



Overview of automated 3d space construction methods from videos using artificial intelligence

Phan Thi Xuan Trang¹, Dang Manh Huy¹, Bui Thi Diem Trinh¹, Tran Thanh Nam^{1*}

¹College of Digital Technology and Artificial Intelligence, Nam Can Tho University

*Corresponding author: ttnam@nctu.edu.vn

DOI: 10.64632/jsde.eng.4.2026.906

ABSTRACT

Received: Apr 29, 2026

Revised: May 9, 2026

Accepted: May 16, 2026

Keywords: 3D

reconstruction, 3D Gaussian
splatting, artificial
intelligence, computer vision,
monocular video, neural
SLAM, real-time rendering

This research focuses on studying, reviewing, and analyzing methods for automatically reconstructing 3D spatial structures from single-image video data (monocular video) through the integration of modern artificial intelligence models such as Deep Learning and 3D Gaussian Splatting. This solution overcomes the cost and time limitations of traditional photometric methods, providing real-time processing capabilities with superior photorealistic image quality. Simultaneously, the research applies a multi-dimensional evaluation system based on metrics such as PSNR, LPIPS, and FPS to optimize the balance between geometric accuracy and operational performance on common devices.

1 INTRODUCTION

1.1 Research background

With the explosion of digital transformation and virtual reality/arrival (VR/AR) technology today, the need to digitize physical spaces into 3D data is becoming increasingly urgent across many fields in Vietnam and around the world. From a professional perspective in the field of Computer Vision and Graphics, traditional 3D reconstruction methods (such as photometry or laser scanning) still face significant barriers in terms of equipment costs and processing complexity, and the lack of automation tools is extremely common. Specifically in the case of video reconstruction, current solutions often

struggle to balance speed and quality, or models reconstructed using older algorithms often suffer from noise, loss of detail, and require significant manual intervention. This paper focuses on surveying and evaluating technological solutions for reconstructing 3D spatial structures from single-image video data (monocular video) using modern representation methods, through an integrated artificial intelligence system (Deep Learning and 3D Gaussian Splatting). Through this, a useful tool will be provided for developers and end users to quickly create high-quality digital replicas, as well as optimize real-time processing capabilities and overcome the limitations of current offline rendering methods.

This research will be presented in five sections. First, an overview of the research history and technological approaches will be provided. Next, the research will describe in detail the process, system architecture, implementation methods, and theory for building the training dataset from video. Finally, the research will briefly describe the experimental process of training the model and the metrics used to evaluate the results. Finally, there is a summary and a section outlining the article's future direction.

1.2 Product overview

The history of 3D video rendering systems has gone through several technological phases. In the past decade, commercial solutions such as Agisoft Metashape (2010) and RealityCapture (2016) dominated the market based on Structure-from-Motion (SfM) algorithms, offering high accuracy but requiring significant processing time. The emergence of Deep Learning ushered in a new era with Polycam (2020) and Luma AI (2022), pioneering applications bringing Neural Radiance Fields (NeRF) technology to mobile and cloud computing platforms. Most recently, the introduction of 3D Gaussian Splatting (2023) and Spatial Intelligence models from World Labs (2025) is reshaping the field, enabling real-time scene rendering with superior photometric quality.

Luma AI is considered a pioneer in bringing NeRF technology to the general public. This application allows users to record video around an object or space using their smartphone, and then the AI processes that data cloud to create a 3D photorealistic replica. Luma's unique feature is its ability to handle complex light reflections and fine details that traditional photogrammetry often misses. Recently, they have integrated

Gaussian Splatting technology, enabling real-time rendering at extremely high speeds. Following this was the introduction of NVIDIA Instant NeRF. Although initially a research project, Instant NeRF quickly became a standard tool in the industry. Previously, creating 3D models from AI took hours to train, but NVIDIA used multi-resolution hash encoding techniques combined with the power of RTX GPUs to reduce this time to just seconds to minutes. This product demonstrates the ability to build 3D spaces from a group of 2D images or short videos at near-instantaneous speed.

Another product is Polycam from Polycam Inc. Polycam is one of the most popular 3D scanning applications on mobile platforms (iOS/Android). Initially, Polycam focused on the LiDAR sensor on iPhones, but later strongly integrated AI to reconstruct 3D from photos (Photo Mode). Currently, Polycam allows users to upload videos to create 3D models using Gaussian splatting technology. This is a prime example of a commercial product combining mobile hardware and cloud-based AI algorithms to serve architects, game designers, and VFX artists. RealityScan is a direct bridge between the real world and Unreal Engine. Users shoot/capture objects with their phones, and Epic Games' cloud system processes the data. The product's strength lies in its highly automated workflow: from raw video to optimized, high-quality 3D mesh models. AI is used to filter out image blur, balance brightness, and automatically stitch pixels together, making the 3D construction process much more accurate than previous manual methods.

