A Survey for Dynamic 3D Understanding Methods

Kulbir Singh Ahluwalia
UIUC
{ksa5@illinois.edu}

Abstract

Dynamic 3D understanding is a critical area in computer vision and robotics, encompassing the analysis, reconstruction, and manipulation of scenes that change over time. Applications such as autonomous navigation, augmented reality, and robotic manipulation require accurate modeling of dynamic environments to interact effectively with the real world. Challenges in this domain include handling occlusions, dealing with non-rigid deformations, varying illumination, accumulating and processing point clouds over time, and reconstructing dynamic scenes from monocular images or sparse data. Recent advancements in neural rendering, implicit neural representations, and visual foundation models have opened new avenues for capturing and interpreting dynamic scenes with high fidelity and real-time performance.

1. Introduction

This survey reviews state-of-the-art methods in dynamic 3D understanding, focusing on techniques for analyzing, reconstructing, and manipulating scenes that change over time. It covers methods such as point cloud accumulation and occlusion-aware reconstruction to enhance data density and handle occlusions, As-Rigid-As-Possible (ARAP) surface modeling for natural and intuitive mesh deformations, and real-time tracking of non-rigid scenes using DynamicFusion.

The survey also discusses approaches leveraging Neural Radiance Fields (NeRF) to model dynamic and deformable scenes from monocular images, exemplified by methods like Nerfies and DynIBaR. In robotic manipulation, it highlights Dynamic 3D Descriptor Fields (D3Fields), which integrate visual foundation models with implicit representations for zero-shot generalization in rearrangement tasks, and AdaptiGraph, which introduces material-adaptive graph-based neural dynamics to manipulate deformable objects with unknown physical properties.

By evaluating these methods based on their strengths and weaknesses, the survey identifies key challenges such as handling occlusions, non-rigid deformations, and complex motions that traditional methods struggle with. It proposes future research directions to advance the field of dynamic 3D understanding.

2. Related Work

Dynamic 3D scene analysis has seen significant advancements with the development of methods that address occlusions and dynamic elements. OcclusionFusion [2] tackles the challenge of occlusion in single-view RGB-D systems by introducing an occlusion-aware motion estimation approach using a lightweight graph neural network. This method predicts full 3D motion, including occluded regions, enhancing real-time dynamic reconstruction. Its strength lies in handling occlusions effectively; however, it may be computationally intensive for very large scenes or highly dynamic environments.

Point cloud accumulation techniques [3] aim to enhance data density by merging consecutive LiDAR scans. While effective in static environments, their weakness becomes apparent in dynamic scenes where moving objects can introduce alignment errors. Methods that disentangle static background from dynamic foreground improve performance but require accurate motion estimation of dynamic objects, which can be challenging.

Surface modeling approaches like As-Rigid-As-Possible (ARAP) surface modeling [4] focus on preserving local rigidity to achieve natural deformations. ARAP excels in applications like animation and virtual sculpting due to its intuitive manipulation capabilities. However, it struggles with large rotations and may face computational challenges with complex meshes. DynamicFusion [11] extends reconstruction to non-rigid scenes in real-time by jointly solving for shape and motion. Its strength is in achieving real-time performance, but it can face difficulties with data association and robustness in handling rapid or complex deformations.

Neural rendering methods have brought significant improvements in dynamic scene reconstruction. Nerfies [21] and DynIBaR [12] extend NeRF to handle non-rigid deformations from monocular images. These methods offer high-quality reconstructions and novel view synthesis but often at the cost of computational efficiency. Challenges include maintaining photometric consistency and dealing with

Table 1. Comparison of Dynamic 3D Understanding Methods

Method	Approach	Real-Time	Handles Occlusions	Non-Rigid Deformations
OcclusionFusion [14]	Graph-based GNN	Yes	Yes	Partial
ARAP [15]	Non-linear Optimization	No	N/A	Yes
DynamicFusion [11]	Volumetric Fusion	Yes	No	Yes
Nerfies [21]	Deformation Fields	No	N/A	Yes
DynIBaR [12]	Image-Based Rendering	Yes	N/A	Yes
D3Fields [1]	Implicit Neural Fields	Yes	Struggles	Yes
AdaptiGraph [5]	Material-Adaptive Graph Neural Network	No	N/A	Yes

limited data, especially in uncontrolled environments.

In robotic manipulation, Dynamic 3D Descriptor Fields (D3Fields) [1] and AdaptiGraph [5] provide innovative solutions for dynamic, semantic scene representation and manipulation of deformable objects. D3Fields integrates semantic information into 3D representations, enabling zero-shot generalization. Its dependence on visual foundation models can be a limitation if the models lack fine-grained semantic distinctions. AdaptiGraph [5] introduces material-adaptive graph-based neural dynamics to handle unknown physical properties, showing strength in adapting to diverse materials. However, it may face computational demands during real-time adaptation and require extensive data for training.

