Splatting (3DGS) is advancing rapidly in reconstruction quality and compression efficiency, the research community lacks standardized tools to evaluate how these algorithms perform under web deployment constraints, where format fragmentation, browser heterogeneity, and hardware diversity create fundamentally different challenges than desktop evaluation. We present **WebGSBench**, a benchmarking framework for systematic evaluation of 3DGS under web-specific conditions. Our system measures: (1) perceptual quality degradation across web formats (.ply, .splat, .ksplat, .spz), (2) performance characteristics across browsers, and (3) temporal stability during user interaction, a quality dimension invisible to static image metrics. Through experiments on standard benchmark scenes, we reveal trade-offs and failure modes that existing evaluation pipelines cannot capture. Code and data are publicly available.

1 Introduction

The rapid adoption of 3D Gaussian Splatting [17] has fundamentally changed the landscape of real-time novel view synthesis. Unlike previous neural rendering approaches that require expensive volumetric ray marching [21], 3DGS achieves real-time rendering through efficient point-based splatting, making it particularly suitable for web-based applications.

Web deployment has become a primary consumption mode for 3DGS content. E-commerce platforms leverage 3DGS for hyper-realistic product visualization, offering customers immersive experiences without specialized software. Cultural heritage organizations use web-based 3DGS viewers to democratize access to digitized artifacts. Photogrammetry platforms (Polycam [27], Scaniverse, KIRI Engine) enable millions of users to capture and share 3DGS models directly from mobile devices, with processing and viewing handled entirely in the browser. The New York Times R&D deployed 3DGS for immersive journalism [31]. Industry adoption extends to digital twins, virtual tours, and WebXR experiences, with major cloud providers highlighting 3DGS as a transformative technology for web-scale 3D content delivery [1].

Despite this widespread web deployment, the research community currently lacks a unified framework for evaluating how novel 3DGS algorithms perform under web constraints, where file size limitations, browser heterogeneity, and hardware diversity create fundamentally different challenges than offline rendering.

This gap between academic development and practical deployment has led to several critical issues:

1. **Inconsistent Evaluation**: Papers report quality metrics (PSNR, SSIM, LPIPS [32]) on desktop implementations but ignore web-specific constraints
2. **Format Fragmentation**: At least 5+ incompatible formats (.ply, .splat, .ksplat, .spz, compressed variants) exist with no systematic comparison [23, 26]
3. **Non-Reproducible Comparisons**: Researchers lack standardized test scenes and evaluation protocols for web deployment
4. **Missing Performance Metrics**: Loading time, memory footprint, and browser-specific rendering performance are rarely reported

We argue that addressing these gaps requires not just a dataset, but a **living benchmarking system**. This is analogous to how Papers with Code [24] transformed machine learning reproducibility, or how the KITTI benchmark [11] standardized autonomous driving evaluation.

2 Background and Related Work

2.1 Evolution of Neural Rendering and 3D Gaussian Splatting

Neural Radiance Fields (NeRF) [21] introduced the paradigm of learning continuous scene representations via volumetric rendering and neural networks. While achieving photorealistic novel view synthesis, NeRF's volumetric rendering approach requires dense sampling along each ray, making real-time rendering infeasible and limiting practical deployment.

3D Gaussian Splatting [17] addressed this limitation by representing scenes as collections of 3D Gaussians with learnable parameters (position, covariance, opacity, spherical harmonic coefficients). By leveraging differentiable point-based rasterization, 3DGS achieves:

- Real-time rendering (≥30 fps at 1080p resolution)
- Explicit scene representation enabling faster training
- Compatibility with standard graphics pipelines

The real-time capability of 3DGS has triggered explosive research growth, with numerous extensions addressing dynamic scenes, sparse input views, compression, and relighting.

