Broadening View Synthesis of Dynamic Scenes from Constrained Monocular Videos

Le Jiang   Shaotong Zhu   Yedi Luo   Shayda Moezzi   Sarah Ostadabbas
Northeastern University, Boston, MA
{jiang.l, zhu.shaot, luo.ye, moezzi.s, s.ostadabbas}@northeastern.edu

Abstract

In dynamic Neural Radiance Fields (NeRF) systems, state-of-the-art novel view synthesis methods often fail under significant viewpoint deviations, producing unstable and unrealistic renderings. To address this, we introduce Expanded Dynamic NeRF (ExpanDyNeRF), a monocular NeRF framework that leverages Gaussian splatting priors and a pseudo-ground-truth generation strategy to enable realistic synthesis under large-angle rotations. ExpanDyNeRF optimizes density and color features to improve scene reconstruction from challenging perspectives. We also present the Synthetic Dynamic Multiview (SynDM) dataset, which is the first synthetic multiview dataset for dynamic scenes with explicit side-view supervision, created using a custom GTA V-based rendering pipeline. Quantitative and qualitative results on SynDM and real-world datasets demonstrate that ExpanDyNeRF significantly outperforms existing dynamic NeRF methods in rendering fidelity under extreme viewpoint shifts. Further details are provided in the supplementary materials. Code is available at https://github.com/ostadabbas/ExpanDyNeRF.

1. Introduction

Novel view synthesis is essential in applications like mixed reality [7, 31], medical supervision [30, 37], autonomous driving [21, 40], and wildlife observation [20, 41]. Neural Radiance Fields (NeRF) and their dynamic variants have improved 3D reconstruction with high precision [24, 26, 32], speed [3, 4, 10], and style editing [1, 6]. Alternatively, Gaussian splatting [27, 28, 34] offers an efficient and flexible framework for high-quality rendering. While both approaches produce sharp results from primary viewpoints, renderings degrade with significant viewpoint shifts due to the lack of diverse-view supervision during training, which is a limitation of monocular settings (Fig. 1).

To overcome these limitations, we propose Expanded Dynamic NeRF (ExpanDyNeRF), a dynamic NeRF framework designed to expand reliable rendering to large-angle novel views, even under monocular camera constraints. This end-to-end pipeline, illustrated in Fig. 2, incorporates a novel-view pseudo ground truth strategy that optimizes model training from novel view by leveraging Gaussian priors [25], effectively refining dynamic object contours and color consistency across frames.

One of the key challenges in novel view synthesis for dynamic scenes lies in the lack of suitable datasets that offer both dynamic motion and side-view supervision. Existing datasets either have dynamic camera motion without side-view ground truth (e.g., NVIDIA [35]) or include rotated views without camera motion (e.g., DyNeRF [11]). This gap stems from the difficulty of capturing multi-view dynamic scenes in the real world. To address this, we introduce SynDM, a GTA V-based dataset with a novel dynamic camera dome that enables synchronized main-view motion and side-view supervision. Our evaluation focuses on SynDM, with only qualitative comparisons on existing datasets. Our main contributions are:

- We identify and characterize the limitations of current monocular dynamic NeRFs in rendering from significantly deviated viewpoints, highlighting their inability to preserve structure and appearance consistency under angular shifts
- We propose **ExpanDyNeRF**, a novel dynamic NeRF architecture that incorporates pseudo-novel view supervision using Gaussian splatting priors, enabling reliable synthesis at large viewpoint deviations
- We introduce **SynDM**, the first synthetic dataset for dynamic monocular NeRFs with paired primary and rotated views, captured via a custom GTA V pipeline to benchmark novel view synthesis under controlled angular perturbations
- We perform comprehensive experiments on the SynDM, DyNeRF, and NVIDIA datasets, demonstrating improved perceptual quality and geometric consistency over previous dynamic NeRF methods, especially when handling large viewpoint deviations

Figure 1. Comparison of novel-view rendering from constrained-pose monocular video. Leading dynamic NeRF and Gaussian splatting methods—DetRF [39], RoDynRF [19], DecNeRF [36], and D3DGS [34]—all suffer from artifacts and depth errors at viewpoints distant from the training pose, whereas ExpanDyNeRF (ours) maintains accurate shape and color consistency. The first column shows an example frame from the training video; the remaining columns show novel-view renderings obtained by rotating that frame's camera pose.

2. Related Work

NeRF-based Dynamic Novel View Synthesis. NeRF algorithms have emerged as a powerful approach for high-quality 3D scene reconstruction from sparse images. The original NeRF [15] represents a scene using a fully connected neural network that maps 3D coordinates and viewing directions to color and density, enabling novel view synthesis via volume rendering. While NeRF excels in static scenes, recent work has extended it to dynamic settings. Dynamic NeRF methods, including HyperNeRF [17], DetRF [39], and DecNeRF [36], introduce temporal modeling to capture object motion and appearance variations, often leveraging motion fields and temporal consistency constraints. Despite their effectiveness, dynamic NeRFs remain computationally expensive and typically require densely sampled temporal data. For instance, Dynibar [12] reports training times exceeding one day on 8 A100 GPUs for a single scene, underscoring the scalability challenges of dynamic NeRFs for large-scale or real-time applications.

Gaussian Splatting-based Dynamic Novel View Synthesis. 3D Gaussian splatting [9] offers an efficient alternative to NeRF by representing scenes with Gaussian blobs, enabling real-time, high-resolution rendering. Recent methods like 4DGS [27] and D3DGS [34] extend this approach to dynamic scenes using monocular inputs, capturing non-rigid motion via deformation fields. While NeRF provides high-fidelity reconstructions and Gaussian splatting excels in speed, both struggle with novel view synthesis from deviated angles in monocular settings. Our method combines their strengths to overcome these limitations.

3. Method

We present ExpanDyNeRF, a monocular dynamic NeRF framework for synthesizing novel views of 3D scenes under large viewpoint deviations. To address the lack of ground truth in monocular settings, we combine a two-branch dynamic NeRF with pseudo-supervision from 3D Gaussian priors. Section 3.1 details the dynamic NeRF backbone, Section 3.2 describes pseudo-novel view supervision using 3D Gaussian priors, and Section 3.3 introduces our SynDM dataset with multi-view dynamic scenes.

3.1. ExpanDyNeRF Model Architecture

Our preliminary experiments indicate that NeRF provides greater visual consistency than Gaussian Splatting, especially for distant elements like the sky (shown in Fig. 3). Therefore, we adopt NeRF as the backbone of our model.

NeRF vs Gaussian Splatting Analysis: While both methods appear visually consistent in primary views, significant differences emerge in side view reconstruction. As demonstrated in Fig. 3, 3DGS introduces substantial artifacts in side views, with the sky being incorrectly reconstructed as nearby structures, obstructing distant background elements including mountains and bushes. In contrast, Instant-NGP (NeRF-based) retains the expected characteristics of the sky as distant and uniform, achieving higher fidelity and richer scene details. Accordingly, we adopt a NeRF backbone for its stability under large viewpoint rotations, while using Gaussian Splatting only as a prior generator rather than a rendering backbone. We do not claim NeRF to be universally superior, but find this combination effective for the targeted setting.