3. Methodology

In this section, we present a comprehensive analysis of state-of-the-art methods for dynamic 3D understanding. We categorize these methods based on their key design decisions and techniques, compare their strengths and weaknesses, and highlight current capabilities and gaps. This survey covers seven approaches as seen in Table 1: OcclusionFusion [14], As-Rigid-As-Possible (ARAP) surface modeling [15], DynamicFusion [11], Nerfies [21], DynIBaR [12], D3Fields [1] and AdaptiGraph [5].

Table 2 summarizes the advantages, disadvantages, strengths, and weaknesses of the surveyed methods, highlighting areas for potential improvement.

3.1. Occlusion-Aware Motion Estimation Methods

OcclusionFusion [14] focuses on real-time dynamic 3D reconstruction in the presence of occlusions. It employs a graph-based representation to parameterize object motion using sparse nodes sampled on the object's surface. The method combines visible motion estimation using optical flow and depth information with occluded motion prediction via a graph neural network (GNN) enhanced by temporal integration using an LSTM module. The strength of Occlusion Fusion is that it can effectively handle occluded regions in real-time, incorporate temporal information for improved motion prediction and models per-node motion confidence to enhance robustness. Main drawbacks are that it does not handle topology changes. Also, performance depends on the quality of optical flow estimation.

3.2. Surface Deformation Methods

As-Rigid-As-Possible (ARAP) surface modeling [15] aims to preserve local geometric details during mesh deformation. Non-linear ARAP methods minimize an energy function that measures deviations from local rigidity, enabling natural and intuitive deformations. Linear methods, such as Laplacian surface editing, simplify computations but are less effective in handling large rotations. The strength of ARAP is that it preserves local details and handles significant rotations. It also provides intuitive controls for mesh manipulation. Main drawbacks are that it is computationally intensive due to non-linear optimization and requires careful initialization to ensure convergence.

3.3. Template-Free Non-Rigid Reconstruction Methods

Dynamic Fusion [11] is a pioneering template-free method for real-time reconstruction of non-rigid scenes without prior models. It incrementally builds a canonical model while estimating a dense volumetric warp field to align new observations. The method handles arbitrary objects and deformations but may accumulate drift over time and struggles with large deformations or topological changes. The strength of DynamicFusion is that it eliminates the need for a pre-scanned template and is capable of real-time performance with open world objects. Main drawbacks are that it is susceptible to drift accumulation over time and is limited in handling large deformations or topology changes.

Table 2. Comparison of dynamic 3D understanding methods.

Method	Strengths	Limitations	Future Research Directions
Dynamic 3D Scene Analysis	Accurate motion estimation and segmentation; suitable for autonomous driving applications	Limited to rigid scenes; struggles with occlusions and sparse data	Incorporate self-supervised learning; improve robustness to occlusions; integrate multi-modal data
As-Rigid-As-Possible (ARAP) Surface Modeling	Preserves local geometry; intuitive deformations; efficient computations	May produce artifacts in extreme deformations; high computational cost for large models	Accelerate optimization algorithms; adaptive mesh refinement; integrate machine learning
DynamicFusion	Real-time non-rigid reconstruction; handles occlusions to some extent	Sensitive to rapid motions; cannot handle topology changes; requires depth data	Incorporate global optimization; handle topology changes; integrate machine learning for correspondence estimation
Nerfies (Deformable Neural Radiance Fields)	High-quality rendering of deformable scenes from monocular images; captures complex deformations	Computationally intensive; struggles with large or rapid motions; lacks physical constraints	Improve computational efficiency; handle large deformations and topology changes; incorporate physical constraints
DynIBaR (Neural Dynamic Image-Based Rendering)	Dynamic novel view synthesis from monocular videos; good temporal consistency	High computational cost; requires large amounts of data; struggles with large motions	Improve computational efficiency; handle larger and faster motions; enhance generalization to unseen scenes
D3Fields (Dynamic 3D Descriptor Fields)	Zero-shot generalization; handles dynamic environments; integrates visual semantics	May not capture fine-grained semantic distinctions; struggles with occlusions; scalability issues	Enhance semantic granularity; improve occlusion handling; scale to larger scenes; integrate language models
AdaptiGraph	Material-adaptive dynamics; effective for manipulation of deformable objects with unknown properties	Computationally intensive; may not generalize to entirely new materials or multiple properties	Extend material types and properties; improve computational efficiency; incorporate uncertainty estimation

Table 3. Comparison of dynamic 3D understanding evaluation metrics and datasets used for evaluation.

