

FLAG-4D: Flow-Guided Local-Global Dual-Deformation Model for 4D Reconstruction

Guan Yuan Tan^{1*}, Ngoc Tuan Vu^{1*}, Arghya Pal¹, Sailaja Rajanala¹, Raphaël C.-W. Phan¹, Mettu Srinivas², Chee-Ming Ting¹

¹ Monash University

² National Institute of Technology Warangal

Abstract

We introduce FLAG-4D, a novel framework for generating novel views of dynamic scenes by reconstructing how 3D Gaussian primitives evolve through space and time. Existing methods typically rely on a single Multilayer Perceptron (MLP) to model temporal deformations, and they often struggle to capture complex point motions and fine-grained dynamic details consistently over time, especially from sparse input views. Our approach, FLAG-4D, overcomes this by employing a dual-deformation network that dynamically warps a canonical set of 3D Gaussians over time into new positions and anisotropic shapes. This dual-deformation network consists of an Instantaneous Deformation Network (IDN) for modeling fine-grained, local deformations, and Global Motion Network (GMN) for capturing long-range dynamics, refined via mutual learning. To ensure these deformations are both accurate and temporally smooth, FLAG-4D incorporates dense motion features from a pretrained optical flow backbone. We fuse these motion cues from adjacent timeframes and use a deformation-guided attention mechanism to align this flow information with the current state of each evolving 3D Gaussian. Extensive experiments demonstrate that FLAG-4D achieves higher-fidelity and more temporally coherent reconstructions with finer detail preservation than state-of-the-art methods.

Code — <https://github.com/tgy1221/FLAG-4D>

Introduction

4D reconstruction has become a crucial advancement for capturing and reconstructing dynamic real-world objects and scenes, incorporating the temporal dimension t within the traditional 3D spatial coordinates (x, y, z) . This approach enables the modeling of continuous changes, movements, and deformations in objects and environments over time, an essential feature for applications in Augmented Reality (AR) and Virtual Reality (VR), where accurate motion capture enhances immersive experiences. Recent works in Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) have been extended to enable 4D reconstruction by integrating time-dependent deformation networks to capture and predict scene

changes across temporal frames (Das et al. 2024; Katsumata, Vo, and Nakayama 2023; Lu et al. 2024; Wu et al. 2024).

There are prior works that tried to extend 3D models for 4D through per-frame optimization, but they are costly, not very accurate for predicting Gaussian parameters, and limited ability to handle dynamic scene elements or adapt to continuous deformations in complex object geometries. Existing methods (Yang et al. 2024; Huang et al. 2024; Liang et al. 2023; Sun et al. 2024) mitigate the first two limitations by using mean Gaussian location. However, these approaches largely fail to address the third limitation due to a fundamental dilemma in their design. A common underlying strategy involves employing a Multilayer Perceptron (MLP) to model all temporal deformations. This creates an inherent tension: the network capacity required to capture intricate local details is in disagreement with the need for smooth, globally coherent motion. Consequently, the models tuned with local details often produce temporally inconsistent results, while models tuned with global details often tend to over-smooth the scene, missing the high-frequency dynamics. This results in reconstructions that lack temporal consistency or fail to preserve dynamic details.

While the principle of decomposing motion into local and global components has been explored, prior works often treat them as independent or loosely coupled systems. We introduce FLAG-4D, a dual-network of synergistic specialists to resolve the tension between local detail and global consistency. Our key contribution is not merely on the decomposition itself, but the tightly integrated mechanism by which these components interact. The framework consists of: (1) an Instantaneous Deformation Network (IDN), which excels at forecasting fine-grained, highly localized deformations, which are crucial for preserving texture and subtle movements. (2) a Global Motion network (GMN), which is designed to capture broader, scene-level dynamics and ensure long-range temporal consistency.

The tight coupling between these specialists is achieved by the GMN, which leverages our proposed Contextual Deformation Alignment (CDA) mechanism. This deformation-guided attention process introduces a dynamic, query-based approach where the IDN’s “look-ahead” local predictions act as a query that probes the GMN. The GMN then provides tailored, globally-aware guidance to the final deformation by using its CDA module to selectively retrieve relevant con-

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

text from rich optical flow embeddings. This architecture is the first to use local motion forecast to guide the selection of global information, differs from the simple feature concatenation or indirect loss-based supervision. This unique CDA process allows the specialists to synergize: the IDN preserves intricate local details from ambiguous monocular cues, while GMN encourages global coherence needed to solve for large-scale motion ambiguities. Our contributions can be summarized as follows:

- We propose a novel, tightly-coupled dual-network architecture of synergistic specialists: an Instantaneous Deformation Network (IDN) to model local details and a Global Motion Network (GMN) for global coherence.
- We introduce a novel query-based mechanism, Contextual Deformation Alignment (CDA), where a local deformation forecast is used as a dynamic query to selectively retrieve context from optical flow embeddings.
- We introduce a synergistic mutual learning strategy to harmonize the specialists networks and enables them to learn complementary representations, enhancing temporal coherence and reconstruction quality.