The most significant breakthrough in solving the 3D rendering performance problem recently

is the introduction of the 3D Gaussian Splatting method (Kerbl et al., SIGGRAPH 2023). Addressing the drawback of the high computational cost of traditional ray-marching methods in NeRF, this research proposes a completely new approach: representing scenes using a collection of millions of unstructured 3D Gaussian points. Each Gaussian point carries parameters for position, covariance (determining elliptical shape), opacity, and Spherical Harmonics coefficients (to describe color dependence depending on viewing angle). The breakthrough of this method lies in the combination of a process for optimizing the differential radiation field with a custom Tile-

based Rasterizer. This mechanism allows for the projection and alpha-blending of 3D points onto a 2D image plane at extremely high speed, while maintaining backpropagation for model training. Experimental results show that 3D Gaussian Splatting not only achieves image quality comparable to or superior to the best NeRF methods (such as Mip-NeRF 360) but also provides real-time rendering (achieving ≥ 100 FPS at 1080p resolution) and reduces training time from tens of hours to just a few minutes. This is currently considered the most advanced (State-of-the-Art) technique forming the foundation for current real-time 3D rendering systems.

Table 1. Comparison table of product criteria

Criteria	LumaAi	NVIDIA Instant network	Polycam	Reality scan
Core technology	Gaussian Splatting & NeRF (Neural Radiance Fields).	Multi-resolution Hash Encoding combined with NeRF.	LiDAR & Photogrammetry (traditional) + Gaussian Splatting support (new).	AI-assisted photogrammetry (based on Capturing Reality algorithms).
Operating platform	iOS, Android, Web (Cloud Processing).	Windows/Linux (Runs locally).	iOS, Android, Web	iOS, Android
Platform price	Freemium (Basic free, paid to export high-quality files or use APIs).	Open source (Free) for non-commercial/research purposes.	Freemium (Limited free offer, Pro plan ~\$15-18/month).	Free (Integrated into the Unreal Engine/Sketchfab ecosystem).
Strengths	- Extremely high photorealistic quality. - Capable of processing reflective/transparent surfaces (glass, water). - Real-time rendering speeds are very fast	- Fastest training speed on the market. - Allows for deep intervention into technical specifications (good for research). - High security due to offline operation.	- Versatile: Features both LiDAR and Photo mode. - The built-in editing, cropping, and measuring tools are very powerful.	- Process automation is extremely beneficial for game developers. - Outputs a very clean mesh model with few errors.

Criteria	LumaAi	NVIDIA Instant network	Polycam	Reality scan
	thanks to Gaussian Splatting.			- Direct connection with Unreal Engine.
		- Requires extremely high-end hardware (NVIDIA RTX GPU).		- Processing time in the cloud can be long.
	- Completely dependent on cloud servers (requires internet).	- The interface is difficult to use (command-line or simple GUI), not suitable for the average user.	- The quality of glossy/reflective surface reproduction is inferior to Luma AI.	- Limited customization options for input parameters.
Weakness	- It's difficult to manually edit the geometry after it's created.	- It's difficult to export files for use in other 3D software.	- Premium features are locked behind a paid "firewall".	- Primarily focuses on creating static objects (asset), not strong in recreating large scenes.

1.3 Source code for creating 3d spaces from video

The first method, Neural Radiance Fields (NeRF), with source code at: <https://github.com/bmild/nerf>, approaches the problem through implicit representation. Instead of storing data as a polygonal mesh or a finite point cloud, NeRF encodes all geometric and color information of the space into the weights of a multi-layer neural network (MLP). NeRF's display mechanism is based on volume ray marching, allowing the synthesis of volume density at any coordinate to create frames with extremely high photo-realistic fidelity, particularly effective in reproducing complex reflective surfaces. However, the biggest drawback of NeRF is its enormous computational cost; the training and display process is often slow, making it difficult to meet real-time demands on devices with limited resources.