Following the architecture in [39], we employ two inter-

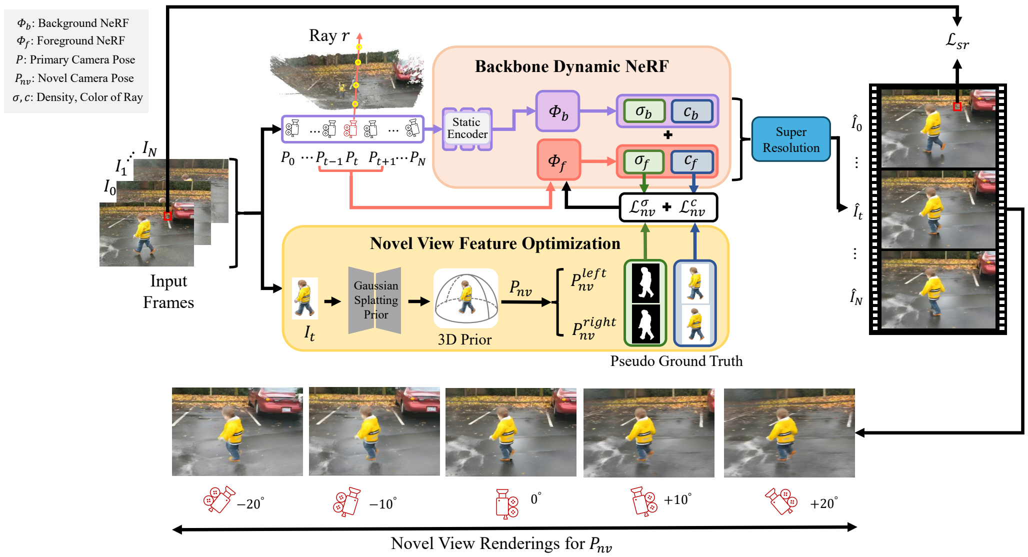

Figure 2. The ExpanDyNeRF architecture is structured into two main components: **(1) Backbone dynamic NeRF model** that processes rays to extract density (σ) and color (c) features from both background (Φ_b) and foreground (Φ_f) models, generating rendering predictions $\hat{I}_t$ from primary camera positions, supervised by the super-resolution loss $\mathcal{L}_{sr}$ **(2) Novel View Feature Optimization** uses the Gaussian Splatting based method to generate a 3D prior for each frame, facilitating the optimization of density and color features via pseudo ground truth for novel views. This includes updating σ_f and c_f using novel view loss metrics $\mathcal{L}^{\sigma}_{nv}$ and $\mathcal{L}^{c}_{nv}$, respectively, enhancing feature representation across different perspectives.

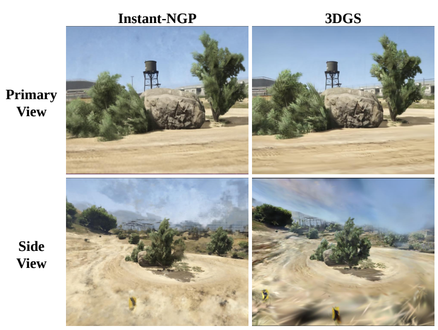

Figure 3. Comparison between Instant-NGP [16] and 3DGS [9] from primary and side views. While both methods appear consistent in primary views (top), 3DGS introduces significant artifacts in side views (bottom right), whereas Instant-NGP maintains better reconstruction quality.

twined neural networks: Φ_b for modeling the static background and Φ_f for the dynamic foreground. As illustrated in Fig. 2, our framework is structured into two main components: (1) a backbone dynamic NeRF model that processes rays to extract density (σ) and color (c) features from both background and foreground models, generating rendering predictions $\hat{I}_t$ from primary camera positions, supervised by the super-resolution loss $\mathcal{L}_{sr}$; (2) novel view feature optimization that uses Gaussian splatting priors to generate 3D representations for each frame, facilitating optimization of density and color features via pseudo ground truth for novel views. This includes updating σ_f and c_f using novel view loss metrics $\mathcal{L}^{\sigma}_{nv}$ and $\mathcal{L}^{c}_{nv}$, respectively, enhancing feature representation across different perspectives.

Preliminaries: We use N video frames $I_t, t \in [1, N]$, to reconstruct a point cloud and estimate primary camera poses P. For each pose $P_t \in P$, we compute ray trajectories to sample points $\mathbf{x} = (x, y, z)$ along ray r at time t in direction $\mathbf{d}$. The sampling process is defined as $\mathcal{F}(P_t) \rightarrow (\mathbf{x}, \mathbf{d})$, linking the camera orientation to the sampled points.

Static Background Representation: The static background module Φ_b takes all N frames and encodes static scene components using a distribution-based representation. This improves alignment between camera projections and the background geometry, allowing simplified renderings of low-variance static features. The module, $\Phi_b(\mathcal{F}(P), \theta) \rightarrow (c_b, \sigma_b)$, predicts color c_b and density σ_b of spatial points from all poses in P, using a prior distribution ($\theta \sim P_{\Theta}(\theta)$).

Dynamic Foreground Representation: The dynamic foreground module Φ_f captures temporal variation using a three-frame sliding window. It integrates spatial coordinates, viewing directions, and the timestamp t: $(\Phi_f(\mathcal{F}(P_t), t) \rightarrow (c_f, \sigma_f))$. Temporal consistency is maintained by encoding time t directly and using optical flow to estimate scene flow, predicting future states of dynamic objects. Continuity constraints are applied to maintain smooth attribute transitions across frames, expressed as: $\mathcal{L}_{cont} = \sum \|\sigma_f(t+1) - \sigma_f(t)\|^2$.

Primary View Reconstruction Loss: The system employs a reconstruction loss to optimize Φ_b and Φ_f by minimizing the discrepancies between the features $\hat{C}(r)$ from rendered images and $C(r)$ from the ground truth images, defined as $\mathcal{L}_{rec} = \sum_{i=1}^{N} \sum_{r \in R} \|\hat{C}(r) - C(r)\|_2^2$. This

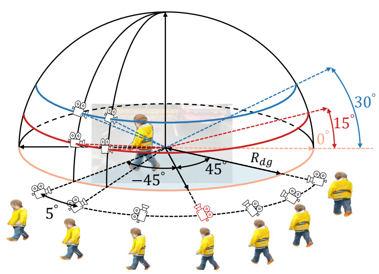

Figure 4. Pseudo ground truth generation demonstration. Novel viewpoints (black cameras) are sampled around a dome centered on the foreground object, with the primary view shown in red. Example renderings and shape masks are shown for different viewing angles.

loss ensures the renderings from the primary views closely match the ground truth frames, setting up a baseline for the following novel view optimization.

Super-Resolution Loss: Inspired by SOTA super-resolution methods [22, 23], we incorporate a super-resolution loss ($\mathcal{L}_{sr}$) to enhance image quality. Rendered patches from ExpanDyNeRF are processed by a pre-trained super-resolution model, which preserves fine textures while increasing resolution. The loss is computed by comparing sampled patches from the predicted and corresponding high-resolution reference. The formulation of $\mathcal{L}_{sr}$ is:

$$\mathcal{L}_{sr} = \sum_{k=1}^{K} \left\| \hat{Q}_k - Q_k \right\|_1 + \sum_{k=1}^{K} \sum_{l} \lambda_l \left\| F_{vgg}^{l}(\hat{Q}_k) - F_{vgg}^{l}(Q_k) \right\|_1 .$$

Here, $\hat{Q}_k$ and Q_k represent the super-resolution prediction and reference patches, respectively, F_{vgg}^{l} is a set of layers in a pretrained VGG-19 feature extractor, and λ_l is the reciprocal of the number of neurons in layer l, combining reconstruction and perceptual losses.

