International Journal of Information Technology Vol 30 No 2 2024

Object Dynamics Reconstruction from Monocular Videos

Yangsen Chen[1], Yu Feng[1] and Hao Wang*[1]

[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[1]ychen950@connect.hkust-gz.edu.cn, haowang@hkust-gz.edu.cn

Abstract

In this paper, we investigate the challenging task of reconstructing object 3D dynamics and estimating physics parameter from monocular videos. Without any known object geometry or materials, we reconstruct the 3D collision motions for various object types. The key of achieving accurate physics reconstruction is to generate multi-view frames having consistent object positions, such that the motions can be correctly modeled. To this end, we first propose a novel spatial-aware novel-view synthesis module to generate spatially consistent multi-view videos. Specifically, we input bounding boxes with various positions to regularize the novel-view synthesis process. Further, we propose the temporal-aware physics augmentation module. It seamlessly connect the generated pseudo multi-view videos with a voxel-based dynamic neural radiance field, to model the interactions between object material points, in which our proposed view adaptive 3D reconstruction loss and temporal information is adopted. Our experiments demonstrate the efficacy of our model, in various scenarios, showcasing its ability to handle complex object structures and dynamics with high fidelity.

Keywords: 3D Reconstruction, Neural Radiance Fields.

*Corresponding author

Yangsen Chen, Yu Feng, Hao Wang

I. Introduction

The field of 3D content generation has made significant progress, especially in the development of 3D object generation. However, there are still substantial gaps that need to be addressed to effectively use AI-generated objects in real-world applications, such as the film and game industry. A notable gap is that the generated objects from the existing methods [25, 18, 26, 2] do not possess any inherent physical properties, including material, density, etc. As a result, the involvement of professional designers is still necessary to provide precise parameters, such that the object properties can be accurately modeled. This aims to ensure the generated objects to interact with the environment and be seamlessly incorporated into film and game creation.

In this paper, we aim to automate the physics parameters estimation process. This task is particularly challenging due to the complex and non-rigid nature of given object shapes, which demands sophisticated modeling to capture their structures and dynamics accurately. Prior studies [8, 10] have shown the potential for modeling object motion dynamics automatically. However, these methods rely on learning from a large volume of videos. For instance, Li et al. [10] require thousands of videos with similar physics patterns for training, making it difficult to collect such data for each specific physical property. Additionally, it is infeasible for us to perform individual modeling for given objects and scenarios.

To address this challenge, we propose to adopt the monocular video of an object only, to reconstruct its corresponding 3D dynamics. Our research introduces a novel framework, that integrates the pretrained multi-view synthesis model [15] with the Neural Radiance Fields (NeRF) [17]. Technically, we first produce multi-view videos from the given monocular videos. Then inspired by [8], we adopt voxel-based NeRF [12] to produce the 3D object representation and obtain the material point clouds to model their physical interactions. This is a non-trivial process, as we observe the object positions in the generated novel views are not consistent, which adds much noise when modeling the dynamic motions between frames. Moreover, the generated novel-view information does not perfectly align with the ground truth, making the 3D reconstruction results inaccurate.

Therefore, our core innovations lie in two folds. Firstly, the spatial-aware novel-view synthesis module, which serves as our proposed controlling mechanism over the object positions on the generated novel views, ensuring that object motions can be modeled without noise. To achieve this, we fine-tune the pretrained multi-view synthesis model [15] by using random bounding boxes as conditions to regularize the novel-view synthesis process. Secondly, the temporal-aware physics augmentation module, which integrates a view-adaptive 3D reconstruction loss and temporal information, seamlessly combining the multi-view generation model with the voxel-based NeRF model. Our design ensures high fidelity in the reconstructed outputs. To the best of our knowledge, this is the first work of successfully reconstructing the physical dynamics of objects from monocular video. Our contributions are summarized as:

- We propose a novel framework that dynamically reconstruct 3D object and estimate physics parameters from monocular video without any object geometry or materials as input.

- We introduce the spatial-aware novel-view synthesis module and the temporal-aware physics augmentation module as solutions aimed at mitigating challenges related to ensuring consistency in reconstruction within the tasks.

- With only monocular video as input, our framework achieves comparable results as the state-of-the-art dynamics reconstruction results [8] from multi-view videos.

II. Related Works

A. Neural Radiance Fields for Dynamic Scene

Neural Radiance Field (NeRF) [16] is promising for 3D generation tasks, and significant progress has been made in accurately rendering dynamic scenes using NeRFs. The Non-Rigid Neural Radiance Fields [23] technique enables high-quality reconstruction of non-rigid dynamic scenes from RGB images. Neural Scene Flow Fields [9] utilizes a time-variant continuous function that encapsulates 3D motion, enhancing the realism and accuracy of dynamic scene rendering. These advancements signify a major leap in the ability to model and recreate complex, dynamically changing environments. While these work is great, they do not fully utilize the real-world physical priors and constraints.