Related Work

Dynamic 3D Reconstruction Several foundation works in 4D reconstruction have advanced the use of Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) for dynamic scenes. Nerfies (Park et al. 2021a), Hypernerf (Park et al. 2021b), D-nerf (Pumarola et al. 2021), and DyNeRF (Li et al. 2022) have pioneered the application of deformation neural fields to extend the static NeRFs to capture dynamic scenes. However, NeRF-based approaches often suffer from high computational costs, limiting their real-time applicability (Stephen et al. 2021; Wang et al. 2023b,a; Park et al. 2021b). Alternatives like explicit voxel grids (Fang et al. 2022) or plane-based factorizations (Shao et al. 2023; Fridovich-Keil et al. 2023; Cao and Johnson 2023) improved real-time flexibility using MLP decoders for deformation.

Dynamic 3D Gaussian Splatting Recently, Dynamic 3D Gaussian Splatting (3DGS) methods have gained significant traction due to their rendering speed and quality. (Wu et al. 2024; Duisterhof et al. 2023) adapted plane-based encodings, or directly integrated temporal components into Gaussian representation (Li et al. 2024). Many subsequent works rely on MLPs to predict Gaussian deformations over time (Yang et al. 2024, 2023; Sun et al. 2024), with variations including sparse control points (Huang et al. 2024), specialized per-Gaussian embeddings (Bae et al. 2024), and models that distinguish between static and deformable scene (Liang et al. 2023). Other approaches target temporal consistency specifically by employing tunable MLPs for varied motion patterns (Liu and Banerjee 2024), using time-calibrated inputs (Kim et al. 2024), recovering deformations with motion bases (Kratimenos, Lei, and Daniilidis 2025), or explicit temporal modeling with polynomials and Fourier series to each Gaussian’s trajectory (Lin et al. 2024). These works are solely depend on the mean Gaussian position and current time step while lacking relevant guidance, stressing on the deformation

network to predict the deformation vectors in the next time step.

Optical Flow in 3D Dynamic Reconstruction Optical flow, refined by deep learning models (Dosovitskiy et al. 2015; Sun et al. 2018; Teed and Deng 2020; Huang et al. 2022), is increasingly used in dynamic 3D reconstruction to enhance temporal consistency and motion accuracy. Optical flow has been applied to supervise 3D Gaussian movements (Dosovitskiy et al. 2015; Sun et al. 2018; Teed and Deng 2020; Huang et al. 2022), or as a component in the loss function to encourage smoother transitions (Liu et al. 2023; Katsumata, Vo, and Nakayama 2025; Zhu et al. 2024; Gao et al. 2024). In distinct from all these approaches, our work is the first to leverage on the rich, pre-trained optical flow embedding as input to guide the deformation process directly.

Methodology

3D Gaussian Splatting 3D Gaussian Splatting represents the scene with a set of explicit 3D Gaussians. Each primitive k is characterized by: center location $\mu_k \in \mathbb{R}^3$, covariance matrix $\Sigma_k \in \mathbb{R}^{3 \times 3}$, opacity $o_k \in [0, 1]$, and view-dependent color c_k , modeled with spherical harmonics coefficient sh_k . The covariance matrix Σ_k can be decomposed as $\Sigma_k = R_k S_k S_k^T R_k^T$, where rotation matrix R_k is represented by quaternion $q_k = [r_w, r_x, r_y, r_z]$, and a scaling factor, represented by $s_k \in \mathbb{R}^3$. Given a 3D point $x \in \mathbb{R}^{3 \times 1}$, the 3D Gaussian can be formulated as: $G(x) = o \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$. During the rendering process, these 3D Gaussians are splatted onto the 2D image plane. This involves a viewing transform matrix W and the Jacobian matrix J of the affine approximation of the projective transformation, yielding a 2D covariance matrix Σ_{2D} through: $\Sigma_{2D} = JW\Sigma W^T J^T$, and the 2D center position $\mu_{2D} = JW\mu$. The final color $C(p)$ for a pixel p is computed by α -blending the projected Gaussians, sorted by depth $C(p) = \sum_{k=1}^N SH(sh_k, v_k) \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j)$, where SH is the spherical harmonic function, v_k is the view direction, and α_k represent the density computed from the k -th 3D Gaussian.