3.2. Pseudo Ground Truth Optimization Strategy

Through empirical experiments, we observed that foreground objects appear blurrier than the background during viewpoint rotation. This is due to affine distortion, where nearby objects exhibit greater apparent motion than distant ones. To address this, we prioritize optimizing foreground representations. Our method leverages FreeSplatter [25] to generate high-quality 3D mesh priors, enabling pseudo-ground truth supervision from novel viewpoints.

Pseudo Ground Truth Generation for Novel Views: For each input frame I_t, we construct a 3D Gaussian prior representing the foreground object in its local coordinate frame. Centered around this object, we define a dome-shaped sampling space, with a radius R_d as shown in Fig. 4. The radius R_d represents the distance from the primary viewpoint to the object. The position of P_t on the dome is denoted as $(elevation = e, azimuth = 0, radius = R_d)$, where e corresponds to the elevation angle of the primary recording view.

Novel Viewpoint Sampling Strategy: We systematically sample novel viewpoints by varying both azimuth and elevation angles while maintaining the fixed radius R_d. Specifically, we generate viewpoints spanning azimuth angles from $-45°$ to $45°$ in $5°$ increments, at three elevation levels: $0°$, $15°$, and $30°$. The forward vector of all camera poses $P_{nv}^{(d)}$ points towards the dome center, ensuring consistent object framing across viewpoints. From these novel camera poses, we render pseudo ground truth images that encode the expected density and color distributions of the foreground object. The resulting renderings provide both RGB color and corresponding shape masks at different viewing angles, creating comprehensive supervision that would be impossible to obtain from real monocular capture.

Mapping Novel Views to the NeRF Coordinate System: To apply pseudo ground-truth supervision at novel viewpoints, we transform the sampled camera poses from the Gaussian prior coordinate system to the NeRF coordinate system using a rigid alignment matrix. Let the primary camera pose P_t in the foreground NeRF coordinate system be $P_t^{(n)}$, and its corresponding pose in Gaussian prior coordinate system be $P_t^{(d)}$. The transformation matrix T that aligns the two coordinate systems is computed as: $T = P_t^{(n)} \cdot (P_t^{(d)})^{-1}$. We apply this transformation, T, to each novel view camera pose, $P_{nv}^{(d)}$, sampled in the Gaussian prior coordinate system, to transfer all new camera positions to the foreground NeRF coordinate system:$P_{nv} = \{P \cdot T, \forall P \in P_{nv}^{(d)}\}$.

Novel View Loss: During each training iteration, two symmetric novel views are randomly sampled per frame from P_{nv}. For each selected novel viewpoint, a set of rays $\mathbf{R}_{nv}$ is sampled from the camera pose. Color and density predictions in the foreground NeRF are obtained by evaluating the network Φ_f on sampled ray $(\mathcal{F}(P_{nv}), t)$, producing outputs (c_f, σ_f). Specifically, $\mathcal{F}(P_{nv}) \rightarrow (\mathbf{x}, \mathbf{d})$ samples points along a ray $r_{nv} \in \mathbf{R}_{nv}$. We then integrate the predicted color and density values along each ray r_{nv} to produce the corresponding pixel-wise predictions $\hat{C}_f(r_{nv})$ and $\hat{\sigma}_f(r_{nv})$. The corresponding novel view loss is defined as:

$$\begin{aligned}\mathcal{L}_{nv} &= \mathcal{L}_{nv}^{c} + \mathcal{L}_{nv}^{\sigma} \\ &= \sum_{r \in R_{nv}} \left(\|\hat{C}(r_{nv}) - C(r_{nv})\|_2^2 + \|\hat{\sigma}(r_{nv}) - \sigma(r_{nv})\|_2^2 \right).\end{aligned}$$

where $C(r_{nv})$ and $\sigma(r_{nv})$ denote the pseudo ground truth color and density values for ray r_{nv}, respectively. This loss encourages the model to match the rendered appearance and structure of the pseudo-supervised novel views, improving generalization to unseen angles. A further explanation and visualization of ray sampling strategies are detailed in Fig. 8 in the supplementary. To manage the risk of exploding gradients early in training, we defer inclusion of $\mathcal{L}_{nv}$ until after a fixed number of epochs. The final total loss function is given by:

$$\mathcal{L} = \mathcal{L}_{cont} + \mathcal{L}_{rec} + \mathcal{L}_{sr} + \mathcal{L}_{nv}.$$

3.3. Synthetic Dynamic Multiview (SynDM) Dataset

To enable quantitative evaluation of novel view synthesis under significant viewpoint deviations, we introduce our SynDM dataset, built using the high-fidelity simulation platform GTA V. The game offers rich dynamic environments and realistic rendering, making it an ideal foundation. However, a core limitation of GTA V is its support for only a single active viewport, which posing a challenge for synchronized multi-view dynamic scene capture. We extend the GTAV-TeFS [14] framework–originally developed for dual-camera capture –into a generalized multi-camera pipeline to simultaneously support both monocular primary camera capture and multi-view stereo camera collection in GTA V's dynamic environment. Traditionally, collecting data from multiple camera views in a single-viewport engine requires frame swapping, where each camera is rendered sequentially. Under a 60 Hz refresh rate, this results in a latency of at least 16.7 ms per camera swap. While acceptable for static scenes, this approach quickly breaks down in dynamic settings and as we add cameras, as the accumulated latency introduces motion misalignment and temporal artifacts. To address this we developed a custom plugin that semi-freezes the game's graphical state while allowing the rendering and physics engine to continue running. This design enables us to cycle through camera views in a controlled and consistent manner during a single logical frame. With precise scheduling, we reduced the per-swap latency from 16.7ms to just 0.2ms, making high-resolution, low-latency multi-view capture of dynamic scenes possible.

Our dataset enables synthetic object tracking via synchronized multi-view recordings, offering a robust ground truth for evaluating dynamic NeRF-based novel view synthesis. It consists of nine distinct scenes spanning three categories–humans, vehicles, and animals (Fig. 5). Each scene is captured using 22 cameras: 19 are distributed horizontally around a reference point at $5°$ intervals from $-45°$ to $45°$, including a central anchor camera, while the remaining three are elevated vertically at $-45°$, $0°$, and $45°$. All frames are rendered at a resolution of 1920×1080 with a $90°$ horizontal and $59°$ vertical field of view.

Necessity and Advantage of Proposed SynDM Dataset As shown in Table 1, existing datasets lack critical features for evaluating dynamic novel view synthesis under large viewpoint deviations. Most importantly, no real-world dataset provides deviated view ground truth, severely limiting quantitative assessment beyond primary viewpoints. Our SynDM dataset uniquely combines all essential features, including multi-view data, deviated GT, full-scene representation, and camera motion, enabling comprehensive evaluations previously impossible with existing datasets alone.

Table 1. Comparison of dynamic 3D reconstruction datasets. Columns: Multi-view (MV), Deviated View Ground Truth (Deviated View GT), Unconstrained Scene (Unconst. Scene), Camera Motion (Cam. Motion), and Background (Bkg.). SynDM is the only dataset supporting all five attributes.

Dataset	MV	Deviated View GT	Unconst. Scene	Cam. Motion	Bkg.
DAVIS [18]	✗	✗	✓	✓	✓
iPhone [2]	✗	✗	✓	✓	✓
NeRFDS [33]	✗	✗	✓	✓	✓
NVIDIA [35]	✗	✗	✓	✓	✓
HyperNeRF [17]	✗	✗	✗	✓	✓
DyNeRF [11]	✓	✓	✗	✗	✓
ActorsHQ [8]	✓	✓	✗	✗	✗
Multi-face [29]	✓	✓	✗	✗	✗
SynDM (Ours)	✓	✓	✓	✓	✓

Figure 5. Gallery of images from our SynDM dataset, showcasing a variety of subjects. Animals are featured in the top row, humans in the middle row, and vehicles in the bottom row.