To overcome the speed limitations of NeRF, 3D Gaussian Splatting (3DGS) was introduced as a breakthrough with explicit representation. The space in this method is composed of a collection of millions of discrete 3D Gaussian spheres, each point carrying information about position, opacity, and color properties. The advantage of 3DGS lies in replacing the expensive ray tracing technique with a rasterization mechanism, allowing for the direct projection of point clouds onto the image plane with superior processing speed. As a result, 3DGS not only shortens training time from tens of hours to just a few minutes but also achieves real-time rendering at over 100 FPS with image quality equivalent to, or even sharper than, NeRF. This is also why 3DGS is gradually becoming the new standard for modern 3D rendering systems. The source link for 3D Gaussian Splatting is: <https://github.com/graphdeco-inria/gaussian-splatting>

Completely different from the offline processing of the two methods above, the Neural SLAM group of methods (<https://github.com/devendrachaplot/Neural-SLAM>) such as NICE-SLAM or Co-SLAM focus on solving the problem of online stream processing. These systems are designed to perform two tasks in parallel: estimating camera trajectories (Tracking) and updating 3D maps (Mapping) while video is being recorded. Architecturally, Neural SLAMs often use a hybrid model combining feature grids and small neural networks to ensure fast local update speeds without the phenomenon of "forgetting" old data. Although it sacrifices some image detail compared to 3DGS or NeRF, Neural SLAM plays an irreplaceable role in autonomous robotics and augmented reality (AR) applications requiring instantaneous response in unknown environments.

Additionally, there is the NeuralRecon method (<https://github.com/zju3dv/NeuralRecon>), which has demonstrated significantly superior reconstruction quality compared to other machine learning-based methods in both 3D geometry and 2D depth indices (Jiaming Sun et al., 2021). ViPE (<https://github.com/nv-tlabs/vipe>) achieves significant improvements in camera pose estimation, with performance 18% and 50% better than uncalibrated baseline methods on TUM and KITTI sequences, respectively (Jiahui Huang et al., 2025). The system maintains accuracy at a real-world scale through an indexed depth prediction module.

The methods for validating real-world data vary depending on the application domain. Some studies use depth maps generated from reconstructed grids, while others use actual

distances measured by divers or reference scans from advanced acquisition setups. For camera posture estimation, established standards such as TUM RGB-D and KITTI provide real-world trajectory data. In short, the choice of method depends heavily on the application's objectives. If the highest image quality and fast display speed for the end user are prioritized, 3D Gaussian Splatting is the optimal solution. Conversely, if the application requires the ability to self-locate and expand the map in real time as the camera moves, a Neural SLAM architecture would be a more suitable approach.

1.4 Training data system

In 3D spatial reconstruction studies based on deep learning, the quality and diversity of input data are crucial to the generalization capability and accuracy of the model. The current research community primarily relies on benchmark datasets classified according to specific environmental characteristics and evaluation objectives. For problems focusing on interior reconstruction and localization (In-door Reconstruction & SLAM), ScanNet (Dai et al., 2017) and TUM RGB-D are considered the most important reference standards. ScanNet provides a massive dataset with over 1,500 scans of real-world room spaces, including complete information on surface geometry, semantic segmentation, and camera trajectories, which is invaluable for training neural networks to complete hidden structures. Conversely, TUM RGB-D focuses on handheld video sequences with high motion blur, serving as a measure of robustness for tracking algorithms in SLAM systems.

Furthermore, the explosion of neural visualization methods such as NeRF and 3D

Gaussian Splatting has placed new demands on image resolution and the complexity of ambient lighting. Responding to this need, the Mip-NeRF 360 dataset (Barron et al., 2022) provides realistic footage (both indoor and outdoor) with camera trajectories surrounding objects, challenging algorithms in handling boundless backgrounds and reflective surfaces. In parallel, the Tanks and Temples dataset (Knapitsch et al., 2017) is widely used to assess geometric accuracy thanks to ground truth data collected from industrial laser scanners, allowing for the measurement of specific errors of reconstructed 3D models compared to reality. Recently, research trends have shifted strongly towards building Foundation Models capable of understanding general spatial context. A prime example of this trend is the emergence of the DL3DV-10K dataset (Ling et al., 2024), containing over 10,000 diverse landscape videos from around the world. Leveraging large-scale datasets like DL3DV-10K allows AI systems to learn rich a priori features in physics and lighting, significantly improving reconstruction quality even in cases of sparse or insufficient input data.

The relationship between training data and real-world performance reveals important patterns. Systems trained entirely on synthetic data (such as multi-perspective CNNs and Geo4D) have succeeded in generalizing to real-world scenarios, but with specific limitations. Multi-perspective CNNs struggle with complex backgrounds, while Geo4D struggles with significant focal length changes or extreme camera movement. Conversely, systems trained on large-scale real-world datasets (96 million frames for ViPE and 17 diverse datasets for SpatialTrackerV2) have demonstrated superior

performance under varying conditions. This model demonstrates that synthetic training provides a solid foundation for geometric inference but requires fine-tuning for specific domains or architectural features (such as the integration of classical SLAM with learned components of ViPE) to handle the full complexity of real-world video. The 18%/50% improvement ViPE achieved over the underlying TUM/KITTI systems demonstrates the value of hybrid approaches combining learned knowledge with geometric constraints.