2.2 Existing Benchmarking Efforts

Traditional NeRF/3DGS Benchmarks focus primarily on reconstruction quality:

- **Mip-NeRF 360** [5]: 9 scenes with outdoor/indoor captures, evaluated on PSNR/SSIM/LPIPS
- **Tanks and Temples** [18]: Multi-view stereo benchmark adapted for NeRF evaluation
- **DTU** [15]: Controlled captures with ground truth geometry

These benchmarks established quality evaluation standards but do not address web deployment concerns.

Recent Compression & Efficiency Benchmarks:

- **Splatwizard** [2]: Unified toolkit for evaluating 3DGS compression methods, measuring file size, rendering FPS, and quality metrics. However, it focuses on offline compression pipelines rather than web-native evaluation.

Author's Contact Information: Anonymous Submission, Anonymous for Review, Anonymous, Anonymous.

2026. ACM XXXX-XXXX/2026/1-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- **GS-QA** [3]: Comprehensive quality assessment analyzing 18 objective metrics across diverse scenes. Lacks performance and deployment evaluation.
- **SIGGRAPH Asia 2025 3DGS Challenge** [30]: Focuses on reconstruction speed (<60s) and PSNR, not web deployment.

Web Viewer Implementations (not benchmarks):

- antimatter15/splat [4]: WebGL viewer with CPU sorting
- mkkellogg/GaussianSplats3D [16]: Three.js with GPU-based sorting
- cvlab-epfl/gaussian-splatting-web [8]: WebGPU implementation

These viewers demonstrate feasibility but lack standardized evaluation protocols or comparative analysis across formats and rendering backends.

2.3 Format Landscape and Fragmentation

The 3DGS ecosystem has fragmented into multiple incompatible formats, each optimizing for different trade-offs, as shown in Table 1.

Table 1. 3D Gaussian Splatting web format comparison

Format	Size	Features
PLY	1.0×	Full SH, uncompressed
Compressed PLY	0.25×	Quantized, trimmed SH
.splat	0.4×	No SH, simplified
.ksplat	Var.	Compression levels
.spz	0.1×	64B/splat, SH kept

Critical Problem: Research papers typically report results on .ply files evaluated offline, but web deployment requires format conversion with poorly understood quality-performance trade-offs. No systematic study compares how novel 3DGS algorithms degrade across different web formats.

2.4 The Dataset Contribution Model

Highly-cited benchmark papers in computer vision typically provide:

1. **Standardized Data**: Curated, diverse, and challenging (ImageNet [10], COCO [19], KITTI [11])
2. **Evaluation Protocols**: Metrics, splits, and comparison methodology
3. **Leaderboards/Infrastructure**: Tools for reproducible comparison
4. **Community Adoption**: Becomes the de facto standard

Examples include ImageNet [10] (14M images, 100k+ citations), COCO [19] (object detection/segmentation benchmark, 50k+ citations), and ScanNet [9] (indoor RGB-D dataset with reconstruction benchmarks, 5k+ citations).

2.5 Benchmarking Systems as Research Infrastructure

Beyond passive datasets, several influential papers have contributed **active benchmarking systems** that fundamentally advanced their fields through standardized evaluation infrastructure.

3D Reconstruction Benchmarks: Tanks and Temples [18] provided not just multi-view stereo data, but an online evaluation server with standardized metrics and public leaderboards, enabling fair comparison of reconstruction algorithms. Similarly, ETH3D [29] combined high-precision laser-scanned ground truth with an automated evaluation platform that computes accuracy, completeness, and F1 scores. This established reproducible protocols that advanced the field. These systems became de facto standards because they provided *infrastructure*, not just data.

Embodied AI Platforms: Habitat [28] exemplifies how evaluation systems drive research progress. By providing a fast simulator (10,000+ fps), standardized tasks, and reproducible environments, Habitat enabled the embodied AI community to iterate rapidly and compare methods fairly. The platform's impact stems from reducing evaluation friction. Researchers can submit agents and receive immediate, comparable performance metrics.

