Dynamic 3D Understanding Research Survey

Anonymous CVPR submission

Paper ID 1

Abstract

Dynamic 3D Understanding is the research focuses on the perception, interpretation, and prediction of three-dimensional environments in motion, involving objects and scenes that change over time.It is critical for applications such as autonomous driving, robotics, augmented reality (AR), and human motion analysis. While significant progress has been made, especially with deep learning advancements in 3D point clouds and neural scene representations, dynamic 3D understanding remains a challenging problem due to its need for real-time processing, handling of non-rigid deformations, and multi-modal sensor fusion. Recent developments include improvements in object tracking, deformable surface modeling, neural rendering techniques, innovative network architectures and so on. These approaches have enabled better real-time scene understanding and smoother handling of dynamic deformations. Applications span across multiple domains, but challenges remain in scaling to complex environments, improving real-time performance, and achieving long-term prediction capabilities. In the future,advancements in real-time processing, self-supervised learning, and enhanced sensor fusion promise to unlock further potential in this field.

1. Introduction

Dynamic 3D understanding, which refers to the ability of systems to perceive and interpret three-dimensional scenes that evolve over time, has gained increasing importance across various fields. With the rise of intelligent systems capable of real-time environmental interaction, the ability to understand dynamic scenes becomes crucial for technologies like autonomous vehicles, robotic navigation, and augmented reality (AR). These systems depend on accurate 3D information to ensure safe and efficient operation, especially when objects or agents in the environment are in motion or undergoing changes over time.

1.1. Background and motivation

Dynamic 3D understanding has become essential in advancing key technologies such as autonomous vehicles, robotic navigation, and augmented reality (AR) [2, 4]. These applications rely on systems that can accurately interpret and respond to complex, evolving environments. Whether it's the motion of surrounding vehicles or the interactions between agents, dynamic 3D information enables these systems to function effectively in real-time, ensuring safety and operational efficiency.

Autonomous vehicles, for instance, must be able to interpret the motion of surrounding vehicles, pedestrians, and cyclists, all of which introduce significant complexity into the vehicle's perception system[1]. Even slight misinterpretations of this dynamic data can lead to safety hazards. Similarly, in AR applications, users interact with both virtual objects and the real world simultaneously. The seamless integration of these elements requires an accurate understanding of how the real-world environment is changing in 3D space, such as how objects are moving or how lighting conditions are shifting. In such cases, dynamic 3D understanding is fundamental to ensuring that the virtual objects behave naturally in relation to the real environment.

The motivation behind advancing research in dynamic 3D understanding is driven by the growing need for systems that are not just reactive but also predictive. Traditional methods, while successful in static 3D scene reconstruction, often fall short in scenarios where objects or agents exhibit non-rigid movements, such as a human walking or a hand grasping an object. As we look towards the future of robotics, virtual environments, and intelligent transportation systems, the ability to model, predict, and analyze dynamic 3D data will become even more critical. Moreover, the increasing availability of real-time 3D sensing technologies, such as LiDAR and depth cameras, underscores the need for algorithms that can fully utilize these rich data streams for more accurate dynamic scene interpretation.

1.2. Related work

Dynamic 3D understanding is an evolving field rooted in early research on visual surveillance, motion detection, and

behavior analysis. [7]. Early projects such as the Visual Surveillance and Monitoring (VSAM) project, funded by DARPA in 1997, laid the groundwork for automatic video understanding in complex environments like battlefields and urban spaces[3]. Following this, DARPA's Human Identification at a Distance (HID) program [7], launched in 2000, pushed forward multimodal surveillance technologies, aiming to detect, classify, and identify human subjects from a distance, further broadening the scope of dynamic scene understanding[7]. These large-scale projects significantly influenced research in both 2D and 3D dynamic scene analysis.

In terms of specific systems, early visual surveillance technologies were primarily 2D and focused on tracking and analyzing moving objects in video streams. For example, the W4 system used shape analysis and tracking to monitor groups of people, even in the presence of occlusions[5]. The Pfinder system developed by Wren et al. [19] was one of the first systems capable of recovering 3D representations of humans in large rooms, although it was limited to tracking single, non-occluded persons. These systems, while pioneering in their time, lacked the spatial depth required to fully capture the complexities of dynamic 3D environments.