4. Experimental Results

We present a comprehensive evaluation of ExpanDyNeRF against state-of-the-art dynamic novel view synthesis methods. Our evaluation strategy progresses from controlled synthetic environments to challenging real-world scenarios, demonstrating robustness across diverse settings. More results can be find in our Supplementary Material.

4.1. Experimental Setup

Datasets. We evaluate ExpanDyNeRF on three datasets with complementary characteristics. Our **SynDM Dataset** provides complete ground truth for quantitative analysis across three scene categories (human, animals, vehicles) spanning rural and urban environments within GTA V simulation. For training, we use the first 24 frames from each scene and evaluate on 12 novel views, uniformly sampled between $-30°$ to $+30°$ at $5°$ intervals. The **DyNeRF Dataset** [11] offers real-world dynamic scenes with multi-view ground truth, enabling quantitative evaluation under challenging viewing conditions. The **NVIDIA Dataset** [35] provides real-world monocular sequences for generaliza-

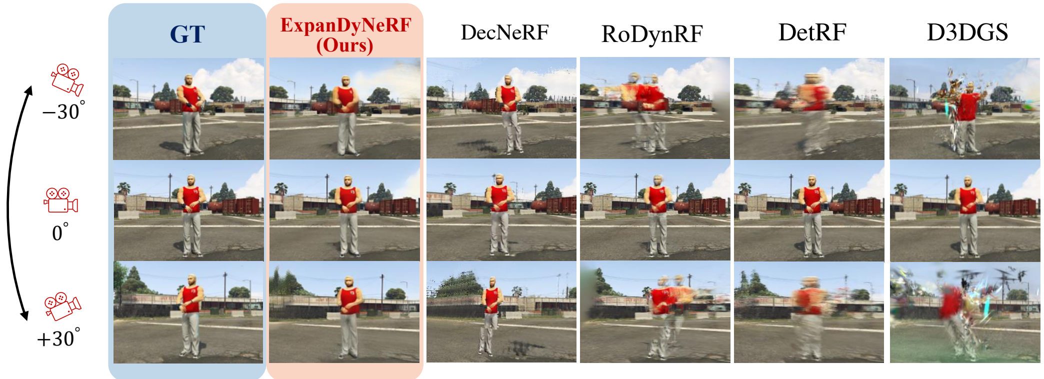

Figure 6. Comparison of dynamic NeRF models on SynDM dataset. Each column shows a method's performance across different rotation angles, with ground truth in blue. Red boxes highlight key differences, showing ExpanDyNeRF's superior color and shape fidelity.

Table 2. Quantitative results on SynDM; best scores are bolded, second-best scores are in blue.

Method	Human			Animal			Vehicle			Average		
	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓
D3DGS [34]	87.83	16.97	0.305	315.3	18.55	0.272	267.8	14.70	0.543	223.6	16.74	0.373
RoDynRF [13]	167.3	19.66	0.318	262.0	21.00	0.302	285.3	16.31	0.395	238.2	18.99	0.338
DecNeRF [36]	178.5	17.46	0.441	287.3	15.11	0.545	211.3	14.76	0.562	225.7	15.78	0.516
DetRF [39]	200.0	21.27	0.408	290.6	22.69	0.378	226.4	17.07	0.521	239.0	20.34	0.436
ExpanDyNeRF (Ours)	**85.61**	**21.71**	**0.182**	**155.8**	**23.66**	**0.142**	**186.7**	**17.21**	**0.341**	**142.7**	**20.86**	**0.209**

tion assessment, though without deviated viewpoint ground truth (see Table 1 for detailed comparison).

Implementation Details. Implementation Details. ExpanDyNeRF is trained with per-scene optimization on 2 A100 GPUs for 300k iterations (approximately 15 hours per scene). The training cost is primarily dominated by NeRF optimization with pseudo-novel view supervision, while Gaussian prior generation is performed once per frame as a preprocessing step and incurs marginal overhead. Consistent with prior dynamic NeRF approaches, our method prioritizes rendering fidelity under large viewpoint deviations over training efficiency. Loss coefficients are set to: $\lambda_{nv}^{c} = 1.0$, $\lambda_{nv}^{\sigma} = 0.1$, and $\lambda_{sr} = 0.5$. All other parameters follow [38].

PSNR Limitation Analysis. Our experiments reveal important limitations of PSNR in evaluating perceptual quality. Despite ExpanDyNeRF producing sharper, more detailed reconstructions, PSNR scores may paradoxically be lower due to localized high-density errors highlighted in pixel-wise error heatmaps. In contrast, baseline methods with blurry results yield smoother transitions, leading to lower MSE despite poorer visual quality. This occurs because blurry regions (e.g., object boundaries, fine textures) blend into backgrounds, minimizing MSE contributions (shown in Fig. 9 in the supplementary). This phenomenon demonstrates why complementary perceptual metrics like LPIPS and FID are essential for comprehensive quality assessment, particularly when evaluating sharpness and clarity improvements.

4.2. Evaluation on SynDM Dataset

The SynDM dataset enables comprehensive quantitative evaluation with complete ground truth across diverse dynamic scenes. We compare ExpanDyNeRF against four state-of-the-art methods using PSNR, LPIPS, and FID metrics (Table 2).

Quantitative Results. ExpanDyNeRF achieves the highest PSNR score of 20.86 while producing the sharpest renderings. Although DetRF also reports competitive PSNR despite producing heavily blurred images, this underscores the importance of complementary perceptual metrics. Our model significantly outperforms competing methods in both LPIPS (38% lower than second-best) and FID (36% lower than second-best), demonstrating superior perceptual alignment and distributional consistency.

Qualitative Analysis. Fig. 6 illustrates ExpanDyNeRF's superior shape coherence and color stability in dynamic regions. Each baseline method exhibits distinct failure patterns: DecNeRF renders sharp details but suffers from poor depth perception, resulting in flat, cardboard-like appearances that fail to maintain 3D structure consistency. D3DGS introduces depth inconsistencies causing foreground objects to fracture under rotation, particularly evident after 30° viewpoint changes where object parts appear disconnected. RoDynRF struggles with consistent object placement, often producing floating artifacts or mis-

Table 3. Quantitative results on DyNeRF dataset. We show Coffee and Beef scenes which represent the two primary scenarios in this dataset as shown in the first and second scenes of DyNeRF in Fig. 1. Best scores are bolded, second-best scores are in blue.

Method	Coffee			Beef		
	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓
D3DGS [34]	178.5	28.45	0.227	172.8	32.78	0.268
RoDynRF [13]	156.3	29.13	0.268	148.2	33.45	0.252
DecNeRF [36]	152.8	28.87	0.304	155.4	33.12	0.285
DetRF [39]	165.2	29.52	0.275	151.3	33.99	0.259
ExpanDyNeRF (Ours)	**132.4**	**30.32**	**0.189**	**135.8**	**34.92**	**0.195**

Table 4. Quantitative comparison on NVIDIA dynamic scenes. We show Skate and Truck scenes which correspond to the first and second scenes of NVIDIA dataset in Fig. 1. Best is bold; second-best is blue.