Performance Benchmarking: MLPerf [20] demonstrates how standardized evaluation infrastructure can establish industry-wide standards. Beyond datasets, MLPerf provides reference implementations, measurement protocols, and leaderboards that enable objective comparison of ML systems across hardware and software configurations.

Critical Pattern: These highly-cited benchmarking papers share common elements: (1) standardized evaluation protocols, (2) automated measurement infrastructure, (3) online platforms for comparison, and (4) open-source tools for reproducibility. Their impact comes not from data collection, but from *lowering the barrier to fair, reproducible comparison.*

The Gap for Web-Based 3DGS: Despite 3DGS's real-time capability making it ideal for web deployment, *no analogous benchmarking system exists* for evaluating web-based performance. Researchers developing novel 3DGS methods have no standardized way to assess how their algorithms perform when deployed to browsers, compressed to web formats, or rendered on diverse hardware. This infrastructure gap limits the field's ability to optimize for real-world deployment.

Our proposed system extends this model of benchmarking infrastructure to web-based 3DGS evaluation, providing researchers with automated tools to submit outputs and receive comprehensive deployment metrics. This is analogous to how Tanks and Temples standardized 3D reconstruction evaluation and MLPerf standardized ML performance benchmarking.

3 Motivation: The Web Deployment Gap

3.1 Why Web-Based Evaluation Matters

Real-World Consumption Pattern: Web-based deployment represents a significant mode of 3DGS content consumption:

- E-commerce 3D product visualization (Amazon, Shopify)
- Virtual tours and digital twins (real estate, museums)
- AR/VR experiences via WebXR
- Photogrammetry sharing platforms (Scaniverse, Polycam)

Web Constraints Differ Fundamentally from Desktop:

1. **File Size**: Network bandwidth limitations favor 10–100× compression

(2) **Memory**: Mobile browsers have strict memory limits (often <4GB)
(3) **API Variability**: WebGL 2.0 vs WebGPU performance varies drastically
(4) **Browser Diversity**: Chrome, Safari, Firefox have different optimization characteristics
(5) **Hardware Range**: From high-end desktop GPUs to integrated mobile GPUs

Current Research Gap: 95%+ of 3DGS papers evaluate only on desktop with .ply files, ignoring the conversion/compression pipeline required for web deployment. This creates a "reproducibility crisis" for practitioners attempting to deploy academic methods.

3.2 Fragmentation Hinders Progress

Format Incompatibility: Converting a novel sparse-view 3DGS method to .spz format may destroy the quality gains reported in the paper. Researchers currently have no way to know without manual implementation and testing.

Non-Standard Evaluation: Different papers use different:

- Test scenes (Mip-NeRF 360 vs custom captures)
- Quality metrics (some report LPIPS, others don't)
- Performance metrics (FPS on NVIDIA A100 vs RTX 3090)
- Rendering implementations (CUDA vs WebGL vs WebGPU)

This makes it **impossible to fairly compare** methods across papers.

3.3 What's Missing: A Unified Benchmarking System

Drawing inspiration from successful benchmarking systems in adjacent fields:

MLPerf [22] (ML performance benchmarking):

- Standardized model implementations
- Hardware-agnostic measurement protocols
- Automated submission and validation
- Public leaderboards

Hugging Face Spaces [13] (ML model deployment):

- Unified interface for model comparison
- Interactive visualization
- Reproducible environment

Our Proposed System combines these ideas for 3DGS:

(1) **Standardized Test Scenes**: Curated dataset spanning diverse scene types, complexity levels, and capture conditions
(2) **Multi-Format Pipeline**: Automated conversion to all major web formats with quality/size reporting
(3) **Web-Native Evaluation**: Performance profiling across WebGL/WebGPU, multiple browsers, and devices
(4) **Interactive Comparison**: Side-by-side visualization of quality-performance trade-offs
(5) **Automated Metrics**: PSNR, SSIM, LPIPS, file size, load time, FPS, memory footprint
(6) **Researcher-Friendly Submission**: Upload .ply → receive comprehensive benchmark report

4 Research Impact and Community Value

4.1 Enabling Reproducible Research

By providing a standardized platform, we enable:

- **Fair Comparison**: All methods evaluated on identical test scenes, formats, and rendering conditions
- **Reproducibility**: Researchers can verify claims by submitting their outputs
- **Transparency**: Public metrics and visualizations reveal true deployment trade-offs

This addresses the reproducibility crisis identified in recent ML/CV meta-research [25].