2. RESEARCH METHODS

The shift from traditional geometric methods to deep learning models has shaped fundamental technological approaches, beginning with the advent of Neural Radiance Fields (NeRF). Introduced by Mildenhall et al., NeRF completely revolutionized how computers represent space through the concept of Implicit Representation (Mildenhall et al., 2020). Instead of storing data as polygonal meshes or point clouds, NeRF encodes all geometric and color information into the weights of a multi-layer neural network (MLP), taking 5-dimensional spatial coordinates as input and predicting the corresponding volume density and RGB color (Mildenhall et al., 2020). While achieving high photometric quality with PSNR values often exceeding 30 dB on standard datasets, the biggest limitation of NeRF is its speed; The training process can take up to 18-20 hours, and the display speed only reaches about 0.06 - 5 FPS on high-end GPUs, making real-time applications impossible (Kerbl et al., 2023).

Although NeRF achieves high photometric quality, its biggest limitation is the slow training and display speed. To address this problem, a series of optimized variants have emerged.

Typical examples include Mip-NeRF and Mip-NeRF 360, which use cone casting to eliminate aliasing and accurately reconstruct infinite spaces (Barron et al., 2022). Another significant advancement is Instant-NGP, which uses multi-resolution hash encoding to shorten training time from hours to seconds (Müller et al., 2022). Furthermore, hybrid methods such as TensorRF or K-Planes have proposed tensor decomposition to minimize memory overhead while maintaining high accuracy (Chen et al., 2022; Fridovich-Keil et al., 2023).

However, to thoroughly address the real-time rendering problem, a new breakthrough emerged in 2023 with 3D Gaussian Splatting (3DGS) technology. Unlike NeRF, this method returns to Explicit Representation, composed of millions of discrete 3D Gaussian spheres. The core advantage of 3DGS lies in its custom rasterizer, allowing for the direct projection of point clouds at superior speeds. Experiments have shown that 3DGS shortens training time to just 2-7 minutes (hundreds of times faster than NeRF) and achieves real-time display speeds of up to 150+ FPS at 1080p resolution, while maintaining image quality comparable to or higher than NeRF (PSNR ranging from 31-33 dB) (Kerbl et al., 2023). Recently, optimized variants like Minimal 3DGS have even reached 612 FPS on RTX 4090 GPUs thanks to efficient data compression algorithms (Li et al., 2025).

Alongside static scene reconstruction methods, Neural SLAM research focuses on solving the problem of online positioning and mapping for robots. The combination of SLAM and 3DGS has created a new wave of research in the 2024-2025 period. SplatAM is one of the pioneering systems that successfully applied 3DGS to the

RGB-D SLAM problem, achieving a positioning error of less than 1 cm (Keetha et al., 2024). Following that, MonoGS demonstrated the ability to perform SLAM with only a monocular camera without the need for a depth sensor (Matsuki et al., 2024). Notably, ViPE research improved the camera pose estimation performance by up to 18% compared to basic methods (Huang et al., 2025). To handle dynamic environments with moving objects, WildGS-SLAM proposes an uncertainty-aware mapping mechanism, which more effectively eliminates disturbances than traditional methods (Zheng et al., 2025).

Furthermore, the trend of integrating semantics into 3D mapping is also gaining attention, with models like LangSplat enabling robots to not only "see" images but also "understand" objects in space through Gaussian semantic fields (Qin et al., 2024). Finally, the support from Vision Foundation Models cannot be overlooked. Depth estimation models such as Depth Anything V2 are playing a crucial role in providing geometric priors for 3DGS optimization, especially in areas lacking surface texture (Yang et al., 2024). Simultaneously, generalized reconstruction models such as DUST3R or LGM are opening new avenues for creating instant 3D content from a few single images (Wang et al., 2024; Tang et al., 2024).

Especially for single-image video input (Monocular Video), algorithms need to integrate advanced estimation modules. NeuralRecon has proposed a method for consistently reconstructing 3D surfaces from real-time video (Sun et al., 2021). Meanwhile, new research in 2025 focuses on improving camera pose accuracy; for example, ViPE, which improves pose

estimation performance by up to 18% on TUM datasets and 50% on KITTI datasets compared to basic methods, with an average processing speed of 3-5 FPS on a single GPU (Huang et al., 2025). Similarly, Geo4D has demonstrated superior 4D reconstruction capabilities, reducing the absolute relative depth error (Abs Rel) to 0.205 on the Sintel dataset, significantly lower than the 0.270 of the DepthCrafter model, while achieving an accuracy ($\delta < 1.25$) of up to 73.5% (Jiang et al., 2025).