FLAG-4D The main objective of this work is to model a 4D reconstruction model, $F_\omega(\cdot)$:

$$F_\omega(G_0^i(x, r, s, c, \sigma), t) \rightarrow \\ G_t^i(x + \delta x, r + \delta r, s + \delta s, c + \delta c, \sigma + \delta \sigma), \quad (1)$$

that takes an initial set of n -number of Gaussians, $G_0^i(x, r, s, c, \sigma)|_{i=0}^n$, where x defines the center, r is the rotation, s is the scaling, c is the color, and the σ defines the opacity of a single Gaussian, and produces deformed Gaussians, $G_t(x + \delta x, r + \delta r, s + \delta s, c + \delta c, \sigma + \delta \sigma)$, that models the dynamics of the initial set of Gaussians G_0 , by learning a deformation field by tuning the parameter, ω , of the model $F_\omega(\cdot)$. We get the initial set of Gaussians G_0 , from the structure from motion (SfM) point cloud (Kerbl et al. 2023). The SfM point cloud is generated from a sequence of video frames, $\{v_0, v_2, \dots, v_T\}$ as the time varies $t = 0, \dots, T$, captured using a monocular camera (Wang et al. 2021).

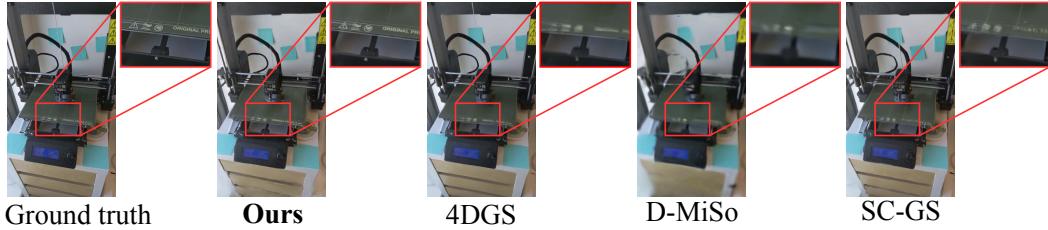


Figure 1: Visual Comparison of our method against very recent methods, such as: 4DGS (Wu et al. 2024), SC-GS (Huang et al. 2024), and D-MiSo (Waczyńska et al. 2024) on the HyperNeRF dataset. Our method demonstrates finer detail preservation across timesteps, particularly in the texture and edges of the zoomed-in regions. This results in a higher fidelity and coherent reconstruction across dynamic viewpoints compared to the baseline.

While complementary methodologies such as (Das et al. 2024; Katsumata, Vo, and Nakayama 2023; Lu et al. 2024; Wu et al. 2024; Li et al. 2024) have demonstrated promising results in 4D reconstruction, a major research gap remains: these approaches largely fail to capture local consistency while preserving global dynamics. The FLAG-4D employs a novel dual-deformation framework enabled with both local and global deformation, is used to model dynamic scenes through evolving 3D Gaussian primitives. To this end, we adapt the standard formulation in Eqn. 1 to incorporate our dual-network design:

$$F_\omega(\text{IDN}(\cdot), \text{GMN}(\cdot)) \rightarrow \\ G_t^i(x + \delta x, r + \delta r, s + \delta s, c + \delta c, \sigma + \delta \sigma) \quad (2)$$

This dual-network design is a deliberate response to the fundamental trade-off where a single network struggles to model fine-grained local dynamics and preserve global temporal coherence simultaneously. Our framework resolves this tension with Instantaneous Deformation Network, IDN(\cdot) to capture high-frequency local details that monolithic models often blur, while the Global Motion Network GMN(\cdot), guided by optical flow, maintains the long-range dependencies. A mutual learning strategy is used to harmonize these two network predictions. The necessity of this decomposition is quantitatively confirmed by ablation studies (Tab. 3) and qualitatively demonstrated by the detail preservation and coherent motion in our visual results (Figs. 1,3,4).

Instantaneous Deformation for Local Deformation

Given the time embedding, Time(t), and the initial set of Gaussians G_0 , the IDN provides the fine-grained local deformation, see Fig. 2 bottom left, i.e.:

$$\text{IDN}(G_0^i(x, r, s, c, \sigma), \text{Time}(t)) \rightarrow \\ G_{t, \text{local}}^i(x + \delta x, r + \delta r, s + \delta s, c + \delta c, \sigma + \delta \sigma), \quad (3)$$

where $i = 0, \dots, n$. The Time(t) represents the time-driven positional encoding of the time stamp t and the future time window $\{t+1, \dots, t+W-1\}$ of window size W . The IDN takes initial Gaussians G_0 and Time(t) as inputs and produces offset for initial Gaussian parameters.