Method	Skate			Truck		
	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓
D3DGS [34]	142.8	25.67	0.258	95.7	26.39	0.239
RoDynRF [13]	103.5	27.89	0.087	78.4	29.13	0.063
DecNeRF [36]	112.9	26.83	0.134	85.6	27.56	0.115
DetRF [39]	**84.34**	**29.45**	**0.072**	**67.69**	**31.75**	**0.041**
ExpanDyNeRF (Ours)	90.83	28.91	0.079	69.37	30.60	0.034

aligned body parts. These failure modes stem from insufficient side-view supervision during training. Further visualizations of results are provided in Fig. 10 and 11 in the supplementary.

4.3. Evaluation on DyNeRF Dataset

The DyNeRF dataset provides real-world dynamic scenes with multi-view ground truth, but presents challenges for monocular training due to its stationary camera setup. Unlike our SynDM dataset, which provides natural camera motion, DyNeRF's fixed camera positions prevent direct COLMAP reconstruction from monocular sequences. To address this limitation, we construct synthetic monocular sequences by selecting frames from central cameras (cam0, cam4, cam5, cam6) across different timestamps, creating the necessary camera motion for COLMAP initialization. We then adopt a holdout validation strategy, using these central cameras for training while reserving the geometrically challenging outer cameras (cam01 and cam10) as test views. These cameras are specifically selected as the most deviated viewpoints from the training set's central viewing positions.

Quantitative Results. Table 3 shows our method's consistent superiority across all baseline methods on representative DyNeRF scenes. ExpanDyNeRF achieves the best performance across all metrics in both scenes. Our method demonstrates particularly strong performance on the Coffee (FID: 132.4, PSNR: 30.32, LPIPS: 0.189) and Beef (FID: 135.8, PSNR: 34.92, LPIPS: 0.195) sequences, with consistent improvements across all three evaluation metrics in both representative scenarios.

Qualitative Analysis. As shown in Fig. 1, our method maintains superior performance on DyNeRF scenes. Leading dynamic NeRF and Gaussian splatting methods, including DetRF, RoDynRF, DecNeRF, and D3DGS, all suffer from artifacts and depth errors at viewpoints distant from the training pose, whereas ExpanDyNeRF maintains accurate shape and color consistency. The improvements are especially significant for challenging viewpoints that demand accurate geometric reasoning due to their large displacement from the training views, confirming that our pseudo-ground-truth supervision strategy successfully addresses fundamental geometric consistency challenges in dynamic 4D scene reconstruction.

4.4. Generalization to NVIDIA Dataset

The NVIDIA dataset contains real-world monocular sequences captured by stationary cameras and serves to evaluate generalization under practical capture conditions. However, it is not well suited for assessing large viewpoint deviations: its 12 cameras form a compact frontal array with limited angular separation, providing no ground-truth supervision for challenging side views. As a result, unlike DyNeRF and SynDM, quantitative evaluation on NVIDIA is restricted to modest viewpoint changes. Performance on this dataset should therefore be viewed as a complementary indicator of real-world robustness rather than a definitive benchmark for large-angle novel view synthesis.

Quantitative Analysis. Table 4 reports competitive results on representative NVIDIA scenes. Although our method does not achieve the highest scores, this is largely attributable to the dataset bias toward primary viewpoints, as the same 12 stationary front-facing cameras are used for both training and testing. ExpanDyNeRF performs strongly on Skate (FID: 90.83, PSNR: 28.91, LPIPS: 0.079) and Truck (FID: 69.37, PSNR: 30.60, LPIPS: 0.034), consistently ranking second-best. DetRF attains superior performance on these scenes due to its emphasis on depth estimation from forward-facing views, whereas our more balanced performance across viewpoints indicates improved stability beyond the primary training poses. This trade-off aligns with our design goal of robustness over dataset-specific optimization.

Qualitative Results. Fig. 1 highlights failure cases of existing methods under camera rotation. We compare against DetRF [39], RoDynRF [19], DecNeRF [36], and D3DGS [34]. All baselines exhibit artifacts or depth errors at viewpoints far from training poses. RoDynRF often preserves local grounding, such as foot placement, but suffers from global misalignment, with upper bodies appearing anchored to background structures in the skate scene. DecNeRF produces cardboard-like foregrounds or complete object disappearance in side views due to insufficient depth

Figure 7. The images above, rendered from a view rotated by -30 degrees, illustrate the impact of different loss functions on the quality of novel view synthesis. The best performance is achieved when all loss functions ($L_{nv}^{\sigma} + L_{nv}^{c} + L_{\text{sr}}$) are applied simultaneously, highlighting the complementary role each loss plays in enhancing rendering quality and achieving a closer match to the ground truth.

Table 5. Quantitative Evaluation of optimization strategies on the SynDM Dataset. Baseline is without any optimization, L_{nv}^{σ} only uses density optimization, L_{nv}^{c} only applies color optimization, and $L_{nv}^{\sigma} + L_{nv}^{c}$ is trained with both. $L_{nv}^{\sigma} + L_{nv}^{c} + L_{sr}$ includes all modules including super resolution.

Method	**Human**			**Animal**			**Vehicle**			**Average**		
	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓
Baseline	200.0	21.27	0.408	290.6	22.69	0.378	226.4	17.07	0.521	239.0	20.34	0.436
L_{nv}^{σ}	158.2	21.56	0.399	237.8	23.57	0.374	186.7	16.69	0.524	194.2	20.61	0.432
L_{nv}^{c}	147.5	20.10	0.395	162.8	19.88	0.442	210.9	16.17	0.525	173.7	18.72	0.454
L_{sr}	147.8	20.85	0.212	194.6	21.30	0.183	176.6	16.62	0.331	173.0	19.59	0.242
$L_{nv}^{\sigma} + L_{nv}^{c}$	146.2	21.66	0.395	207.9	**23.69**	0.339	198.8	17.14	0.537	184.3	20.83	0.424
$L_{nv}^{\sigma} + L_{nv}^{c} + L_{sr}$	**85.61**	**21.71**	**0.182**	**155.8**	23.66	**0.142**	**144.9**	**17.21**	**0.304**	**142.7**	**20.86**	**0.209**

estimation under oblique angles. D3DGS shows spatial fragmentation caused by the lack of side-view supervision. In contrast, ExpanDyNeRF maintains consistent shape and color across rotation angles, effectively mitigating fractured reconstructions and temporal instability. Additional NVIDIA results are provided in Figs. 12 and 13 in the supplementary material.

4.5. Ablation Study

We conduct a comprehensive ablation study to validate the contribution of each component in our optimization strategy. The analysis examines the effects of our novel view losses and super-resolution enhancement.

Component Analysis. Fig. 7 illustrates the effects of different loss combinations on novel view synthesis. Without novel view supervision, the baseline produces blurred renderings with poorly defined shapes, highlighting the importance of side-view guidance. Applying only the color-based novel view loss L_{nv}^{c} alleviates blurring but introduces white artifacts due to missing density-based shape constraints. In contrast, using only L_{nv}^{σ} preserves object geometry but leads to faded or distorted colors. Combining L_{nv}^{c} and L_{nv}^{σ} yields clear improvements in structural integrity and temporal stability, though fine textures remain limited. The super-resolution loss L_{sr} effectively enhances texture sharpness but relies on the geometric consistency provided by novel view losses. When used alone, L_{sr} introduces artifacts and inconsistencies, indicating that geometric supervision must precede texture refinement. Notably, L_{sr} improves perceptual quality without adding geometric constraints and may amplify existing structural errors in the absence of L_{nv}^{c} and L_{nv}^{σ}.