In addition, support from Foundation Models such as Depth Anything V2 also plays a crucial role. In in-the-wild tests, Depth Anything V2 achieved an average absolute error (MAE) of only 0.454m, far surpassing older models like ZoeDepth (MAE 3.087m), providing an accurate depth map as a "prior" input to help 3DGS converge faster and more accurately in areas lacking surface texture (Yang et al., 2024). While standard 3D Gaussian Splatting (3DGS) methods have excelled at reconstructing static scenes, practical video applications often involve complex moving objects (such as pedestrians, expressive faces, or swaying leaves). To address this challenge, recent studies have proposed 4D Gaussian Splatting models, incorporating an additional time dimension into the representation space. One of the pioneering methods is the use of space-time hash grids (Hexplane) to encode the distortion of Gaussian points in real time. This technique allows for the separation of static and dynamic geometric information, helping to maintain impressive rendering speeds of up to 70 FPS at 800x800 resolution, while shortening training time to under 1 hour for a short video, significantly better than previous dynamic NeRF

methods that often took days to converge (Wu et al., 2024).

Furthermore, the Deformable 3D Gaussian model approaches the problem by using a Deformation Network to adjust the position and blur of Gaussian points frame by frame. This method achieves an average PSNR of 31.5 dB on the D-NeRF dataset, demonstrating the ability to reproduce non-rigid structures with high fidelity without complex physical constraints (Yang et al., 2024). Beyond image reconstruction, modern research trends also focus on "understanding" space through Semantic Integration. Instead of just storing color (RGB), models like LangSplat have proposed embedding language features from the CLIP model into each Gaussian point. This allows users to search for or segment objects in 3D space using natural text (e.g., "find the red chair"). LangSplat uses a Variational Autoencoder to compress the linguistic feature space, increasing query speeds by up to 144 times compared to the NeRF-based Language Embedded Radiance Fields (LERF) method, while maintaining a segmentation accuracy (mIoU) of 84.3% on the test dataset (Qin et al., 2024).

The combination of 4D animation and semantic understanding is reshaping the future of 3D video rendering systems. It not only creates "realistic-looking" digital replicas but also "meaningful" and "living" digital entities, opening up enormous potential for applications in Robotics (robots that understand commands to find objects) and the Metaverse (real-time interaction with virtual environments). Besides per-scene optimization methods, a new trend emerging in 2024 is the introduction of Generalizable Reconstruction Models, aiming to

eliminate reliance on traditional camera position estimation algorithms like COLMAP, which are time-consuming. A prime example is DUST3R (Dense Unconstrained Stereo 3D Reconstruction), a new network architecture that approaches 3D reconstruction using pixel-to-pixel regression without requiring input information on camera posture or internal parameters. DUST3R has demonstrated robust processing capabilities for in-the-wild, hand-cranked videos with superior global stability compared to traditional stereo pairing methods (Wang et al., 2024). Furthermore, to address the challenge of moving objects in real-world video – something static 3DGS haven't handled well – 4D Gaussian Splatting variants have been proposed. Studies such as Deformable 3DGS incorporate an additional deformation field parameterized by a small neural network or hexplane, allowing Gaussian points to move and change shape over time t . This mechanism helps the system maintain real-time rendering speed while accurately reproducing complex nonlinear movements of the subject in video (Yang et al., 2024).

Finally, support from Vision Foundation Models (VSMs) is playing a crucial role in improving the input quality for 3D reconstruction algorithms from single-video. Notably, Depth Anything, a single-eye depth estimation model trained on massive datasets, provides depth maps with significantly higher edge detail and robustness than previous generations such as MiDaS. Integrating Depth Anything as a geometric prior signal helps guide the optimization process of 3D Gaussian Splatting, significantly reducing geometric noise in areas far from the camera or lacking surface texture (Yang et al., 2024).

3. RESULTS AND DISCUSSION

To quantify the effectiveness and practical applicability of automated 3D spatial reconstruction systems from video, the study uses a multidimensional evaluation system encompassing three core aspects: photometric quality, geometric accuracy, and computational performance. For the aspect of image display quality (Novel View Synthesis), three widely applied standard indices are PSNR, SSIM, and LPIPS. Of these, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) measure signal-to-noise discrepancies and structural similarity at the pixel level (Wang et al., 2004), reflecting the ability to preserve color information and detailed edges of objects. However, for modern deep learning models, LPIPS (Learned Perceptual Image Patch Similarity) is considered the most important metric; By using neural networks to simulate human visual perception, LPIPS accurately assesses the naturalness and realism of reconstructed images more effectively than purely mathematical formulas.