Enabling the IDN with *only* time embedding, Time(t) and G_0 are necessary but not a sufficient condition. We examined

that the local deformation across time for minute objects, small colors, and discontinuous geometry requires a more informed notion of time and consistency. To this end, IDN is tasked with a two-step process. First, its core deformation network, IDN_{core} is repeatedly queried for the window, W , to get a more future-tailored timesteps relative to the current time t (e.g., for timesteps $\{t+1, t+2, \dots, t+(W-1)\}$). This yields a sequence of provisional future deformation vectors, $\mathcal{H}_{\text{future}}^{\text{IDN}} = [\delta\theta_t^{\text{IDN}}, \dots, \delta\theta_{t+(W-1)}^{\text{IDN}}]$. This sequence is then processed by a Gated Recurrent Unit (GRU) (Dey and Salem 2017) to distill the temporal dynamics of the hypothesized local trajectory into a single, compact representation:

$$\delta\theta_{\text{rep}} = \text{GRU}(\mathcal{H}_{\text{future}}^{\text{IDN}}) \quad (4)$$

In this way, $\delta\theta_{\text{rep}}$ encapsulates a more learned summary of the IDN’s hypothesized local motion trajectory over a short future horizon. For clarity, we refer the querying process of the IDN_{core} and GRU as a function of the IDN module. The GRU role in IDN is to distill the pattern of anticipated local dynamics from each forecast and provide GMN a predictive context. Crucially, the input sequence of the GRU is generated on-the-fly by the feed-forward IDN for each frame independently. Therefore, the use of GRU does not create inter-frame dependencies, and supports fast, random time access for rendering.

Global Motion Network for Global Deformation

While the FLAG-4D is refined for local motions, it is still a question of how we could leverage the global motion to model the global deformation. Prior works (Kerbl et al. 2023; Kwak et al. 2025) have shown results by learning the global motion directly from the sfm or by learning anchor points. In contrast, our GMN is designed as a multi-stage pipeline. First, a Temporal Fusion Encoder is used to process and fuse optical flow embeddings. These global motion features are then aligned with IDN’s hypothesized local deformation using a Contextual Deformation Alignment. Finally, a terminal Deformation Refinement Network synthesizes these aligned features to predict the final deformation for the next timesteps. **Temporal Fusion Encoder.** We note that the global motion (and, hence global deformation) can be achieved directly by leveraging the optical flow. To explicitly capture inter-frame motion for more accurate Gaussian deformations, the FLAG-4D incorporates a *Temporal Fusion Encoder* that processes