Method	Evaluation Metrics	Datasets Used
Dynamic 3D Scene Analysis	EPE, AccS, AccR, Outlier Ratios	Waymo Open Dataset, nuScenes, KITTI
As-Rigid-As-Possible (ARAP) Surface Modeling	Edge Length Preservation, Angle Preservation, Energy Minimization, Visual Quality, Computational Time	Standard Meshes, Synthetic Models, Real-world Scans
DynamicFusion	Reconstruction Accuracy, Tracking Robustness, Computational Performance, Completeness	Synthetic Datasets, Real-world Captures, Berkeley MHAD
Nerfies	PSNR, SSIM, LPIPS, Depth Error Metrics	Synthetic Datasets, Real-World Captures, DeepDeform
DynIBaR	PSNR, SSIM, LPIPS	Nvidia Dynamic Scene Dataset, UCSD Dynamic Scenes Dataset, DYNACT Dataset, Real-World Videos
D3Fields	Correspondence Accuracy, Reconstruction Quality, Manipulation Success Rate, Computation Time	Real-World Scenes, Simulated Environments, Benchmark Tasks
AdaptiGraph	MSE, Chamfer Distance, Task Success Rate	Deformable Objects, Physical Property Variations, Manipulation Tasks

3.4. Deformable Neural Radiance Fields

Nerfies [21] extend Neural Radiance Fields (NeRF) to dynamic scenes by introducing deformation fields that map observations to a canonical space. This method can handle complex, non-rigid deformations without requiring multi-view or depth data. However, the training process is computationally intensive and may encounter local minima during optimization. The strength of Nerfies is that it can handle complex non-rigid deformations effectively. Also, it does not require multi-view or depth inputs. Main drawbacks are that it is computationally costly during training. Also, optimization may be prone to getting stuck in a local minima.

3.5. Dynamic Image-Based Rendering Methods

DynIBaR [12] introduces a volumetric image-based rendering framework that aggregates features from nearby views in a motion-aware manner. It is scalable to long videos with complex motions and produces high-quality renderings. The method requires careful feature aggregation and may need additional regularization to avoid artifacts. The strength of DynIBaR is that it is capable to run long sequences with complex motions and produces high-quality renderings suitable for novel view synthesis. Main drawbacks are that it requires precise feature aggregation techniques and may need additional regularization to prevent artifacts.

3.6. Dynamic 3D Descriptor Fields

D3Fields [1] integrate visual foundation models with implicit representations to provide dynamic, 3D, and semantic scene understanding. They offer zero-shot generalization to new objects and scenes without retraining and efficiently capture dynamics for manipulation tasks. However, they depend on the quality of the underlying visual models and may struggle with fine-grained semantic distinctions. Main strengths are zero-shot generalization to unseen objects and scenes, efficient computation suitable for real-time applications and being able to captures dynamics relevant for robotic manipulation tasks. Limitations include dependency on the performance of visual foundational models and inheriting their limitations such as poor fine-grained semantic understanding.

4. Evaluation

The evaluation of dynamic 3D understanding methods is crucial for assessing their performance in terms of accuracy, robustness, and applicability to real-world scenarios. This section discusses the evaluation metrics, datasets, and protocols used by various methods and provides a comparative analysis of their strengths and weaknesses.

Table 3 summarizes the evaluation metrics and datasets used to evaluate different aspects of dynamic 3D understanding methods.

The methods surveyed employ a variety of evaluation metrics tailored to their specific goals. For instance, methods focusing on motion estimation and segmentation in dynamic scenes use metrics like EPE and accuracy thresholds (Dynamic 3D Scene Analysis), whereas methods dealing with image synthesis and rendering (Nerfies, DynIBaR) utilize image-based metrics such as PSNR, SSIM, and LPIPS.

Datasets also vary significantly among the methods. Autonomous driving datasets like Waymo Open Dataset and KITTI are prevalent for motion estimation tasks, while synthetic datasets and real-world captures are used for evaluating non-rigid reconstruction and deformation methods.

5. Future Work

Future research should focus on developing robust deformation models and real-time adaptive algorithms to handle complex deformations and topology changes, enhancing methods like DynamicFusion and Nerfies. Improving computational efficiency through optimization techniques and hardware acceleration can make methods like ARAP surface modeling suitable for real-time applications.

Enhancing generalization and data efficiency via self-supervised learning and domain adaptation can help methods like D3Fields and AdaptiGraph operate effectively without extensive retraining. Integrating multi-modal data and physical constraints can improve accuracy and realism. Advances in deep learning can enhance dynamics modeling and correspondence estimation, while developing collaborative systems for interactive 3D environments could open new applications.