Regarding spatial structure reconstruction, when 3D reference data (Ground Truth) is available, geometric accuracy is determined through Chamfer Distance and F-Score (Knapitsch et al., 2017). Chamfer Distance measures the shortest average error between the reconstructed point cloud and reality, while F-Score helps assess the balance between accuracy and completeness of the surface, eliminating extraneous noise. For integrated localization and mapping systems (SLAMs), the ATE (Absolute Trajectory Error) index is mandatory to verify the deviation between the estimated camera trajectory and the actual movement trajectory.

Finally, the feasibility of the system in a real-world environment is measured through performance metrics. These parameters include Frame Rate (FPS) to determine real-time operational capability, Training Time to assess the cost of learning a new scene, and hardware resource consumption (VRAM). An optimal system today, such as variations of 3D Gaussian Splatting, is often defined as achieving a balance between display quality (low LPIPS) and high processing speed ($\text{FPS} > 30$), ensuring a smooth interactive experience for the end user (Kerbl et al., 2023).

Besides the basic metrics mentioned above, the specifics of the spatial reconstruction problem from video require special attention to the accuracy of depth maps (Depth Estimation Metrics). In architectures such as NeuralRecon (Sun et al., 2021) or SplatAM (Keetha et al., 2024), the quality of 3D geometry depends directly on the ability to estimate the distance from the camera to the object. Commonly used metrics include Abs Rel (Absolute Relative Error), which measures the relative error between predicted depth and actual depth, and RMSE (Root Mean Square Error), which helps detect unusually large deviations. Another important metric is $\Delta < 1.25$, which indicates the percentage of pixels with depth errors within acceptable limits; a higher percentage indicates a more reliable system for spatial localization. Furthermore, since the input data is a continuous sequence of images, temporal consistency becomes a mandatory criterion for evaluating the quality of the experience. A good 3D model not only needs to be sharp in each static frame but also must ensure that there is no flickering or jittering when the camera moves. To measure this

factor, research can use the Optical Flow Warping Error (Lai et al., 2021) or video variations of LPIPS to assess the stability of reconstructed details over successive time steps (Jiahui Huang et al., 2025).

In addition to quantitative measures, qualitative evaluation through visual perception plays an indispensable supplementary role, especially when processing real-world data without a standard reference (in-the-wild). Studies often conduct User Studies to collect Mean Opinion Scores (MOS). In these studies, participants evaluate results based on realism, level of detail, and the presence of artifacts such as floaters or areas of tearing (Barron et al., 2022). Combining rigorous mathematical metrics with subjective human evaluation provides the most comprehensive view of the system's real-world performance.

Finally, the model's generalization and robustness should also be considered. The system needs to be evaluated on its ability to handle challenging situations such as sudden changes in lighting conditions, moving objects in a static scene, or highly reflective surfaces (mirrors, glass). A superior AI system is one that maintains stable PSNR and SSIM values even in the face of these environmental disturbances, rather than performing well only on ideal laboratory datasets (Martin-Brualla et al., 2021).

4. CONCLUSION

This overview of research is conducted within the specific context of the fields of Computer Vision and Computer Graphics, where the need to digitize the real world and the limitations of computing resources pose a significant challenge to building real-time spatial reconstruction systems. The difficulty in implementing these

systems stems from three main factors: the complexity of handling diverse object surfaces (reflective, transparent) in a natural context, the excessive time cost of traditional photometric methods or early NeRF models, and the instability of manually recorded video data (shaking, blurring) leading to cumulative errors in the positioning process. In the field of 3D reconstruction, a well-built initial model can quickly become outdated or inconsistent when the camera moves to a new viewing angle or when lighting conditions change. Continuously updating 3D maps based on video streams requires a delicate balance between retaining prior knowledge and acquiring new data. This complexity makes maintaining a stable system difficult in the face of the "catastrophic forgetting" phenomenon often seen in deep neural networks. Therefore, these situations pose challenges for the field of artificial intelligence, particularly the application of explicit representation methods and online learning mechanisms to 3D modeling problems. This paper presents an overview of artificial intelligence methods that enable the automatic reconstruction of 3D spatial structures from monocular video (monocular video), through the integration of 3D Gaussian Splatting techniques and advanced depth estimation algorithms. The system is an attempt to apply high technology to simplify the digital content creation process, a technological leap forward compared to expensive laser scanning methods or manual photogrammetry, in order to keep pace with the rapidly developing "Spatial Intelligence" trend worldwide. AI-based 3D rendering technology integrated into mobile devices is an inevitable trend in the digital content industry and the future

of the Metaverse. The research results have demonstrated the feasibility of creating photogrammetric-quality digital replicas in real time, opening up prospects for widespread application for non-expert users. Although this system performs well in recreating static scenes, further support and in-depth research are needed to effectively handle complex moving objects and scale up to larger urban spaces.