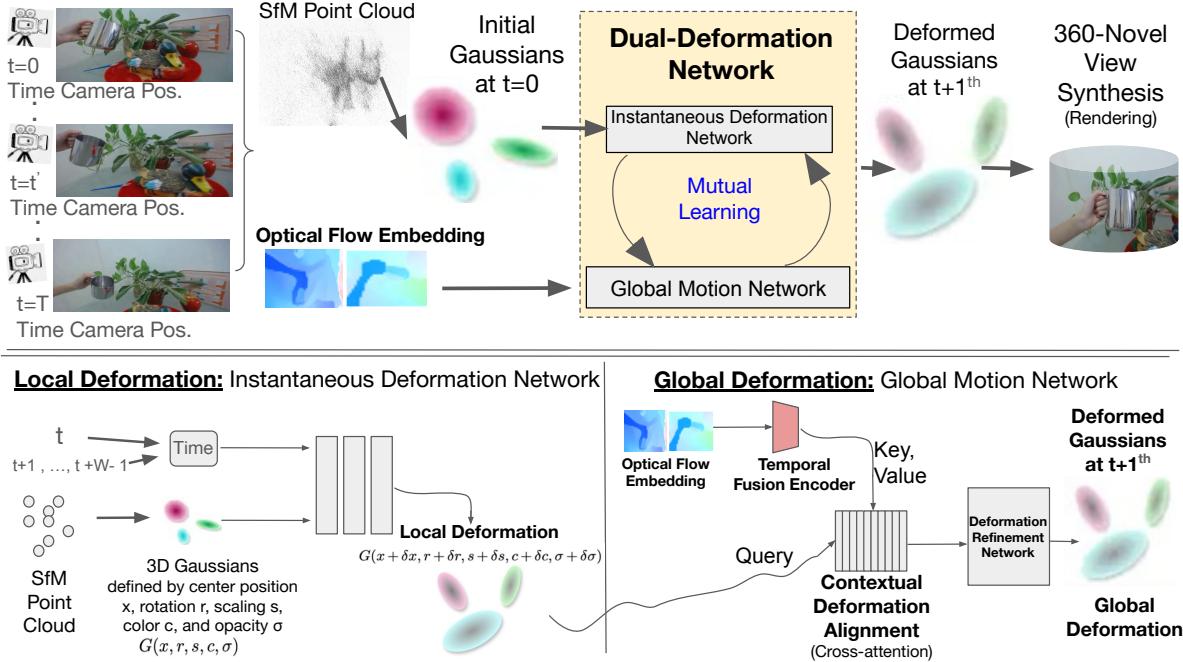


Figure 2: FLAG-4D Methodology: Our dual-deformation framework for 4D reconstruction. **Top:** The overall pipeline: A monocular video sequence is used to generate an initial SfM point cloud, from which a canonical set of 3D Gaussians at $t = 0$ is derived. The Dual-Deformation Network consists of an Instantaneous Deformation Network (IDN) and a Global Motion Network (GMN), which are trained synergistically via Mutual Learning. **Bottom Left:** The IDN processes the canonical Gaussians and a window of future-oriented time embeddings to produce a hypothesized local deformation. **Bottom Right:** The GMN integrates this local deformation hypothesis (as Query) with fused optical flow embeddings (as Key/Value) via a cross-attention mechanism, producing the final globally consistent deformation.

motion embeddings extracted from a pretrained optical-flow network. This feature extraction is a pre-computation step for the entire video sequence, and the resulting embeddings are cached. Thus, it does not impact the final interactive rendering speed, which only relies on loading these pre-computed features. Instead of the raw, and often noisy, 2-D optical-flow field, we use high-dimensional intermediate embeddings $flow_{t-1 \rightarrow t}$, $flow_{t \rightarrow t+1} \in \mathbb{R}^{B \times H \times W \times 128}$ obtained from the backbone. Two optical-flow embeddings—(1) past-to-current ($flow_{t-1 \rightarrow t}$) and (2) current-to-future ($flow_{t \rightarrow t+1}$)—are concatenated along the feature dimension to form a composite tensor $flow_{emb} = \text{Concat}[flow_{t-1 \rightarrow t}, flow_{t \rightarrow t+1}]$. The Temporal Fusion Encoder $Fusion_{enc}$ then produces $M_{final}^{t+1} = Fusion_{enc}(flow_{emb}, t+1)$, which summarizes both the motion that led to the current state and the anticipated motion of the next frame, and thus serves as the primary image-based scene-dynamics representation.

Contextual Deformation Alignment. Once the global motion is captured over a time using the temporal fusion encoder, the remaining task is to align the global motion and the local motion encoding. We design a network that aligns local deformation with global motion cues through multi-head cross-attention (MHA). The main purpose of the MHA is to synchronize the local deformation with the global motions, yielding temporally consistent deformation:

$$M_{\text{scene},i}^{t+1} = \text{softmax}\left(\frac{G_{\text{unified}} M_{\text{final}}^{t+1\top}}{\sqrt{d}}\right) M_{\text{final}}^{t+1}, \quad (5)$$

where $G_{\text{unified}} = \delta\theta_{\text{rep}} + G^i$ is the Gaussian state across the window $[t+1, \dots, t+W-1]$. G_{unified} probes how the global motion should act, given the Gaussian’s anticipated local evolution. The fused embedding M_{final}^{t+1} is further enriched with sinusoidal positional encodings to preserve spatio-temporal order and served as the *key* and *value* in MHA, while G_{unified} acts as the *query*, essentially asks: “For this primitive at this location with the anticipated local motion, what is the relevant global motion information?” The result, $M_{\text{scene},i}^{t+1}$, is an optical-flow feature that is selectively filtered and weighted according to the IDN-predicted future trajectory of Gaussian i , giving a context-aware, dynamic motion descriptor.

Deformation Refinement Network (DRN) At $t+1$, DRN forms the final predictive stage of FLAG-4D. Acting as an *integrator*, it combines the context-aligned motion representation M_{scene} with the noise-augmented temporal embedding $S(t+1)$ to generate the refined deformation:

$$G_{t+1}^{\text{GMN}} = \text{DRN}(\text{Concat}(S(t+1), G_t, M_{\text{final}}^{t+1})), \quad (6)$$

where $G_t = G_0 + \delta\theta_t$ and $\delta\theta_t$ is the direct forecast of IDN_{core} at time t . This design ensures that the GMN’s prediction is conditioned on explicit geometric prior for the target state (IDN’s direct forecast), while this prior is itself interpreted and modulated by context from both the IDN’s hypothesized future trajectory and observed optical flow.