The shift toward dynamic 3D understanding began with advancements in depth sensors and multiple-camera setups. Systems like the VIEWS project at the University of Reading employed 3D models for vehicle tracking, while the system at Carnegie Mellon University (CMU) used multiple connected cameras to monitor large areas, tracking both people and vehicles [7]. These multi-camera approaches allowed for better object tracking and behavior understanding in dynamic, cluttered environments, but still suffered from challenges such as occlusions and the high computational costs of integrating data from multiple views.

Recent advancements have taken this research a step further by integrating deep learning and neural representations into dynamic 3D understanding. Spatio-Temporal Neural Networks (STNs) [8] and 3D Convolutional Neural Networks (3D-CNNs) [18] have been employed to capture both spatial and temporal features in dynamic scenes, significantly improving object recognition and motion prediction capabilities. Meanwhile, Implicit Neural Representations (INR) [17], neural radiance fields (NeRF) [13] and their extensions, such as D-NeRF [15], have demonstrated the ability to model highly detailed 3D scenes over time, overcoming the limitations of traditional geometric approaches.

1.3. Challenge

Dynamic 3D understanding presents several significant challenges. First, accurately capturing and modeling complex, non-rigid object deformations in real-time remains difficult due to the high computational costs and the need for large datasets. Techniques like NeRF [13] and its extensions, while promising, are computationally expensive and often struggle with highly dynamic scenes where rapid changes occur. Second, maintaining temporal consistency in motion sequences and occlusion handling, particularly with multiple interacting objects, is a persistent challenge. Additionally, achieving photorealistic rendering of dynamic environments under varying lighting conditions requires further improvement in neural rendering techniques. Lastly, integrating spatial and temporal information in a unified model is still a key area for research, as balancing resolution and efficiency is critical for real-world applications.

2. Taxonomy of key design decisions and techniques

Dynamic 3D understanding has evolved significantly, with early approaches relying on mathematical and physics-based methods such as level set method [14], optical flow methods [6], and structure from motion (SFM) [12], which provided the foundation for modeling and interpreting 3D motion. These approaches were instrumental in capturing geometric and kinematic information but often faced limitations in handling complex, real-time, and large-scale dynamic environments.

With the advent of deep learning, methods like Implicit Neural Representations (INR)[17], spatio-temporal neural networks (STNs) [8] [20], neural radiance fields (NeRF) [13] have dramatically improved the ability to model spatial and temporal dynamics in 3D scenes. These newer models offer significant advancements in handling non-rigid object deformations, complex motion patterns, and photorealistic rendering, pushing the field toward more accurate, real-time understanding of dynamic environments. By integrating both historical and modern approaches, this section provides a comprehensive view of the key design decisions and techniques that have shaped the progress in dynamic 3D scene analysis.

2.1. Mathematical and Physics-Based Methods for 3D Understanding

Mathematical and physics-based methods for 3D understanding leverage principles from geometry, optics, and physics to infer three-dimensional information from two-dimensional images or sensor data. These approaches typically involve the formulation of mathematical models that describe the relationship between 2D observations and the underlying 3D structure. Common tools used in these methods include differential equations, optimization techniques, and geometric transformations.

Despite their foundational significance, these methods encounter several challenges. A primary limitation is their adaptability to complex dynamic scenes. Many traditional

methods assume that the scene is static and that the shapes and lighting conditions are fixed, whereas real-world scenarios often involve non-rigid deformations of objects and variations in lighting that can lead to misinterpretations. Additionally, these methods are often sensitive to the quality of input data, where noise and occlusion can significantly impact the results. The followings are some methods that play significant roles in the development of dynamic 3D understanding.

Level Set Method (LSM) is a mathematical approach used to track and model evolving interfaces and shapes in various contexts. The core principle involves representing a curve or surface as the zero level set of a higher-dimensional function, typically a scalar function defined over the entire space. This allows for the natural handling of topological changes, such as merging or splitting of shapes, as the curve evolves according to a partial differential equation. LSM is widely applied in image segmentation, fluid dynamics, and computer vision, particularly for modeling object boundaries and dynamic shapes in 3D reconstruction. However, it can be computationally intensive and may require careful tuning of parameters to achieve accurate results in complex dynamic scenes[14].