Recently, physics-aware dynamic NeRFs have received high attention, PAC-NeRF [8] represents a significant advancement in estimating both the unknown geometry and physical parameters of highly dynamic objects using multi-view video recordings. Similarly, the work [3] of Chu et al. in reconstructing dynamic fluid phenomena leverages the governing physics for end-to-end optimization from sparse video frames. While these methods offer valuable insights for our project, they rely on highly calibrated, simultaneously recorded multi-view videos, thus limiting their accessibility and practical use. In contrast, our approach utilizes single-view videos without requiring detailed knowledge of the camera. This difference is crucial as it allows us to preserve most of the dynamic effects while ensuring our method is accessible and practical, broadening its applicability.

B. Diffusion Model for 3D Generation Tasks

Recent advancements have witnessed the innovative application of 2D diffusion models such as [20] in executing 3D tasks. Foremost among these pioneering efforts are DreamFusion [19] and

SJC [24], which introduced the concept of distilling a 2D text-to-image generation model to create 3D shapes from text descriptions. These foundational works have inspired a series of subsequent studies [14, 27] that extend to more 3D tasks.

In the single-view reconstruction field, Zero123 [14] represents a seminal work in enabling open-world single-image-to-3D conversion through zero-shot novel view synthesis. However, despite its impressive performance, there remains a notable challenge in addressing geometric inconsistencies across generated images, a critical aspect in bridging the divide between multi-view imagery and cohesive 3D scenes. Recent initiatives, including One-2-3-45 [13], SyncDreamer [15], Consistent123 [11] and Zero123++ [21], have made strides in overlaying additional layers onto Zero123 to achieve more 3D-consistent outcomes. In our work, we build upon these developments with targeted modifications to Zero123, tailoring it to meet the specific demands of our project. Our focus lies in optimizing performance for the unique challenge of single-view dynamic body object reconstruction. Distinct from other methods that concentrate on static single-view reconstruction, our approach prioritizes achieving positional stability and temporal consistency across frames, particularly for videos of highly dynamic objects. This focus is crucial for ensuring realistic physical effects in our reconstructions.

III. Method

In Figure 1, we present our framework. It aims to achieve precise 3D reconstruction of highly dynamic continuum objects from monocular video, showing its physical interaction with solid ground. To sum up, our framework has two key components: a spatial-aware novel-view synthesis module (Section A.) and a temporal-aware physics augmentation module (Section B.). Figure 2 illustrates the training pipeline of our proposed spatial-aware novel-view synthesis module.

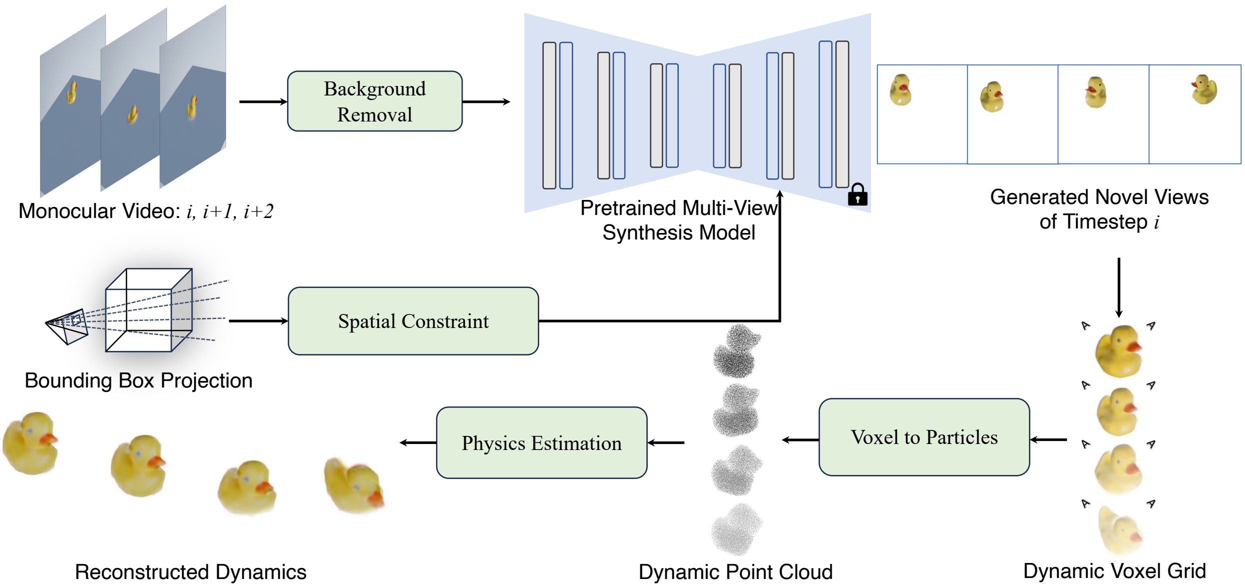

Fig 1 The pipeline of our proposed framework, the background removed images are fed into the pretrained multi-view synthesis model, regularized with our proposed spatial-aware novel-view synthesis module, to generate novel views for each video frame. Then we reconstruct 3D voxel grids from the generated multi-view results, which are further converted to dynamic point clouds. Based on these particles, we further estimate the physical parameters and reconstruct object dynamics from a monocular video.