Quantitative Validation. Table 5 confirms these visual observations across all metrics. The best performance is achieved when integrating L_{nv}^{σ}, L_{nv}^{c}, and super-resolution loss L_{sr}. Notably, using L_{sr} in isolation fails to produce acceptable renderings, underscoring the essential role of novel view supervision in mitigating artifacts during viewpoint rotations. The full combination achieves optimal perceptual quality, confirming our optimization strategy's effectiveness in reducing visual artifacts and enhancing overall realism.

5. Discussion and Conclusion

Limitations. Despite outperforming existing models, ExpanDyNeRF has limitations, particularly in handling extreme viewing angles (beyond 45 degrees) and unseen background generation. Also, ExpanDyNeRF requires per-scene optimization, resulting in higher computational cost compared to feed-forward or purely Gaussian-based methods.

Conclusion. ExpanDyNeRF advances dynamic NeRF by significantly improving novel view synthesis, particularly at wider viewing angles, by extending the range of stable visualization. Our SynDM dataset, based on GTA V for dynamic multiview scenarios, provides a strong foundation for evaluating dynamic scene reconstructions from varied angles. Our evaluations demonstrate ExpanDyNeRF's superior ability to render dynamic scenes.

References

[1] Yaosen Chen, Qi Yuan, Zhiqiang Li, Yuegen Liu, Wei Wang, Chaoping Xie, Xuming Wen, and Qien Yu. Upst-nerf: Universal photorealistic style transfer of neural radiance fields for 3d scene. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

[2] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 5

[3] Xinyu Gao, Ziyi Yang, Yunlu Zhao, Yuxiang Sun, Xiaogang Jin, and Changqing Zou. A general implicit framework for fast nerf composition and rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1833–1841, 2024. 1

[4] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 1

[5] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009. 1

[6] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1

[7] Jiaming Gu, Minchao Jiang, Hongsheng Li, Xiaoyuan Lu, Guangming Zhu, Syed Afaq Ali Shah, Liang Zhang, and Mohammed Bennamoun. Ue4-nerf: Neural radiance field for real-time rendering of large-scale scene. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[8] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 5

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[10] Byeonghyeon Lee, Howoong Lee, Usman Ali, and Eunbyung Park. Sharp-nerf: Grid-based fast deblurring neural radiance fields using sharpness prior. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3709–3718, 2024. 1

[11] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 1, 5

[12] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2

[13] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 6, 7

[14] Yedi Luo, Xiangyu Bai, Le Jiang, Aniket Gupta, Eric Mortin, and Hanumant Singh Sarah Ostadabbas. Temporal-controlled frame swap for generating high-fidelity stereo driving data for autonomy analysis. *arXiv preprint arXiv:2306.01704*, 2023. 5

[15] 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

[16] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 3

[17] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2, 5

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5

[19] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20626–20636, 2023. 2, 7

[20] Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4881–4891, 2023. 1

[21] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 1

[22] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 4

[23] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 4

[24] Yifan Wang, Yi Gong, and Yuan Zeng. Hyb-nerf: A multiresolution hybrid encoding for neural radiance fields. In

2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 3677–3686. IEEE, 2024. 1

[25] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free-view synthesis of indoor scenes. *arXiv preprint arXiv:2405.17958*, 2024. 1, 4

[26] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions. *arXiv preprint arXiv:2212.04701*, 2022. 1

[27] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 1, 2

[28] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *arXiv preprint arXiv:2403.11134*, 2024. 1

[29] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 5

[30] Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: neural radiance fields for ultrasound imaging. In *Medical Imaging with Deep Learning*, pages 382–401. PMLR, 2024. 1

[31] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, et al. Vr-nerf: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 1

[32] Yiwei Xu, Tengfei Wang, Zongqian Zhan, and Xin Wang. Mega-nerf++: An improved scalable nerfs for high-resolution photogrammetric images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:769–776, 2024. 1

[33] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8285–8295, 2023. 5

[34] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 1, 2, 6, 7

[35] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 1, 5

[36] Meng You and Junhui Hou. Decoupling dynamic monocular videos for dynamic view synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2, 6, 7

[37] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 1

[38] Boyu Zhang, Wenbo Xu, Zheng Zhu, and Guan Huang. Detachable novel views synthesis of dynamic scenes using distribution-driven neural radiance fields. *arXiv preprint arXiv:2301.00411*, 2023. 6

[39] Boyu Zhang, Zheng Zhu, and Wenbo Xu. Detrf: Detachable novel views synthesis of dynamic scenes using backdrop-driven neural radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9860–9868, 2025. 2, 6, 7

[40] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7178–7186, 2024. 1

[41] Shangzhan Zhang, Sida Peng, Yinji ShenTu, Qing Shuai, Tianrun Chen, Kaicheng Yu, Hujun Bao, and Xiaowei Zhou. Dyn-e: Local appearance editing of dynamic neural radiance fields. *arXiv preprint arXiv:2307.12909*, 2023. 1

Broadening View Synthesis of Dynamic Scenes from Constrained Monocular Videos

Supplementary Material

Overview of Supplementary Materials This supplementary document provides extended experimental results, ablations, and visual comparisons that complement the main paper. In particular, we include detailed ablation studies on our ray sampling and padding strategies (Fig. 8) and error heatmaps illustrating the limitations of standard metrics such as PSNR and MSE (Fig. 9). Furthermore, we include additional novel view synthesis visualizations across both the SynDM and NVIDIA datasets in Figs. 10–13. These results further validate the effectiveness of ExpanDyNeRF in capturing sharper, more consistent scene details across diverse scenarios. For an interactive overview with richer visualizations, readers are encouraged to view the accompanying `IO` page, which showcases additional qualitative examples and video results.

Ray Sampling Strategies We compared various ray sampling strategies for novel view density and color optimization in Equation 3.2. Examples are shown in Fig. 8. Global sampling over the whole frame yields results in Panel (e) similar to the base output in Panel (a), due to the small proportion of dynamic segments in the frame, causing generalized and ineffective updates. Alternate strategies sample within the foreground object's area shown white in Panel (c), which may overlook updates outside this zone. Panel (f) demonstrates that sampling from various viewpoints for dynamic density updates can unintentionally extend beyond the intended mask, causing non-dynamic areas to obscure the background. Panel (g) shows the third strategy where the GaussianBlur [5] expands the foreground boundary, creating a zero gray-scale edge. Sampling within this blurred mask improves results, yet areas adjacent to the person still see undue dynamic density updates beyond the motion mask. Our final strategies focused on ray sampling within the padded area of the mask's bounding box (bounding boxes in Panel (c)), which outperforms the other strategies. Experimentation showed that while larger padding, like 10 pixels in Panel (h), achieves comparable foreground optimization to smaller padding, such as 2 pixels in Panel (d), it adversely affects background clarity.