Optical Flow is a technique used to estimate the motion of objects between consecutive frames in a video sequence. The underlying principle is based on the assumption that the intensity of pixels remains constant over time, allowing the calculation of motion vectors that represent the displacement of pixels across frames. This method is commonly applied in various computer vision tasks, including object tracking, motion analysis, and 3D reconstruction. While optical flow can effectively capture motion in many scenarios, it may struggle in situations with significant occlusion, large displacements, or non-rigid motion, leading to inaccuracies in motion estimation [6].

Structure from Motion (SfM) is a technique used to reconstruct three-dimensional structures from two-dimensional image sequences. The core principle involves estimating the camera positions and the three-dimensional coordinates of points in the scene by analyzing the motion and features across multiple images. This is typically achieved through feature extraction, matching, and triangulation processes. SfM is widely utilized in applications such as 3D modeling, augmented reality, and robotics, where understanding the geometry of a scene is crucial. However, SfM can be sensitive to noise, lighting conditions, and the quality of feature matching, which may lead to incomplete or inaccurate reconstructions, particularly in poorly textured environments. [12]

There are also some techniques that do not directly apply to dynamic 3D understanding. However, they provide essential theoretical foundations and played pivotal roles before the advent of deep learning. Shape from Shading (SFS) [21] is a technique that infers the 3D shape of an object based on the shading in a single 2D image, estimating surface orientation through light interaction analysis, which is particularly useful when depth information is sparse. Active Contours (Snakes) [9] are curves that evolve within an image to minimize an energy function, enabling accurate contour fitting to complex object boundaries; however, they may struggle with noisy images or poorly defined boundaries. Lastly, Surface Reconstruction from Point Clouds [11] involves creating continuous surface representations from discrete 3D points, with methods like Poisson surface reconstruction generating smooth surfaces by utilizing spatial relationships among points, which is critical in 3D modeling but challenged by incomplete data or noise.

2.2. Deep Learning Approaches to 3D Understanding

Deep learning techniques offer significant advantages over traditional mathematical and physics-based methods in the realm of 3D understanding. Unlike traditional methods that rely on hand-crafted features and specific assumptions, deep learning models can automatically learn complex representations from large datasets, allowing for greater adaptability to diverse scenarios. Neural networks, particularly convolutional and spatio-temporal networks, excel at capturing intricate spatial and temporal dependencies, making them well-suited for dynamic 3D understanding. By processing sequential data, they can effectively model the motion and interaction of objects in real-time, improving accuracy in tasks like object detection, tracking, and scene reconstruction. The ability to leverage vast amounts of annotated data and benefit from GPU acceleration has led to substantial advancements in the field, enhancing the robustness and precision of dynamic 3D analyses. The followings are some important models in the realm of dynamic 3D understanding.

Spatio-Temporal Graph Convolutional Network (STGCN) [20]offers a powerful framework for capturing the complex relationships in spatiotemporal data. By integrating graph convolutional networks with temporal convolutions, STGCN effectively models both spatial structures and temporal dynamics, addressing limitations of traditional neural networks that often neglect the spatial topology inherent in the data. This dual modeling capability enhances prediction accuracy. It was originally used for traffic prediction, but later its potential in the realm of 3D understanding such as human action recognition was also discovered [8].

Yan et al. introduce the Spatial Temporal Graph Convolutional Network[8], which innovatively combines graph neural networks with convolutional structures to enhance human action recognition using skeleton data. A key innovation of STGCN is its modeling of human skeletons as dy-

namic graphs, where nodes represent joints and edges signify their connections. This approach enables the model to effectively capture the crucial spatial relationships necessary for accurate action recognition. By utilizing hierarchical learning through spatial-temporal convolutional blocks, STGCN integrates information across both spatial and temporal dimensions, eliminating the reliance on manually crafted features and thereby improving expressiveness and generalization capabilities. The model has shown superior performance on benchmark datasets like Kinetics and NTU-RGBD, leading to significant improvements in action recognition accuracy.

STGCN's architecture is particularly well-suited for dynamic 3D understanding because it adeptly models the temporal evolution of 3D structures while capturing intricate spatial dependencies. Its flexible design allows for adaptation to various tasks in dynamic scene understanding, making it a promising approach for advancing research in this field by leveraging the spatial topology of joints and the temporal dynamics of motion.