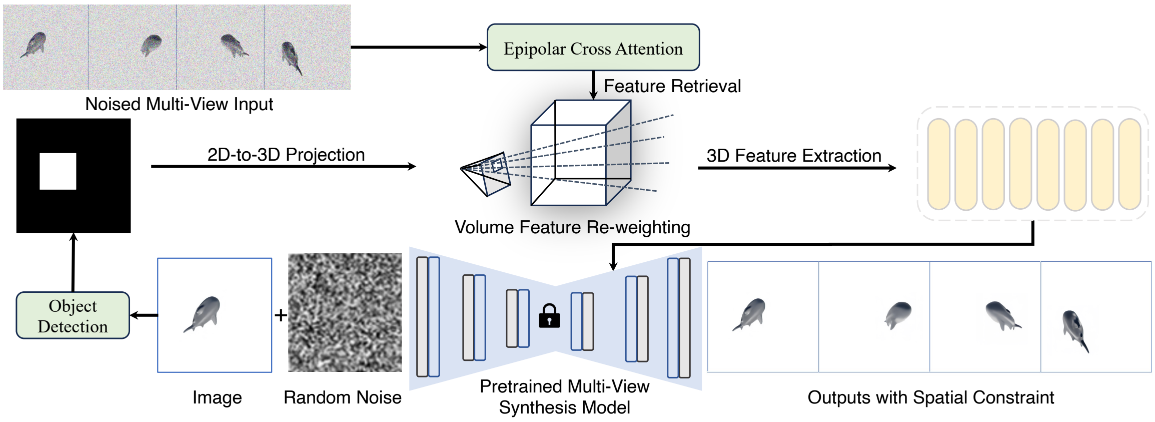

Fig 2 The training pipeline of spatial-aware novel-view synthesis module. Firstly, we use an off-the-shelf bounding box detection model to get the 2D bounding box of the object from the input monocular image. Then we conduct 2D-to-3D projection to allow the volume feature to only activate the feature vectors inside the bounding box region. In terms of the noised multi-view input, we extract their features with the epipolar cross attention. We further retrieve the epipolar features of the bounding box region, which are summed with activated features. The combined features are fed into a 3D CNN to learn spatial feature, which is the condition for the diffusion process to force the model to generate the objects at the correct area. The multi-view synthesis backbone weights is fixed to preserve existing generation ability.

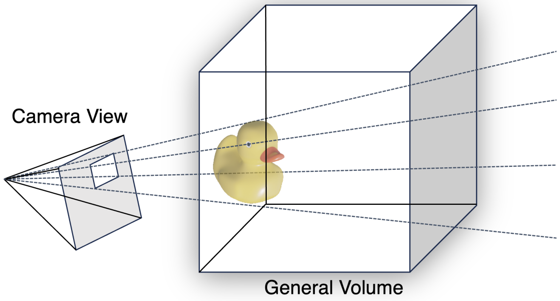

Fig 3 The bounding box constraint illustration. The bounding box are projected to the general feature volume, depicting the boundary between the region where the object is present and the region having no object. Following such guidance, more weights can be assigned to the feature grid within the boundary.

A. *Spatial-Aware Novel-View Synthesis*

From the existing multi-view synthesis models [15, 14], we know that they always generate objects located at the center of the novel-view images. Since there is no position change between the adjacent frames, the estimated speed would be zero, making the motion speed estimation infeasible. To address this issue, we propose the spatial-aware novel-view synthesis module to help SyncDreamer [15] generate images with objects in their correct 3D positions. Also, it can be seamlessly integrated into the SyncDreamer model.

In Figure 2, we show our training process of the spatial-aware novel-view synthesis module. Firstly, given an input image, we can generate the bounding box for locating the object, based on the object detection tool [1]. To utilize this bounding box as conditioning, we treat our spatial-aware novel-view synthesis module as conditioning. This module accepts both the noised target view and the bounding box. In the following, we give a detailed illustration of the spatial-aware novel-view synthesis module, including five parts.

Bounding Box Unprojection. As depicted in the Figure 3, we unprojected the 2D bounding box into 3D spaces through matrix manipulation. This procedure can be integrated flawlessly into the training and inference process. Naturally, this view outline derived from our bounding box shades also affects the positional information of the object to be generated.