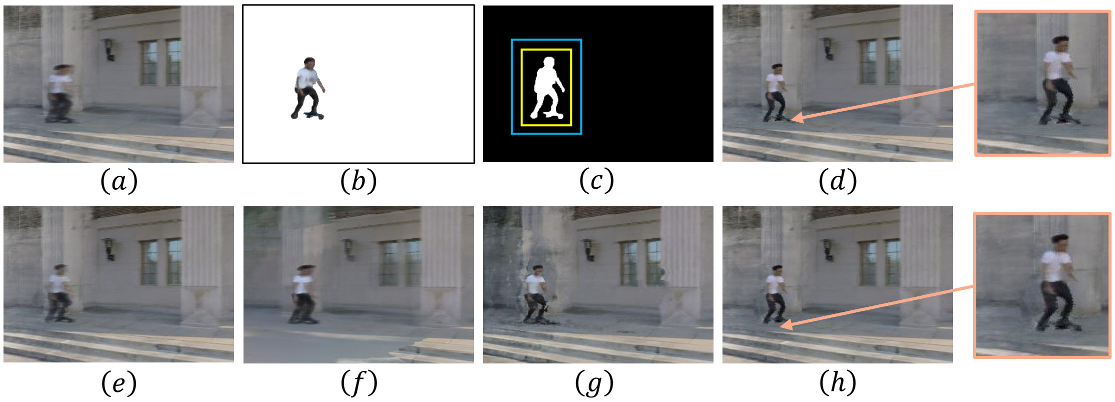

Figure 8. This figure presents an ablation study on our ray sampling strategy. Panel (a) displays the base model's output without optimization for color and density. Panel (b) depicts the pseudo-ground-truth of novel views from the created 3D mesh. Panel (c) illustrates the density mask derived from pseudo ground truth, where the yellow and blue boxes represent the bounding box with 2-pixel padding, and 10-pixel padding, respectively. Panel (e) shows predictions from global ray sampling on the mask, while panel (f) shows predictions from ray sampling within the foreground object area only. Panel (g) demonstrates the GaussianBlur strategy's prediction. Panels (d) and (h) showcase predictions with 2-pixel and 10-pixel padding, respectively. The comparison between panels (d) and (h) reveals that employing 2-pixel padding leads to enhanced quality in reconstructing novel view details with minimum background distortion.

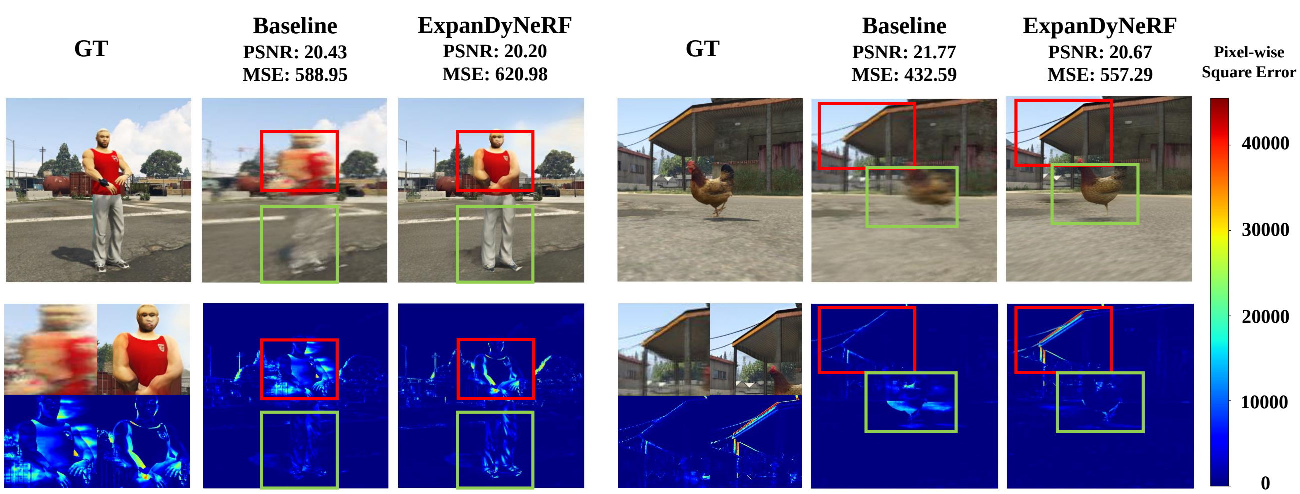

Figure 9. Comparison of visual and quantitative results between the baseline and ExpanDyNeRF models, evaluated using PSNR and MSE metrics. The "GT" column represents the ground truth images. The red and green boxes highlight critical regions of interest for analysis. The red box demonstrates areas with sharp and detailed reconstruction by ExpanDyNeRF, whereas the baseline exhibits significant blur. Despite ExpanDyNeRF producing clearer outputs, PSNR scores are lower due to localized high-density errors (highlighted in the pixel-wise error heatmap). In contrast, the baseline's blurry results yield smoother transitions, leading to lower MSE despite poorer visual quality. The green box further illustrates how blurry regions (e.g., human pants or chicken body) blend into the background, minimizing MSE contributions. This demonstrates the limitation of PSNR in capturing perceptual quality, particularly when evaluating sharpness and clarity.

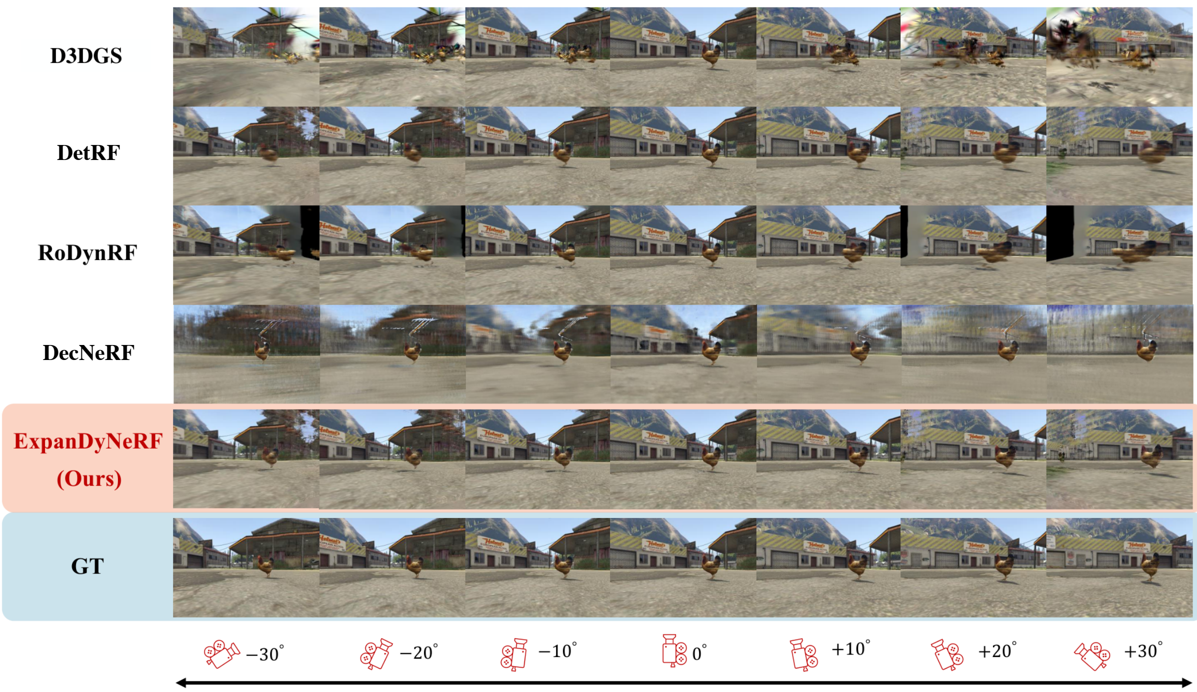

Figure 10. This figure presents comparison on the novel view synthesis performance of leading dynamic NeRF models and our ExpanDyNeRF training on the animal data from our SynDM dataset.