4.2 Accelerating Algorithm Development

Feedback Loop: Currently, researchers spend weeks implementing web deployment to test their methods. Our system provides instant feedback, accelerating iteration cycles.

Multi-Objective Optimization: Researchers can simultaneously optimize for:

- Visual quality (PSNR/SSIM/LPIPS)
- File size (.ply → .spz conversion efficiency)
- Rendering performance (FPS across devices)
- Loading time (critical for UX)

Identifying Failure Modes: Automated testing across formats may reveal degradation patterns invisible in desktop evaluation.

4.3 Bridging Academia and Industry

Industry practitioners currently struggle to:

(1) **Select Methods**: Which 3DGS variant works best for their use case?
(2) **Predict Performance**: Will this method meet our 60 FPS target on mobile?
(3) **Deployment Guidance**: Which format should we use for our constraints?

Our benchmark provides evidence-based answers, increasing the real-world impact of academic research.

4.4 Driving Standardization

Similar to how COCO metrics became the de facto standard for object detection, our benchmark can:

- Establish standard web deployment metrics
- Encourage format convergence or interoperability
- Inform development of future web standards (WebGPU features, compression APIs)

5 Proposed Contributions

Our system makes the following novel contributions:

5.1 Infrastructure Contributions

(1) **Curated Test Dataset**: Diverse, challenging scenes with ground truth for reproducible evaluation
(2) **Automated Benchmarking Pipeline**: End-to-end conversion, rendering, and metric computation

(3) **Web-Based Comparison Interface**: Interactive visualization of quality-performance trade-offs, including arena-style anonymous A/B comparison for perceptual assessment
(4) **Open-Source Toolkit**: Extensible codebase for community-driven improvement

5.2 Scientific Contributions

(1) **Comprehensive Format Study**: First systematic comparison of 3DGS web formats across quality/performance axes
(2) **Deployment Performance Characterization**: Profiling across browsers, devices, and rendering backends
(3) **Compression Trade-off Analysis**: Quantifying quality degradation vs file size for different 3DGS variants
(4) **Best Practices**: Evidence-based guidelines for web deployment of 3DGS content

5.3 Community Impact

(1) **Reproducible Comparisons**: Researchers can verify and compare methods objectively
(2) **Lower Barrier to Entry**: Newcomers can quickly understand format/deployment landscape
(3) **Industry Adoption**: Practitioners gain confidence in deploying academic methods
(4) **Future-Proofing**: Extensible system can incorporate new formats and metrics as the field evolves

6 Related Systems and Differentiation

6.1 Machine Learning Benchmarks

Papers with Code [24] provides leaderboards for ML tasks but:

- Relies on author-reported metrics (not automated)
- Doesn't provide deployment/inference benchmarking
- *Our system*: Automated evaluation with web-specific metrics

Hugging Face Spaces [13] enables model deployment but:

- Focuses on NLP/vision models, not 3D rendering
- No standardized performance benchmarking
- *Our system*: Standardized 3DGS-specific evaluation

6.2 3D Vision Benchmarks

ScanNet/Matterport3D [9] provide datasets but:

- Focus on RGB-D reconstruction, not novel view synthesis
- No rendering performance evaluation
- *Our system*: Neural rendering quality + web deployment performance

KITTI/Waymo [11] for autonomous driving:

- Excellent model but domain-specific
- *Our system*: Adapts multi-metric, leaderboard-driven approach to 3DGS

6.3 Graphics Performance Benchmarks

GFXBench/3DMark measure GPU performance but:

- Use fixed synthetic workloads
- Not relevant to neural rendering
- *Our system*: Real 3DGS content with scientifically meaningful metrics

7 Methodology

7.1 Scene Selection Strategy

Our objective is not to benchmark datasets or establish method rankings, but to examine how Web-based execution conditions influence the perceptual and systems-level behavior of 3D Gaussian Splatting during interactive use. We therefore adopt a **curated, phenomenon-driven scene selection strategy**, drawing scenes from multiple widely used datasets rather than relying on a single benchmark.