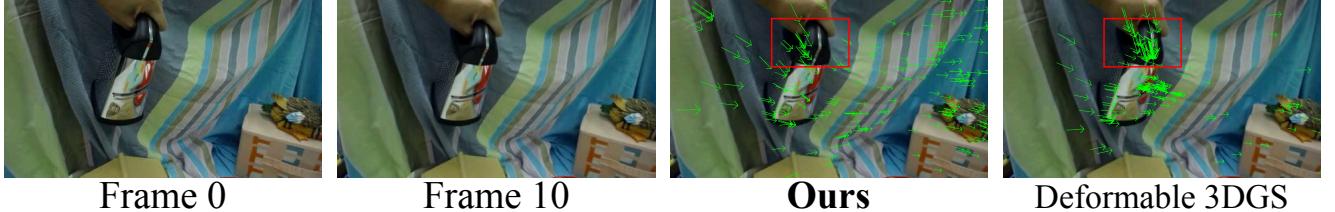


Figure 3: **Comparison of Predicted Gaussian Deformation ($t \rightarrow t + 10$) for the "Bell" Scene.** (a) Frame 0 (GT). (b) Frame 10 (GT). (c) FLAG-4D (Ours) accumulated flow from $t = 0$ to $t = 10$, overlaid on Frame 0. (d) Deformable 3D Gaussians(Yang et al. 2024) accumulated flow. FLAG-4D produces a more accurate and coherent deformation field. The highlighted region (red box) demonstrates our method’s superior preservation of the bell’s local rigidity during its motion towards the state in Frame 10.

Mutual Learning of GMN and IDN

We create a feedback mechanism to bring the IDN and GMN to learn from each other. This mechanism acts as a regularization in both networks, ensuring a more stable deformation prediction. The feedback mechanism is based on Mutual Learning (Zhang et al. 2018), encouraging both networks to share representations and learn complementary features that are hard to capture by a network independently. The mutual learning objective $\mathcal{L}_{\text{mutual}}$ is formulated to encourage agreement between their respective predictions, $\delta\theta_{t,i}^{\text{GMN}}$ and $\delta\theta_{t,i}^{\text{IDN}}$.

$$\mathcal{L}_{\text{mutual}} = \frac{1}{N_{\text{vis}}} \sum_{i \in G_{\text{vis}}} \left(\|\text{sg}(\delta\theta_{t,i}^{\text{IDN}}) - \delta\theta_{t,i}^{\text{GMN}}\|_2^2 + \|\delta\theta_{t,i}^{\text{IDN}} - \text{sg}(\delta\theta_{t,i}^{\text{GMN}})\|_2^2 \right) \quad (7)$$

where $\text{sg}(\cdot)$ represents the stop-gradient operator. The first term treats the IDN’s prediction as a pseudo-target to supervise the GMN, encouraging GMN to learn how to produce deformations that are globally consistent which still align with fine-grained local details. The second term treats the GMN’s prediction as a pseudo-target for the IDN, allowing IDN to learn from GMN’s broader contextual understanding extracted from optical flow and its look-ahead capabilities. This bidirectional distillation harmonizes their distinct inductive bias, encouraging the predictions to become consistent and complementary.

Optimization

Smoothen Opacity Regularization Managing the lifecycle of the Gaussian primitive is crucial for both quality and efficiency. Inspired by the opacity resetting strategy (Rota Bulò, Porzi, and Kortschider 2024), we design a smoother opacity regularization to promote sparsification by gradually reducing opacity. This strategy offers more stability than conventional periodic hard opacity resets, which can disrupt training and densification. The current opacity logit for a primitive k , α_k , is converted to its corresponding opacity value $o_k = \sigma(\alpha_k)$, where $\sigma(\cdot)$ is the sigmoid function. This opacity value o_k is then decayed by a small factor δ_o and clamped within a defined range $[o_{\min}, o_{\max}]$ to obtain the new opacity value o'_k :

$$o'_k = \text{clamp}(o_k - \delta_o, o_{\min}, o_{\max}) \quad (8)$$

where $\delta_o = 0.001$, and the clamping range $[o_{\min}, o_{\max}]$ is set to $[0.01, 1.0]$. This update occurs after each densification step. Finally, to facilitate optimization, this new opacity value o'_k is converted back to logit space, $\alpha'_k = \text{logit}(o'_k)$, where $\text{logit}(\cdot)$ is the inverse sigmoid function, and α'_k is stored as the updated opacity logit for the primitive.

Depth Enhanced Loss We obtain the normalized depth map D_{gt} and it acts as a guidance to enhance the depth estimation. We rendered the depth map D during the rasterization process, and normalize the depth value to $[0, 1]$, to make the supervision robust to scale and shift differences. Averaged over all pixels \mathcal{P} in the view, the loss is:

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left| \frac{D(p) - \min(D)}{\max(D) - \min(D)} - D_{gt}(p) \right| \quad (9)$$

This encourages the rendered geometry is align with the structure provided by the external depth prior.