References

[1] Y. Wang, Z. Li, M. Zhang, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D^3Fields: Dynamic 3D Descriptor Fields for Zero-Shot Generalizable Robotic Manipulation," *arXiv preprint arXiv:2309.16118*, 2023. 2, 4

[2] D. Zhiqiang, W. Wang, F. Dai, *et al.*, "OcclusionFusion: Occlusion-aware Motion Estimation for Real-time Dynamic 3D Reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[3] A. Dewan, C. Wolf, and A. Knoll, "Dynamic 3D Scene Analysis by Point Cloud Accumulation," *IEEE Transactions on Robotics*, 2016. 1

[4] O. Sorkine and M. Alexa, "As-Rigid-As-Possible Surface Modeling," in *Symposium on Geometry Processing*, 2007. 1

[5] K. Zhang, B. Li, K. Hauser, and Y. Li, "AdaptiGraph: Material-Adaptive Graph-Based Neural Dynamics for Robotic Manipulation," *arXiv preprint arXiv:2407.07889*, 2024. 2

[6] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2446–2454. 6

[7] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631. 6

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. 6

[9] W. Dai, W. Wang, F. Dai, and Y. Li, "OcclusionFusion: Occlusion-aware motion estimation for real-time dynamic 3D reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3477–3486, 2020.

[10] A. Dewan, C. Wolf, and A. Knoll, "Dynamic 3D scene analysis by point cloud accumulation," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1334–1347, 2016.

[11] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 343–352, 2015. 1, 2, 6

[12] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "DynIBaR: Neural dynamic image-based rendering," *arXiv preprint arXiv:2205.15838*, 2022. 1, 2, 4

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020.

[14] Y. Dong, S. Zhang, H. Bao, and X. Zhou, "OcclusionFusion: Occlusion-aware Motion Estimation for Real-time Dynamic 3D Reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11937–11946, 2022. 2

[15] O. Sorkine and M. Alexa, "As-Rigid-As-Possible Surface Modeling," in *Symposium on Geometry Processing*, vol. 4, no. 2, pp. 109–116, 2007. 2

[16] *The Stanford 3D Scanning Repository*. [Online]. Available: http://graphics.stanford.edu/data/3Dscanrep/ 6

[17] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 213–230, 2008. 6

[18] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, 2006, pp. 61–70.

[19] G. Varol, D. Cremers, and I. Laptev, "Learning Scene Flow with CNNs," in *2017 IEEE International Conference on Computer Vision*, 2017, pp. 2921–2929. 6

[20] F. Ofli *et al.*, "Berkeley MHAD: A Comprehensive Multimodal Human Action Database," in *2013 IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53–60. 6

[21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1, 2, 4

[22] Aljaž Božič, Pablo Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7002–7012, 2020.

6

6. Appendix Section

6.1. Evaluation Metrics

- **End-Point Error (EPE)**: Measures the mean Euclidean distance between estimated and ground truth motion vectors for each point.
- **Accuracy Metrics (AccS, AccR)**: Represent the percentage of points with EPE below certain thresholds, indicating strict or relaxed accuracy.
- **Outlier Ratios**: The proportion of points with large errors, highlighting the robustness of the method.
- **Edge Length and Angle Preservation**: Assess how well deformation methods preserve the original mesh geometry.
- **Peak Signal-to-Noise Ratio (PSNR)**: Evaluates reconstruction quality in image synthesis tasks; higher values indicate better quality.
- **Structural Similarity Index (SSIM)**: Assesses perceived image quality based on structural similarity; higher values are better.
- **Learned Perceptual Image Patch Similarity (LPIPS)** [4]: Measures perceptual similarity using deep neural network features; lower values are better.
- **Mean Squared Error (MSE)**: Quantifies the error between predicted and actual particle positions in dynamics modeling.
- **Chamfer Distance (CD)**: Computes the distance between predicted and ground truth point clouds.
- **Task Success Rate**: Evaluates the effectiveness of methods in completing specific manipulation or rearrangement tasks.
- **Computational Performance**: Includes metrics like frames per second (FPS) and computational time, indicating real-time capabilities.

6.2. Datasets

The following datasets are commonly used to evaluate dynamic 3D understanding methods:

- **Waymo Open Dataset** [6]: LiDAR sequences with annotations for autonomous driving scenes.
- **nuScenes** [7]: Multimodal sensor data with annotations for dynamic scenes.
- **KITTI** [8]: Benchmark data for autonomous driving, including point clouds and annotations.
- **Standard Meshes**: Armadillo, Bunny, Cactus models [16] used for surface modeling evaluations.
- **Synthetic Datasets**: Generated using rendering engines or custom shapes to test specific scenarios [17, 19].
- **Real-World Captures**: Sequences recorded with depth cameras showing various deformations [11].
- **Benchmark Datasets**: Berkeley MHAD [20], DeepDeform [22], Nvidia Dynamic Scene Dataset [3].
- **Simulated Environments**: Physics engines like OmniGibson [1] for manipulation tasks.