Each selected scene is chosen to isolate a specific rendering or perceptual stress factor relevant to browser-mediated execution, including memory pressure, depth complexity, temporal stability, and sensitivity to high-frequency detail. By curating scenes across diverse sources, we avoid overfitting our analysis to the characteristics of any single dataset and instead focus on behaviors that consistently emerge under Web deployment constraints.

We emphasize that our selection does not aim for exhaustive dataset coverage or statistical representativeness. Instead, **each scene functions as a measurement probe** designed to reveal failure modes and trade-offs that are obscured by conventional offline evaluation pipelines.

7.2 Curated Scene Portfolio

Table 2 presents our curated scene portfolio. We selected six scenes from four datasets, each isolating a specific Web-relevant stress factor:

Stress Factor Coverage. Our selection ensures orthogonal coverage of Web-specific failure modes: (1) **High-frequency detail** (bonsai, flower): compression artifacts on fine texture; (2) **Depth complexity** (garden, train): GPU overdraw from large depth range; (3) **Memory footprint** (playroom, truck): browser memory limits and tab crashes; (4) **Temporal stability** (truck, train): frame drops during camera motion; (5) **Load time** (train, playroom): parsing overhead on large files; (6) **Compression sensitivity** (flower, bonsai): quantization errors on thin geometry.

Cross-Dataset Justification. Using scenes from four datasets (Mip-NeRF 360, Tanks & Temples, Deep Blending, real-world captures) avoids dataset-specific biases. Single-dataset benchmarks risk mistaking dataset characteristics for fundamental algorithmic behavior. Our cross-dataset selection ensures findings reflect Web deployment realities rather than dataset artifacts.

8 Potential Challenges and Solutions

8.1 Dataset Curation

Challenge: Selecting representative test scenes that span the diversity of 3DGS applications.

Solution: Multi-tier dataset:

- **Tier 1**: Standard scenes (Mip-NeRF 360) for compatibility with existing papers
- **Tier 2**: Challenging cases (reflective surfaces, sparse views, dynamic content)
- **Tier 3**: Domain-specific (product scanning, faces, large-scale outdoor)

Table 2. Curated scenes and evaluation rationale. Each scene isolates a specific stress factor relevant to browser-mediated 3DGS execution.

Scene	Dataset	Characteristics	Stress Factor	Relevance to Web-3DGS	Size
bonsai	Mip-NeRF 360 [5]	Indoor object, fine foliage	High-freq. detail	Tests compression artifacts on complex texture; small file ideal for mobile	56 MB
garden	Mip-NeRF 360	Outdoor, vegetation, depth	Depth complexity	Exposes sorting artifacts during camera motion; mid-range memory	98 MB
playroom	Deep Blending [12]	Indoor room, multi-object	Memory footprint	Stresses GPU memory limits; tests browser tab stability under load	453 MB
truck	Tanks & Temples [18]	Outdoor vehicle, specular	Temporal stability	Reveals frame drops during interaction; high splat count challenges sorting	400 MB
train	Tanks & Temples	Large outdoor, geometric	Load time	Tests parsing and upload overhead; extreme case for network constraints	175 MB
flower	Real-world capture	Small, thin geometry	Compression sens.	Exposes quality degradation in .spz; thin structures prone to artifacts	6 MB

8.2 Maintaining Relevance

Challenge: Fast-moving field with new formats and methods appearing constantly.