Local Rigidity Loss We use the local rigidity loss (Huang et al. 2024; Stearns et al. 2024; Yu et al. 2024), i.e. $\mathcal{L}^{\text{rigid}}$, to encourage the network to learn physically plausible motion and ensure locality in dynamic scenes.

$$\mathcal{L}_{i,j}^{\text{rigid}} = w_{i,j} \|(\mu_{j,t-1} - \mu_{i,t-1}) - R_{i,t-1} R_{i,t}^{-1} (\mu_{j,t} - \mu_{i,t})\|_2, \quad (10)$$

$$\mathcal{L}^{\text{rigid}} = \frac{1}{k|G|} \sum_{i \in G} \sum_{j \in \text{knn}_{i,k}} \mathcal{L}_{i,j}^{\text{rigid}}. \quad (11)$$

In this formulation, $\mu_{i,t}$ denotes the position of Gaussian i at time t , with the same notation used for the time step $t - 1$. For each Gaussian i , neighboring Gaussian j are encouraged to move in alignment with the rigid-body transform of the coordinate system of i between timesteps. Here, $w_{i,j}$ represents a weighting factor, and $\text{knn}_{i,k}$ denotes the k -nearest neighbors of Gaussian i within the Gaussians set G .

Total Loss The total loss can be summarized as:

$$\mathcal{L} = \mathcal{L}_{\text{mutual}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{rigid}} + \mathcal{L}_{\text{render}} \quad (12)$$

representing the mutual learning loss, depth-enhanced loss, local rigidity loss, and rendering loss. The rendering loss is a combination of L_1 loss and D-SSIM loss to balance sharpness and structural coherence.

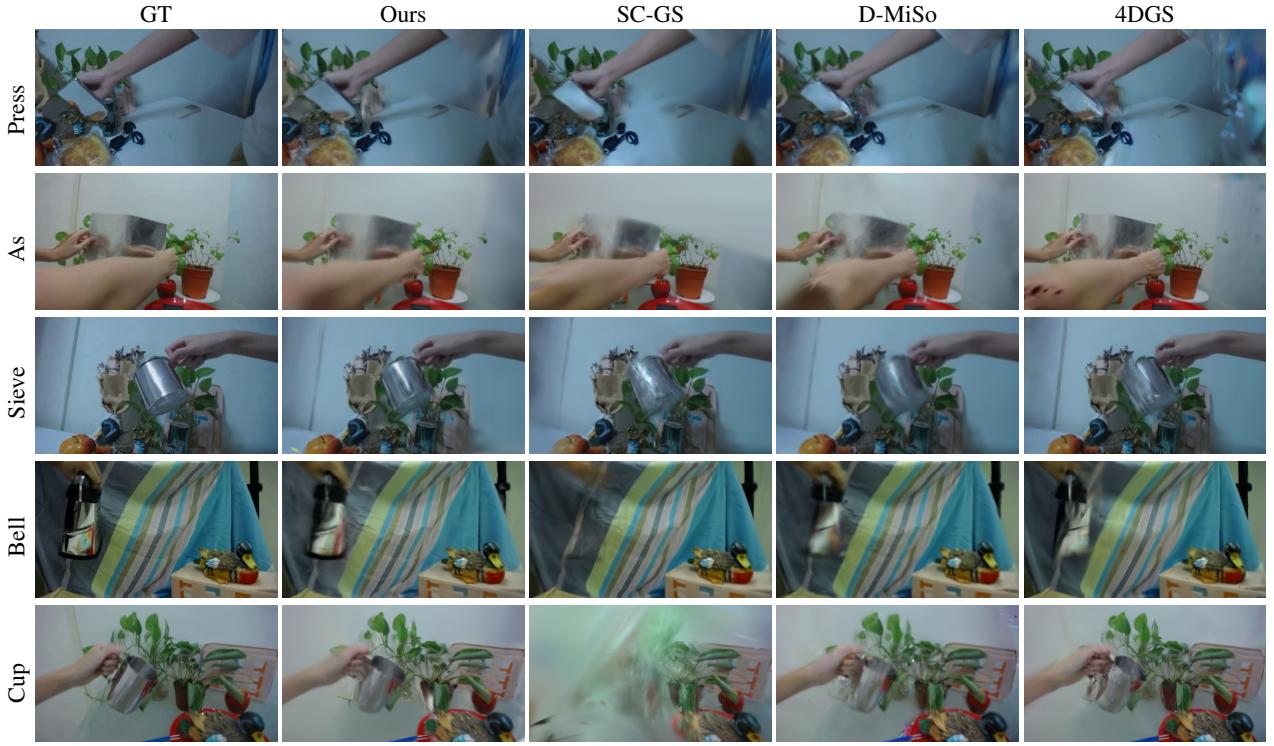


Figure 4: Qualitative comparisons between baseline methods and our approach on the NeRF-DS real-world dataset. Results show that our method delivers superior rendering quality in the case of complex scene dynamics. Our method is capable to capture finer details, preserve complex structure, handling the dynamic scene elements more effectively than baseline methods such as SC-GS(Huang et al. 2024), D-MiSo(Waczyńska et al. 2024), and 4DGS(Wu et al. 2024).