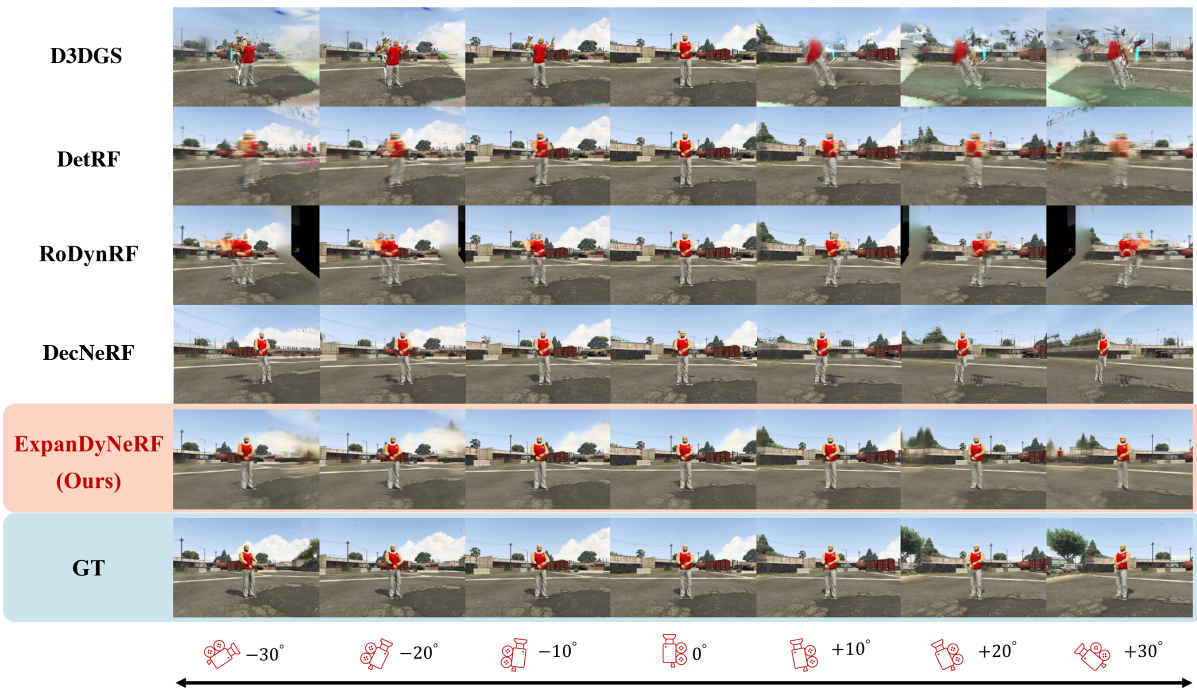

Figure 11. This figure presents comparison on the novel view synthesis performance of leading dynamic NeRF models and our ExpanDyNeRF training on the human data from our SynDM dataset.

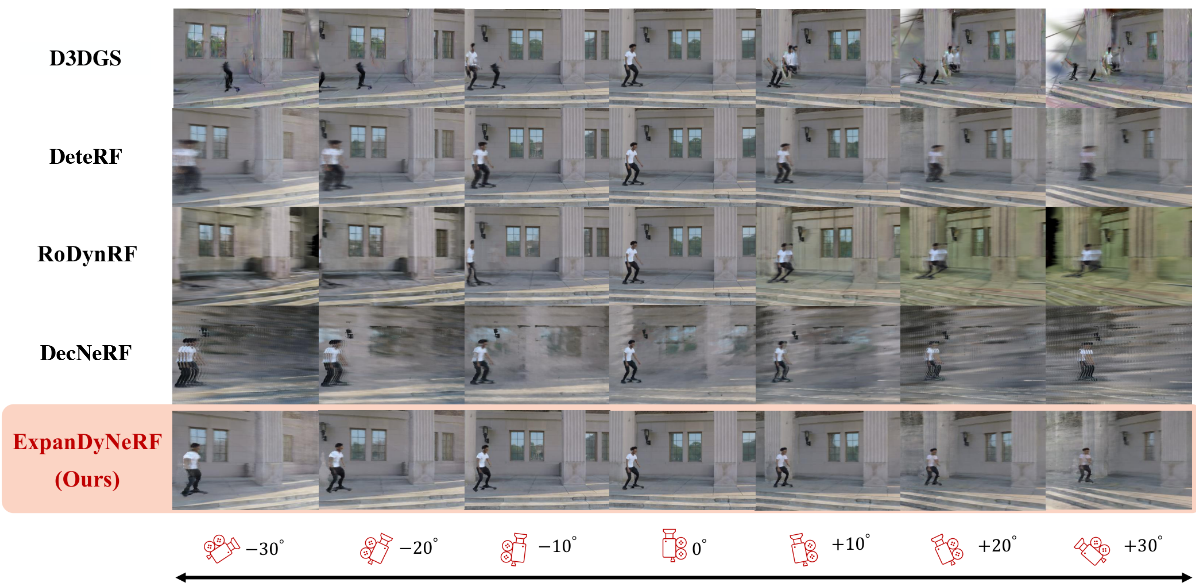

Figure 12. This figure presents comparison on the novel view synthesis performance of leading dynamic NeRF models and our ExpanDyNeRF training on the truck data from the NVIDIA dataset.

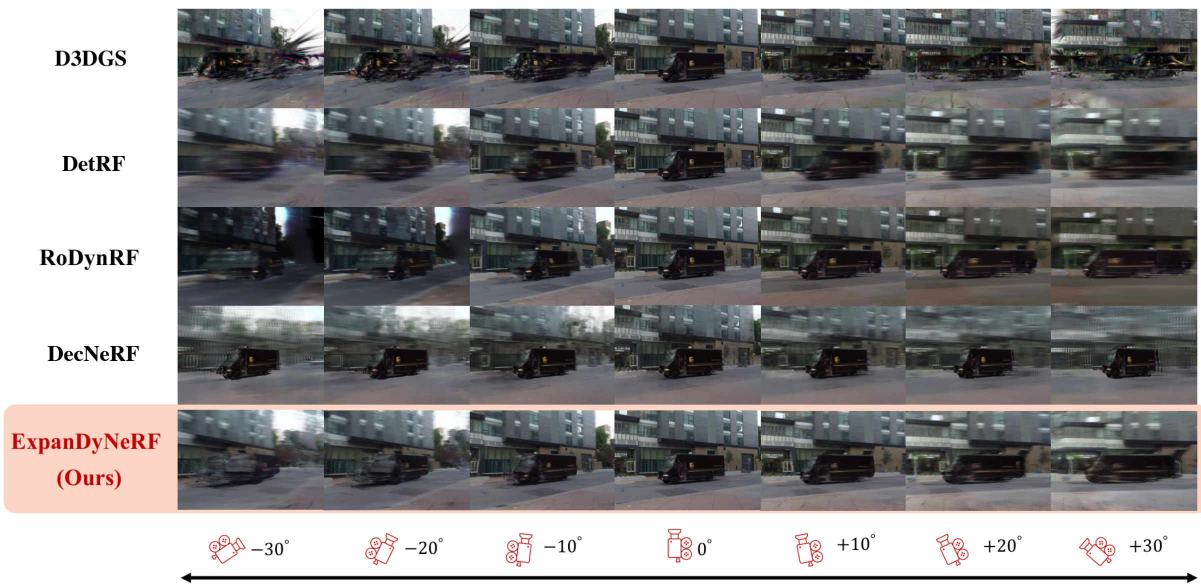

Figure 13. This figure presents comparison on the novel view synthesis performance of leading dynamic NeRF models and our ExpanDyNeRF training on the skating data from the NVIDIA dataset.