Solution:

- Modular architecture allowing easy integration of new formats
- Community contribution model (like Hugging Face)
- Regular benchmark updates (quarterly)

8.3 Ground Truth Acquisition

Challenge: Some metrics require ground truth images not always available.

Solution:

- Multi-view captures with held-out views for PSNR/SSIM/LPIPS
- Reference-free metrics (FID, KID) for cases without GT
- User studies for perceptual quality validation

8.4 Browser/Device Variability

Challenge: Infinite combinations of browsers, OS, and hardware.

Solution:

- Focus on representative configurations (Chrome/Safari/Firefox on desktop/mobile)
- Provide raw performance data for community analysis
- Crowdsource additional device testing via open API

9 Future Work

While our current system provides objective quality metrics (PSNR, SSIM) and performance benchmarks, perceptual quality assessment remains an important area for future investigation.

9.1 Perceptual User Studies

Our system includes an arena-style interface for anonymous A/B comparison, inspired by Chatbot Arena [7]. This enables side-by-side perceptual evaluation where users select which rendering appears better without knowing the underlying format. Future work will extend this to comprehensive user studies following established methodologies:

Two-Alternative Forced Choice (2AFC) [14]: The ITU-R BT.500 standard defines rigorous protocols for subjective video quality assessment. Applying these methods to 3DGS format comparison would provide human preference data to validate objective metrics.

Pairwise Preference Aggregation: The Bradley-Terry model [6] enables conversion of pairwise comparisons into global rankings. Similar to how Chatbot Arena uses Elo ratings for LLM evaluation, we can compute format rankings from crowdsourced preference data, revealing perceptual quality differences that PSNR/SSIM may miss.

Research Questions:

- How well do objective metrics (PSNR, SSIM, LPIPS) correlate with human preference?
- At what compression ratio do users notice quality degradation?
- Do temporal artifacts (frame drops, stuttering) outweigh static image quality in user preference?
- Can we identify perceptual quality cliffs where small file size gains cause large preference drops?

These studies would inform compression algorithm design and establish evidence-based format selection guidelines for practitioners.

10 Conclusion

The 3D Gaussian Splatting research community has achieved remarkable progress in reconstruction quality and training speed, but lacks the infrastructure to evaluate real-world web deployment, a significant consumption mode for 3DGS content. Existing benchmarks focus on offline quality metrics while ignoring format fragmentation, browser constraints, and performance variability that practitioners face.

We propose **WebGSBench**, a comprehensive benchmarking system that:

(1) Provides standardized test scenes and evaluation protocols

(2) Automates conversion across web formats with quality and performance profiling
(3) Enables reproducible comparison of research methods under realistic deployment conditions
(4) Bridges the gap between academic development and practical deployment

By following the successful model of ImageNet, COCO, and MLPerf, our system can become the de facto standard for 3DGS web deployment evaluation, accelerating research progress and increasing real-world impact. This contribution is timely. As the field matures, standardized evaluation infrastructure becomes critical for continued advancement.

References

[1] Amazon Web Services. 2024. 3D Gaussian Splatting: Performant 3D Scene Reconstruction at Scale. https://aws.amazon.com/blogs/spatial/3d-gaussian-splatting-performant-3d-scene-reconstruction-at-scale/. Accessed: 2026-01-16.

[2] Anonymous. 2024. Splatwizard: A Benchmark Toolkit for 3D Gaussian Splatting Compression. arXiv:2512.24742 [cs.CV] https://arxiv.org/abs/2512.24742 arXiv:2512.24742.

[3] Anonymous. 2025. GS-QA: Comprehensive Quality Assessment Benchmark for Gaussian Splatting View Synthesis. arXiv:2502.13196 [cs.CV] https://arxiv.org/abs/2502.13196 arXiv:2502.13196.

[4] antimatter15. 2023. WebGL 3D Gaussian Splat Viewer. https://github.com/antimatter15/splat. Accessed: 2026-01-13.