REFERENCES

- Ali, S., Bailey, A., Ash, S., Haghighat, M., Allan, P., Ambrose, T., ... & Braden, B. (2021). A pilot study on automatic three-dimensional quantification of Barrett's esophagus for risk stratification and therapy monitoring. *Gastroenterology*, 161(3), 865-878.
- Chan, E. R., Nagano, K., Chan, M. A., Bergman, A. W., Park, J. J., Levy, A., ... & Weitzstein, G. (2023). Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4217-4229). <https://nvlabs.github.io/genvs/media/genvs.pdf>
- Chen, X., Chen, Y., Xiu, Y., Geiger, A., & Chen, A. (2025). Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645. <https://arxiv.org/pdf/2509.26645>
- Choi, C., Kim, J., Cha, G., Kim, M., Wee, D., & Kim, Y. M. (2025). Humans as a calibration pattern: Dynamic 3d scene reconstruction from unsynchronized and uncalibrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6598-6608). <https://arxiv.org/pdf/2412.19089>

- Condorelli, F., Rinaudo, F., Salvatore, F., & Tagliaventi, S. (2021). A comparison between 3D reconstruction using nerf neural networks and mvs algorithms on cultural heritage images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 565-570.
- Cujó Blasco, J., Bemposta Rosende, S., & Sánchez-Soriano, J. (2023). Automatic real-time creation of three-dimensional (3d) representations of objects, buildings, or scenarios using drones and artificial intelligence techniques. *Drones*, 7(8), 516. <https://www.mdpi.com/2504-446X/7/8/516>
- Doulamis, A. (2018). Automatic 3D reconstruction from unstructured videos combining video summarization and structure from motion. *Frontiers in ICT*, 5, 29. <https://www.frontiersin.org/journals/ict/articles/10.3389/fict.2018.00029/full>
- Dryanovski, I., Klingensmith, M., Srinivasa, S. S., & Xiao, J. (2017). *Large-scale, real-time 3D scene reconstruction on a mobile device. Autonomous Robots*, 41(6), 1423-1445 https://drive.google.com/file/d/1wOmSTi0x2YDq9TVz92uaOsDEHcgHrw_z/view
- Gao, H., Mao, W., & Liu, M. (2023). Visfusion: Visibility-aware online 3d scene reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 17317-17326).
- Hong, Z., & Yue, C. P. (2023). Cross-dimensional refined learning for real-time 3D visual perception from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2169-2178). <https://arxiv.org/pdf/2303.09248>
- Huang, Y. H., He, Y., Yuan, Y. J., Lai, Y. K., & Gao, L. (2022). Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18342-18352). https://openaccess.thecvf.com/content/CVPR2022/papers/Huang_StylizedNeRF_Consistent_3D_Scene_Stylization_As_Stylized_NeRF_via_2D-3D_CVPR_2022_paper.pdf
- Keetha, N., Karhade, J., Jatavallabhula, K. M., Yang, G., Scherer, S., Ramanan, D., & Luiten, J. (2024). Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21357-21366). <https://arxiv.org/pdf/2312.02126>
- Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139-1. arxiv.org/pdf/2308.04079
- Kosiorrek, A. R., Strathmann, H., Zoran, D., Moreno, P., Schneider, R., Mokrá, S., & Rezende, D. J. (2021, July). Nerf-vae: A geometry aware 3d scene generative model. In *International conference on machine learning* (pp. 5742-5752). PMLR. <https://proceedings.mlr.press/v139/kosiorrek21a/kosiorrek21a.pdf>
- Lai, Z., Liu, S., Efros, A. A., & Wang, X. (2021). Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9730-9740). <https://arxiv.org/pdf/2110.02951>