Epipolar Attention Retrieval. To extract 3D features, we perform feature extraction on the noised target views, generated during the training step of the diffusion model, using Epipolar cross-attention. The Epipolar cross-attention mechanism shades light on the 3D geometry information, as cross attention is only made on the corresponding epipolar lines. It grants that attention is made only at 3D-related areas between images, which will benefit our further manipulation of the volume feature. After constructing the feature map built by Epipolar attention, we perform the retrieval of the feature vector from 2D to 3D volume features, this process can be done through the usage of a projection matrix. After retrieval, all the features from multi-view targets are gathered and fused inside our general feature volume. Let E_{ij} be the epipolar line corresponding to a point in F_i in the view I_j. The cross-attention mechanism can be represented as:

$$A_{ij}(p) = \sum_{q \in E_{ij}} \alpha_{pq} \cdot F_j(q), \tag{1}$$

where α_{pq} is the attention weight, and p and q are points in the feature maps of I_i and I_j, respectively.

3D Feature Projection. As shown in Figure 2, we first project the retrieved Epipolar features of dimensions $[H, W, F]$. Here, F represents the length of the feature vector, while H and W are proportional to the original image size, with width and height set to be equal. Subsequently, the feature volume with dimensions $[F, V, V, V]$ is constructed, resembling a high-dimensional 3D cube of size $[V, V, V]$. Each grid in this space stores a feature vector. Similar to projecting 3D coordinates onto 2D images, we project each 3D grid onto the feature map to retrieve its feature vector. This process aggregates features from all feature maps, enabling the fusion of multi-view information. Note that we assume that the objects for the task are always inside this 3D cube of size $[V, V, V]$. We denote this feature volume as general feature volume, and the 3D cube in the normal 3D coordinate system to be general volume for convenience of discussion below.

The resultant general feature volume, enriched with aggregated multi-view information, is then processed to adjust their dimensions to be compatible with the input requirements of the diffusion model's UNet Backbone, which finally perform conditioning on the generation process, ensuring the produced output follow the principle of multi-view consistency.

Volume Feature Re-weighting. Since there is nothing of interest to us outside the boundary, bounding box constraint is essential. Also, we can assign low activation on feature volumes there, while giving more weight to the feature volumes inside the frustum. In this way, the learned general feature volume will contain knowledge about the objects' generating positions. At the same time, the knowledge can be further transformed to the diffusion model as a condition, giving the diffusion model a sense of spatial constraint when generating objects' novel view. The reweighting function $\phi(x,\ B)$ assigns low activation outside the bounding box B and higher activation inside it, if $x \in B$, $\phi(x,\ B) = 1$, else $\phi(x,\ B) = \epsilon$, where ϵ is a small value.

3D Feature Extraction. To extract 3D features efficiently, we need to carefully design the feature extraction pipeline. In our framework, it is initialized by a large volume of multi-dimensional features. These features undergo a process of progressive feature extraction and reshaping, making them suitable for integration with the diffusion UNet and serving as a conditioning mechanism. Inspired by depth-wise attention methods, we utilize 3D Convolutional Neural Networks (CNNs) and depth-wise attention mechanisms to retain a robust 3D understanding the data. 3D CNNs and depth-wise attention [5] play crucial roles in preserving and integrating three-dimensional feature information. Both of them ensure that the nuanced spatial relationships inherent in the data are effectively captured and utilized in the conditioning process for the diffusion UNet.

B. Temporal-Aware Physics Augmentation

As a 3D reconstruction model, our framework has the potential to be applied to a wide range of downstream reconstruction tasks, given monocular video as input. A notable application is to reconstruct the dynamics of objects, leveraging the accurate relative position information provided by our model. In this work, we focus on reconstructing the physical properties of highly dynamic materials, such as rubber, plasticine, and sand, from the monocular video.

Prior work PAC-NeRF [8], has achieved significant success in reconstructing the unknown geometry and estimating physical parameters of highly dynamic objects from multi-view videos, for about more than 10 videos. It first reconstructs a static voxel grid for each frame using a voxel-based neural radiance field [22]. Then it transforms them to particles for estimation of velocity and other physical parameters using differentiable simulation [7]. Inspired by its success in dynamic reconstruction, we utilize PAC-NeRF in our task with critical task-specific modification.

Though PAC-NeRF has shown great performance in 3D dynamic reconstruction from multi-view video, it cannot be directly applied to monocular video dynamic reconstruction scenarios. Specifically, although it is a feasible solution to reconstruct a separate DVGO[22] for each frame, the reconstruction of results in a 4D scenario often lacks coherence. Furthermore, in complex scenes, it is easy to lose temporal consistency within each frame, especially when pseudo ground truth used as input. In a word, there are two main challenges when exploring PAC-NeRF to monocular video-based dynamic reconstruction: how to encode temporal information and how to deal with inaccurate synthetic views. To address these problems above, we propose a unified dynamic reconstruction framework. In this section, we describe how our model, with only monocular video input, leverages synthetic views from a multi-view diffusion model to generate results with temporal consistency and reconstruction performance close to PAC-NeRF. In the following, we describe details of temporal-aware physics augmentation module, including temporal information encoding and utilizing pseudo-ground truth.