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5460–5469. doi:10.1109/CVPR52688.2022.00539

[6] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. doi:10.2307/2334029

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs.AI] https://arxiv.org/abs/2403.04132

[8] CVLAB EPFL. 2024. Gaussian Splatting Web. https://github.com/cvlab-epfl/gaussian-splatting-web. Accessed: 2026-01-13.

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5828–5839. doi:10.1109/CVPR.2017.261

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255. doi:10.1109/CVPR.2009.5206848

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3354–3361. doi:10.1109/CVPR.2012.6248074

[12] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep Blending for Free-Viewpoint Image-Based Rendering. *ACM Transactions on Graphics (SIGGRAPH Asia)* 37, 6 (November 2018), 257:1–257:15. doi:10.1145/3272127.3275084

[13] Hugging Face. 2016. Hugging Face: The AI Community Building the Future. https://huggingface.co/. Accessed: 2026-01-13.

[14] ITU-R. 2023. *Methodology for the Subjective Assessment of the Quality of Television Pictures*. Technical Report BT.500-15. International Telecommunication Union. https://www.itu.int/rec/R-REC-BT.500 Recommendation ITU-R BT.500-15.

[15] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large Scale Multi-view Stereopsis Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 406–413. doi:10.1109/CVPR.2014.59

[16] Mark Kellogg. 2023. GaussianSplats3D: Three.js-based Implementation of 3D Gaussian Splatting. https://github.com/mkkellogg/GaussianSplats3D. Accessed: 2026-01-13.

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). doi:10.1145/3592433

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (2017), 78:1–78:13. doi:10.1145/3072959.3073599

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*. Springer, 740–755. doi:10.1007/978-3-319-10602-1_48

[20] Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim Hazelwood, Andrew Hock, Xinyuan Huang, Yangqing Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Guokai Ma, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St John, Tsuguchika Tabaru, Carole-Jean Wu, Lingjie Xu, Masafumi Yamazaki, Cliff Young, and Matei Zaharia. 2020. MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. In *Proceedings of Machine Learning and Systems*, Vol. 2. 336–349. https://proceedings.mlsys.org/paper/2020/hash/02522a2b2726fb0a03bb19f2d8d9524d-Abstract.html

[21] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science, Vol. 12346)*. Springer, 405–421. doi:10.1007/978-3-030-58452-8_24

[22] MLCommons. 2018. MLPerf: Fair and Useful Benchmarks for Measuring Training and Inference Performance. https://mlcommons.org/. Accessed: 2026-01-13.

[23] Niantic Labs. 2024. Open-sourcing .SPZ: it's .JPG for 3D Gaussian Splats. https://scaniverse.com/news/spz-gaussian-splat-open-source-file-format. Accessed: 2026-01-13.

[24] Papers with Code. 2018. Papers With Code: The Latest in Machine Learning. https://paperswithcode.com/. Accessed: 2026-01-13.

[25] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research* 22, 164 (2021), 1–20. https://jmlr.org/papers/v22/20-303.html

[26] PlayCanvas. 2024. Compressing Gaussian Splats. https://blog.playcanvas.com/compressing-gaussian-splats/. Accessed: 2026-01-13.

[27] Polycam Inc. 2024. Polycam: Cross-Platform 3D Scanning and Photogrammetry. https://poly.cam. Accessed: 2026-01-16.

[28] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9339–9347. doi:10.1109/ICCV.2019.00943

[29] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3260–3269. doi:10.1109/CVPR.2017.272

[30] SIGGRAPH Asia. 2025. 3D Gaussian Splatting Challenge. https://gaplab.cuhk.edu.cn/projects/gsRaceSIGA2025/. Accessed: 2026-01-13.

[31] The New York Times R&D. 2024. A Field Guide To Gaussian Splatting. https://rd.nytimes.com/projects/gaussian-splatting-guide/. Accessed: 2026-01-16.

[32] 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