- Liu, S., Yang, M., Xing, T., & Yang, R. (2025). A Survey of 3D Reconstruction: The Evolution from Multi-View Geometry to NeRF and 3DGS. *Sensors*, 25(18), 5748. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12473764/>
- Lin, C. H., Lv, Z., Wu, S., Xu, Z., Nguyen-Phuoc, T., Tseng, H. Y., ... & Li, Z. (2025). DGS-LRM: Real-Time Deformable 3D Gaussian Reconstruction From Monocular Videos. *arXiv preprint arXiv:2506.09997*. <https://arxiv.org/pdf/2506.09997>
- Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., ... & Bera, A. (2024). DL3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22160-22169). https://openaccess.thecvf.com/content/CVPR2024/papers/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.pdf
- Liu, Y., Dong, S., Wang, S., Yin, Y., Yang, Y., Fan, Q., & Chen, B. (2025). Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 16651-16662). https://openaccess.thecvf.com/content/CVPR2025/papers/Liu_SLAM3R_Real-Time_Dense_Scene_Reconstruction_from_Monocular_RGB_Videos_CVPR_2025_paper.pdf
- Maccarone, A., Drummond, K., McCarthy, A., Steinlechner, U. K., Tachella, J., Garcia, D. A., ... & Buller, G. S. (2023). Submerged single-photon LiDAR imaging sensor used for real-time 3D scene reconstruction in scattering underwater environments. *Optics Express*, 31(10), 16690-16708
- Park, C., Ko, H., & Kim, T. (2003). NAVER: Networked and Augmented Virtual Environment Architecture; design and implementation of VR framework for Gyeongju VR Theater. *Computers & Graphics*, 27(2), 223-230.
- Schöning, J., Behrens, T., Faion, P., Kheiri, P., Heidemann, G., & Krumnack, U. (2017). Structure from motion by artificial neural networks. In Sharma, P. & Bianchi, F.M. (B.t.p.). In *Image Analysis (Proceedings of the 20th Scandinavian Conference on Image Analysis, SCIA 2017, tr. 264–275)*
- Sung, B. Y., & Lin, C. H. (2017). A fast 3D scene reconstructing method using continuous video. *EURASIP Journal on Image and Video Processing*, 2017(1), 18. <https://link.springer.com/content/pdf/10.1186/s13640-017-0168-3.pdf>
- Sun, J., Xie, Y., Chen, L., Zhou, X., & Bao, H. (2021). Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15598-15607). <https://arxiv.org/pdf/2104.00681>
- Stotko, P., Krumpal, S., Hullin, M. B., Weinmann, M., & Klein, R. (2019). Slamcast: Large-scale, real-time 3d reconstruction and streaming for immersive multi-client live telepresence. *IEEE transactions on visualization and computer graphics*, 25(5), 2102-2112. <https://arxiv.org/pdf/1805.03709>
- Teed, Z., & Deng, J. (2021). Droid-slam: Deep visual slam for monocular, stereo, and rgb-d

- cameras. *Advances in neural information processing systems*, 34, 16558-16569. <https://arxiv.org/pdf/2108.10869>
- Tu, T., Li, M. F., Lin, C. H., Cheng, Y. C., Sun, M., & Yang, M. H. (2025, February). Dreamo: Articulated 3d reconstruction from a single casual video. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 2269-2279). IEEE. <https://arxiv.org/pdf/2312.02617>
- Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., & Fragkiadaki, K. (2017). *Sfm-net: Learning of structure and motion from video*. arXiv preprint arXiv:1704.07804. <https://arxiv.org/pdf/1704.07804>
- Wang, H., Wang, J., & Agapito, L. (2023). Co-slam: Joint coordinate and sparse para-metric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13293-13302). <https://arxiv.org/pdf/2304.14377>
- Würlmlin, S., Lamboray, E., & Gross, M. (2004). 3D video fragments: Dynamic point samples for real-time free-viewpoint video. *Computers & Graphics*, 28(1), 3-14.
- Wu, D., Yan, Z., & Zha, H. (2024). PanoRecon: Real-time panoptic 3D reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 21507-21518).
- Wu, Z., Yan, Q., Yi, X., Wang, L., & Liao, R. (2025). *StreamSplat: Towards Online Dynamic 3D Reconstruction from Uncalibrated Video Streams*. arXiv preprint arXiv:2506.08862. <https://arxiv.org/pdf/2506.08862>
- Xie, Y., Gadelha, M., Yang, F., Zhou, X., & Jiang, H. (2022). Planarrecon: Real-time 3d plane detection and reconstruction from posed monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6219-6228). <https://arxiv.org/pdf/2206.07710> https://openaccess.thecvf.com/content/CVPR2024/papers/Wu_PanoRecon_Real-Time_Panoptic_3D_Reconstruction_from_Monocular_Video_CVPR_2024_paper.pdf
- Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., & Jin, Y. (2024, October). Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 132-142). Cham: Springer Nature Switzerland. <https://arxiv.org/pdf/2405.17835>
- Zinat Matin, H. N., & Delle Monache, M. L. (2023). *Near Collision and Controllability Analysis of Nonlinear Optimal Velocity Follow-the-Leader Dynamical Model In Traffic Flow*. arXiv e-prints, arXiv:2309. <https://arxiv.org/pdf/2309.11704>
- Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., ... & Pollefeys, M. (2022). Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12786-12796). <https://arxiv.org/pdf/2112.12130>
- Zhao, C., Sun, L., & Stolkin, R. (2017, July). A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. In *2017 18th international conference on advanced*

robotics (ICAR) (pp. 75-82). IEEE.

<https://arxiv.org/pdf/1703.04699>