Temporal Information Encoding. Drawing on the innovations presented in TiNeuVox[6], we encode temporal information from two perspectives: coarse coordinate deformation and enhancement of temporal information. In the case of coarse coordinate deformation, a compact deformation network, employing only 3-layer multilayer perceptrons (MLPs), modifies the spatial coordinates. This network, notably smaller than those used in prior studies of dynamic scenes, aims to reduce computational demands. To this end, we have optimized the network's structure, reducing both its size and complexity to facilitate faster optimization and rendering. For the enhancement of temporal information, the substantial reduction of the deformation network's size in our approach may lead to inevitable inaccuracies in coordinate adjustment due to its diminished capacity. Moreover, these inaccuracies are likely to increase when neural voxels are accessed based on point coordinates, affecting not only interpolation accuracy but also the fidelity of retrieved vertices. To address this issue, we enhance temporal information by merging interpolated features with both positionally-encoded coordinates and neural-encoded temporal embeddings. These amalgamated inputs are then processed by neural networks, effectively minimizing inaccuracies.

Pseudo Ground Truth Utilization. To maintain the generated results with great reconstruction performance, we introduce a novel loss function, termed ”View Adaptive Loss” to prioritize the real view while assigning lesser weight to synthesized views. Since PAC-NeRF cannot converge stably, we adjust our model by loosening the constraints on physical parameter estimation. Instead of strict figures, we use ranges. In concrete terms, when predictions fall within these ranges, we apply a less stringent penalty to the loss. The loss function is mathematically defined as:

$$L_{\text{va}} = \sum_{i=1}^{N} w_i \cdot L_i. \tag{2}$$

Here, N represents the total number of views, including real and synthesized ones, L_i is the loss for the i-th view, and w_i is the corresponding weight.

The weight w_i for each view is computed from the angular difference from the primary view:

$$w_i = \exp\left(-\alpha \cdot \Delta\theta_i\right). \tag{3}$$

In this formula, $\Delta\theta_i$ denotes the 3D angular difference between the i-th view and the primary view, with α as a scaling factor modulating the influence of angular differences on the weight. This formulation ensures that the primary view (the only real data view) has the smallest angular difference and it receives the highest weight. The exponential decay of weight with increasing angular difference, governed by $\exp(-\alpha \cdot \Delta\theta_i)$, ensures that views closer to the primary view exert greater influence on the overall loss, aligning to emphasize real over synthesized data.

IV. Experiments

In this section, we demonstrate the capacity and advantage of our proposed method. In the following part, we describe the various elements associated with the experiments, including the experimental settings, data, etc.

A. Experimental Settings

Implementation details. We employ a latent diffusion model (LDM) (Rombach et al., 2021) as the foundational framework for generating multi-view images. Additionally, we adhere to the SyncDreamer configuration and conduct fine-tuning utilizing their publicly available checkpoint. Employing the UNet architecture with their suggested partition locked, we conduct training on images of dimensions 256×256. Subsequently, we fine-tune the model to improve the acquisition of objects' relative 3D positions at a designated elevation angle of 30 degrees, revolving in a cycle around the origin point (0,0,0). Our augmentation involves the integration of a novel spatial-aware novel-view synthesis module, while maintaining the UNet architecture and specific components of SyncDreamer unchanged. This fine-tuning procedure encompasses training the model on a sequence of 16 views, uniformly distributed around the central point located at (0,0,0).

Training data. During the training phase of our multi-view diffusion model, object positions undergo random perturbations around the origin point. This stochastic process is meticulously executed to guarantee that the entire object remains within the field of view, thereby preserving the integrity of the image data. We curated a subset of the Objaverse dataset [4], introducing random shifts in object positions to align with our training requirements.

Test data. To assess the efficacy of our model, particularly concerning physically informed 3D dynamics reconstruction, we utilized the PAC-NeRF dataset. This dataset offers a solid basis for quantitative evaluations, enabling us to gauge the model's fidelity in reconstructing 3D dynamics accurately. During testing, object positions were determined using a uniform distribution, facilitating an evaluation of the model's capacity to accommodate diverse spatial configurations and bolstering the generalizability of our findings.