Implicit Neural Representations (INRs) use neural networks to define continuous and differentiable signal representations, offering advantages over traditional discrete methods, particularly in capturing complex details and signal derivatives essential for modeling physical phenomena governed by partial differential equations. A significant innovation in INRs is the use of periodic activation functions, leading to the development of Sinusoidal Representation Networks (SIRENs), which excel at representing intricate natural signals. This capability makes INRs particularly suitable for dynamic 3D understanding applications, such as real-time scene rendering and dynamic object recognition, as they effectively model continuous changes in spatial and temporal dimensions, enhancing their performance in relevant tasks [17].

Neural Radiance Fields (NeRF) represents a significant advancement in synthesizing novel views of complex scenes through the optimization of a continuous volumetric scene function, using a sparse set of input views. One of its key innovations is the use of a fully connected neural network that takes a continuous 5D coordinate as input—encompassing spatial location and viewing direction—and outputs both the volume density and view-dependent emitted radiance at that location. This enables NeRF to generate photorealistic images by querying 5D coordinates along camera rays and applying differentiable volume rendering techniques, allowing for effective optimization based solely on images with known camera poses[13].

However, traditional NeRF is limited to static scenes where spatial locations can be consistently queried across different images. The introduction of Dynamic NeRF (D-NeRF) extends this concept to dynamic scenarios by incorporating time as an additional input. This method splits the learning process into two stages: encoding the scene into a canonical space and mapping this representation to a deformed scene at a specific time. By using fully connected networks for both mappings, D-NeRF can effectively render images of objects in motion—both rigid and non-rigid—by controlling the camera view and time variable[15].

The advantages of employing NeRF in dynamic 3D understanding include its ability to accurately capture intricate motion details and maintain photorealism in generated images. This is particularly beneficial for applications in computer vision and robotics, where understanding dynamic environments is crucial for tasks like scene reconstruction, object tracking, and augmented reality. The flexibility and high fidelity of NeRF and its dynamic extensions provide robust tools for advancing the capabilities of dynamic 3D understanding, offering potential applications across various fields such as virtual reality, film production, and interactive simulations.

3. Evaluation Methods

Dynamic 3D understanding encompasses various tasks, each requiring distinct evaluation metrics tailored to specific objectives. Key tasks in this field include action recognition, which employs metrics such as accuracy, precision, recall, and F1 score to assess performance on datasets like Kinetics and NTU RGB+D [10] [16]. Object tracking focuses on tracking moving objects in videos, utilizing metrics like tracking accuracy, intersection over union (IoU), and success rate to gauge effectiveness. Scene reconstruction aims to create 3D representations from multiple viewpoints, often measured by reconstruction error, structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR). Depth estimation involves inferring depth information from 2D images, typically evaluated using absolute and relative depth errors. Lastly, action prediction aims to forecast future actions based on observed behaviors, employing metrics like mean average precision (mAP) and prediction accuracy. Each task's specific metrics provide valuable insights into the model's performance, enabling comprehensive evaluations in dynamic 3D understanding research.

4. Conclusion

The field of dynamic 3D understanding has made notable progress, yet several gaps remain. While advancements in neural representation methods have improved the synthesis and interpretation of dynamic scenes, issues like occlusions, rapid motion, and complex object interactions are still critical. Furthermore, the absence of standardized evaluation metrics across tasks complicates model benchmarking. This underscores the need for more robust methods that can generalize well across diverse scenarios and create a unified framework for assessing dynamic 3D understanding tasks.

References

[1] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019. 1

[2] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments/MIT press*, 1997. 1

[3] Robert T Collins, Alan J Lipton, Takeo Kanade, Hironobu Fujiyoshi, David Duggins, Yanghai Tsin, David Tolliver, Nobuyoshi Enomoto, Osamu Hasegawa, Peter Burt, et al. A system for video surveillance and monitoring. *VSAM final report*, 2000(1-68):1, 2000. 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1

[5] I Harritaoglu, DAVID Harwood, and LS Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(22): 809–830, 2000. 2

[6] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2, 3

[7] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004. 2

[8] Zhen Huang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2122–2130, 2020. 2, 3

[9] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 3

[10] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[11] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 3

[12] David G Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial intelligence*, 31(3): 355–395, 1987. 2, 3

[13] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 4

[14] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79 (1):12–49, 1988. 2, 3

[15] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2, 4

[16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 4

[17] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2, 4

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[19] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):780–785, 1997. 2

[20] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017. 2, 3

[21] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8): 690–706, 1999. 3