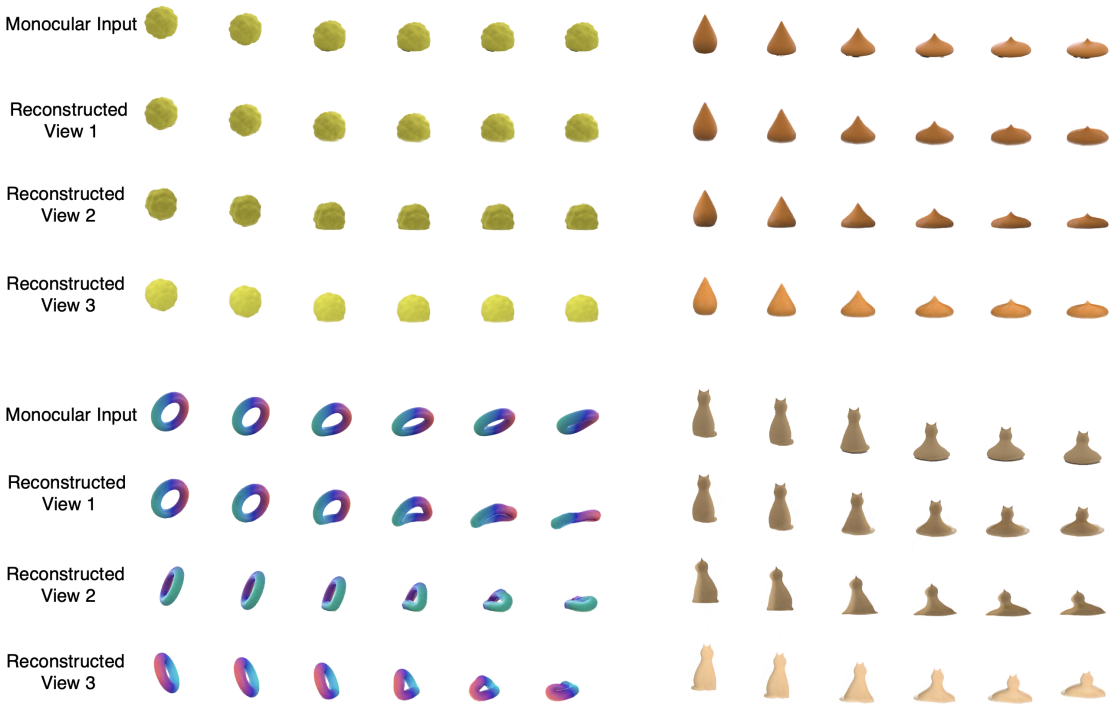

Fig 4 Qualitative result on objects with diverse materials. The top-left object is of Plasticine type, the top-right object is of Newtonian fluid type, the bottom-left object is of elastic type, the bottom-right object is of Plasticine type.

Object	Method	Estimated Parameters	Ground Truth
Droplet	Single View Multi View	$\mu = 218, \kappa = 1.20 * 10^5$ $\mu = 2.09 \times 10^2, \kappa = 1.08 \times 10^5$	$\mu = 200, \kappa = 10^5$
Cream	Single View Multi View	$\mu = 1.95 \times 10^5, \kappa = 1.74 \times 10^6, \tau_Y = 3.41 \times 10^3, \eta = 5.5$ $\mu = 1.21 \times 10^5, \kappa = 1.57 \times 10^6, \tau_Y = 3.16 \times 10^3, \eta = 5.6$	$\mu = 10^4, \kappa = 10^6, \tau_Y = 3 \times 10^3, \eta = 10$
Toothpaste	Single View Multi View	$\mu = 7.93 \times 10^3, \kappa = 2.66 * 10^5, \tau_Y = 247, \eta = 9.50$ $\mu = 6.51 \times 10^3, \kappa = 2.22 \times 10^5, \tau_Y = 228, \eta = 9.77$	$\mu = 5 \times 10^3, \kappa = 10^5, \tau_Y = 200, \eta = 10$
Torus	Single View Multi View	$E = 1.52 \times 10^6, \nu = 0.420$ $E = 1.04 \times 10^6, \nu = 0.322$	$E = 0^6, \nu = 0.3$
Bird	Single View Multi View	$E = 2.60 \times 10^5, \nu = 0.261$ $E = 2.78 \times 10^5, \nu = 0.273$	$E = 3 \times 10^5, \nu = 0.3$
Playdoh	Single View Multi View	$E = 5.14 \times 10^6, \nu = 0.23, \tau_Y = 1.90 \times 10^4$ $E = 3.84 \times 10^6, \nu = 0.272, \tau_Y = 1.69 \times 10^4$	$E = 2 \times 10^6, \nu = 0.3, \tau_Y = 1.54 \times 10^4$
Cat	Single View Multi View	$E = 1.31 \times 10^5, \nu = 0.330, \tau_Y = 3.07 \times 10^3$ $E = 1.61 \times 10^5, \nu = 0.293, \tau_Y = 3.57 \times 10^3$	$E = 10^6, \nu = 0.3, \tau_Y = 3.85 \times 10^3$
Trophy	Single View Multi View	$\theta^0_{\text{fric}} = 33.5°$ $\theta^0_{\text{fric}} = 36.1°$	$\theta^0_{\text{fric}} = 40°$

Table 1 This table compares estimated parameters for various materials and structures using our Single-View and PAC-NeRF Multi-View methods. The objects under analysis include Droplet, Cream, Toothpaste, Torus, Bird, Playdoh, Cat, and Trophy. Each entry lists the estimated values of their specific materials' physical parameters. The results show a comparison between estimates obtained from a single viewpoints and those obtained from multiple viewpoints. The goal is to assess how closely our estimates align with the known ground truth values. Although the estimated results of our single-view method are not as accurate as those of multi-view method PAC-NeRF, since we use the synthesised multi-view images, but we achieve comparable results with them.

B. 3D Dynamics Reconstruction

In Figure 4, we present additional visualizations of our 3D dynamic reconstruction outcomes. Overall, the results in Figure 4 reflects the effectiveness of our proposed spatial-aware novel-view synthesis modules in 3D dynamic reconstruction. Also, these results demonstrate that our method attains a visual fidelity akin to the ground truth for a majority of the objects, particularly those with conventional geometries.

C. Comparison with Multi-view Method

In this study, we focus on a relatively novel task, since there are limited baseline comparisons available. Our principal baseline for comparison is PAC-NeRF, recognized as a state-of-the-art model in the field. It is noteworthy that the comparisons are conducted under differing conditions:

- Our method relies solely on monocular video as its input.
- PAC-NeRF [8], in contrast, utilizes multi-view videos.

Quantitative results of this comparison are presented in Table 1. We adopt PAC-NeRF's measurement method for physical parameters, conducting tests on five different types of materials: elasticity, Plasticine, granular media (e.g., sand), Newtonian fluids, and non-Newtonian fluids. For elasticity, we predict Young's modulus (E) (representing material stiffness) and Poisson's ratio (ν) (indicating the ability to preserve volume under deformation). For Plasticine, we predict Young's modulus (E), Poisson's ratio (ν), and yield stress (τ_Y) (the stress required to cause permanent deformation/yielding). In the case of Newtonian fluids, we predict fluid viscosity (μ) (representing opposition to velocity change) and bulk modulus (κ) (indicating the ability to preserve volume). For non-Newtonian fluids, we predict shear modulus (μ), bulk modulus (κ), yield stress (τ_Y), and

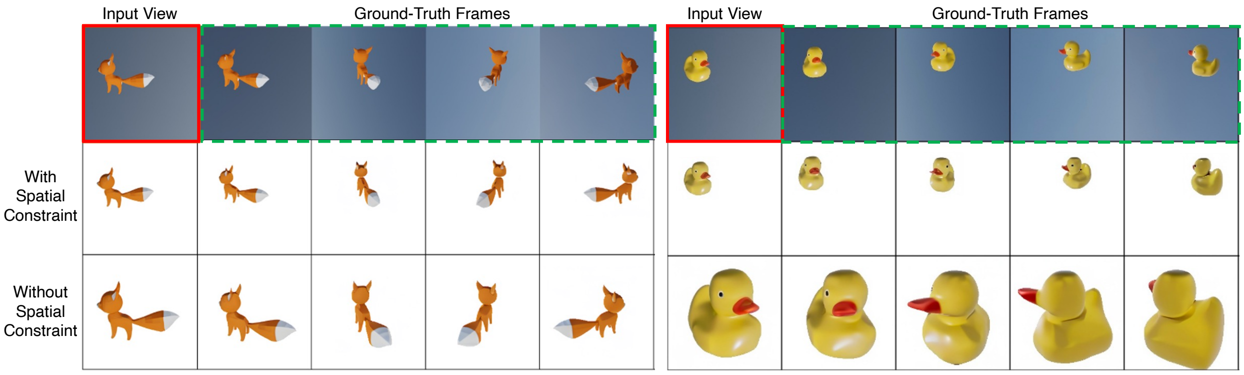

Fig 5 Qualitative analysis of our spatial-aware novel-view synthesis module. We observe that with the proposed spatial constraint, the multi-view synthesis model generates novel views of objects at the original position. However, the model without the spatial constraint does not have such ability, thus demonstrating the efficacy of our module.

plasticity viscosity (η) (representing decayed temporary resistance to yielding). Finally, for sand, we predict the friction angle (θ_{fric}) (a proportionality constant determining the slope of a sand pile). From the Table 1, it can be concluded that the our proposed method, with only monocular video input, can be close to the method PAC-NeRF, with multi-view video input, in the results of each physical parameter estimation. It can demonstrate the excellent performance of our proposed temporal-aware physics augmentation module for dynamic physical parameter estimation.

D. Ablation Study

In this section, we delve into the impact of two fundamental constituents of our framework: spatial control and view-adaptive loss. These factors are instrumental in enhancing the precision and efficiency of object synthesis and simulation convergence within our framework.

Spatial-Aware Novel-View Synthesis. Our enhanced model demonstrates a significant improvement in accurately synthesizing objects at their correct spatial coordinates compared to the original SyncDreamer. This enhancement is evident from the qualitative comparisons illustrated in Figure 5. The absence of spatial-aware novel-view synthesis module results in the clustering of objects towards the center of the image. This central clustering negatively affects velocity es-

Object	with L_{va}	without L_{va}
Droplet	$\mu = 218, \kappa = 1.20 * 10^5$	$\mu = 427, \kappa = 1.82$
Cream	$\mu = 1.95 \times 10^5, \kappa = 1.74 \times 10^6, \tau_Y = 3.41 \times 10^3, \eta = 5.5$	N/A(fail to converge)
Toothpaste	$\mu = 7.93 \times 10^3, \kappa = 2.66 \times 10^5, \tau_Y = 247, \eta = 9.50$	$\mu = 12.33 \times 10^3, \kappa = 4.39, \tau_Y = 136, \eta = 8.53$
Torus	$E = 1.52 \times 10^6, \nu = 0.420$	N/A(fail to converge)
Bird	$E = 2.60 \times 10^5, \nu = 0.261$	$E = 2.44 \times 10^5, \nu = 0.183$
Playdoh	$E = 5.14 \times 10^6, \nu = 0.23, \tau_Y = 1.90 \times 10^4$	$E = 7.29 \times 10^6, \nu = 0.22, \tau_Y = 2.28 \times 10^4$
Cat	$E = 1.31 \times 10^5, \nu = 0.330, \tau_Y = 3.07 \times 10^3$	$E = 0.96 \times 10^5, \nu = 0.13\tau_Y = 2.75 \times 10^3$

Table 2 The ablation Study on the proposed view adaptive loss. The left column records the performance of utilizing the view adaptive loss, and the right column records that of not using the view adaptive loss.

timation, frequently reducing it to nearly zero. Such a situation greatly undermines the model's applicability for simulations. Therefore, the incorporation of spatial control into the generation process is essential for achieving accurate simulations.

View-Adaptive Loss. As illustrated in Table 2, the integration of the view-adaptive loss L_{va} significantly enhances the model's estimation performance and expedites convergence. The utilization of L_{va} demonstrates a notable enhancement in convergence, particularly evident in challenging scenarios involving rare objects such as cream and torus. Without L_{va}, convergence often falters in these cases, emphasizing the critical role of L_{va} in maintaining model stability.

V. Conclusion

In this paper, we present a pioneering method to dynamically reconstruct 3D object and estimate physics parameters from monocular video. Specifically, we introduce a novel framework including two key parts: the spatial-aware novel-view synthesis module and the temporal-aware physics augmentation module. The spatial-aware module allows for precise control over object positions within the generated novel views while concurrently refining the accuracy of motion estimation across frames. The temporal-aware module, with our proposed self-adaptive 3D reconstruction loss and temporal information, adeptly reconstructs the dynamic objects' physics parameters from monocular videos.

Limitation & future work. The primary limitations lie in effectively handling highly complex dynamic scenes. In the future, we intend to research towards addressing more intricate real-world scenarios to reconstruct objects for practical applications.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (No. 62406267), Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3956), Education Bureau of Guangzhou Municipality and the Guangzhou Municipal Education Project (No. 2024312122).

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023.

[3] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics (TOG)*, 41:1 – 14, 2022.

[4] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A uni-

verse of annotated 3d objects. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2022.

[5] Muhammad N. ElNokrashy, Badr AlKhamissi, and Mona T. Diab. Depth-wise attention (dwatt): A layer fusion method for data-efficient classification. *ArXiv*, abs/2209.15168, 2022.

[6] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022.

[7] Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics*, 37(4):150, 2018.

[8] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. *ArXiv*, abs/2303.05512, 2023.

[9] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. *arXiv preprint arXiv:2309.07906*, 2023.

[11] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *ArXiv*, abs/2309.17261, 2023.

[12] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.

[13] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.

[14] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023.

[15] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023.

[16] 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. *Commun. ACM*, 65:99–106, 2020.

[17] 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.

[18] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022.

[20] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[21] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view

diffusion base model, 2023.

[22] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022.

[23] Edgar Tretschk, Ayush Kumar Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12939–12950, 2020.

[24] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Gregory Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12619–12629, 2022.

[25] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023.

[26] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. *arXiv preprint arXiv:2310.03020*, 2023.

[27] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. *ArXiv*, abs/2306.13455, 2023.