Method	PSNR↑	SSIM↑	LPIPS↓
3D-GS (Kerbl et al. 2023)	20.29	0.782	0.292
TiNeuVox (Fang et al. 2022)	21.61	0.823	0.277
HyperNeRF (Park et al. 2021b)	23.45	0.849	0.199
NeRF-DS (Yan, Li, and Lee 2023)	23.60	0.849	0.182
Deformable 3DGS (Yang et al. 2024)	23.76	0.848	0.180
SC-GS (Huang et al. 2024)	22.25	0.824	0.203
D-MiSo (Waczyńska et al. 2024)	23.90	0.851	0.151
4DGS (Wu et al. 2024)	22.54	0.837	0.212
MotionGS (Zhu et al. 2024)	23.71	0.831	0.240
Ours	24.23	0.852	0.198

Table 1: Quantitative comparison on NeRF-DS (Yan, Li, and Lee 2023) dataset. Mean performance across all scenes.

Experiments

In the experiment setup, we use GeoWizard (Fu et al. 2024) as our depth estimation model to guide the depth-enhanced loss, and MemFlow (Dong and Fu 2024) is used as the pre-trained optical flow network to provide guidance. The experiment is conducted on a single A100 GPU. For all datasets, we use Adam optimizer (Kingma 2014) to optimize both the IDN and GMN, along with the 3D Gaussian parameters end-to-end. The learning rate for IDN and GMN parameters are set to $1e-4$. A time window of 4 is used for IDN’s future hypoth-

Method	PSNR↑	SSIM↑
SC-GS (Huang et al. 2024)	20.92	0.63
D-MiSo (Waczyńska et al. 2024)	22.47	0.61
Deformable 3DGS (Yang et al. 2024)	22.06	0.58
MotionGS (Zhu et al. 2024)	20.96	0.49
Ours	22.33	0.78

Table 2: Quantitative comparison on HyperNeRF’s (Park et al. 2021b) vrig dataset. Mean performance across all scenes.

esis sequence. To evaluate the performance of our method, we perform experiments on a monocular real-world dataset from NeRF-DS (Yan, Li, and Lee 2023) and HyperNeRF (Park et al. 2021b). We utilize standard image quality metrics, PSNR, SSIM, and LPIPS, to assess the effectiveness of our method.

Results

Comparisons on NeRF-DS dataset. NeRF-DS dataset is a challenging benchmark that contains complex scene structure and motion. Tab. 1 shows that our approach achieves state-of-the-art (SOTA) performance across a majority of these scene, indicating high reconstruction fidelity. We note a favorable trade-off in the LPIPS perceptual metric, where our framework’s deliberate focus on preserving sharp details and

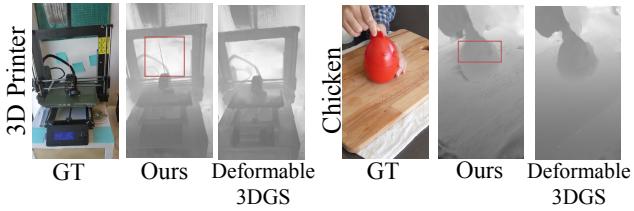


Figure 5: **Rendered Depth Map Comparison** FLAG-4D produces depth maps with visibly greater detail and geometric accuracy (e.g., printer’s string and surface texture).

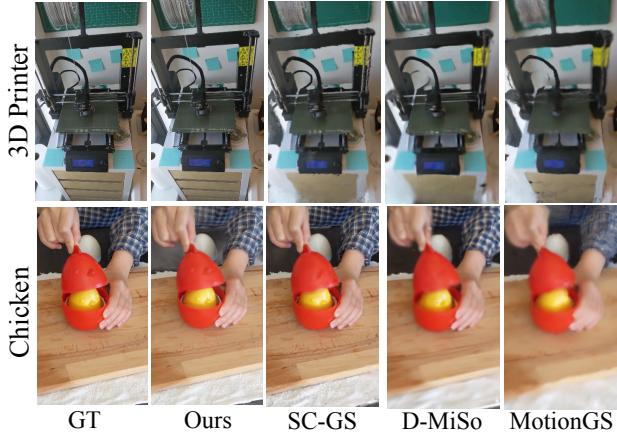


Figure 6: **(Best viewed while zoomed-in.)** Qualitative comparisons between baselines and our approach on the HyperNeRF dataset. The comparison illustrates our model achieves higher rendering quality consistently in challenging real-world scenes where camera pose estimation is less accurate.

eliminating motion blur, results in a competitive score while excelling in PSNR and SSIM (Zheng et al. 2024). The qualitative results in Fig. 4 highlight our model is more effective in capturing finer details and maintaining clarity. This enhanced coherence is not only aesthetic, but is also evident in the underlying learned motion. As Fig. 3 directly visualizes, the accumulated deformation field generated by our framework is visibly more structured and physically plausible than the baseline method. It preserves the object’s rigidity correctly over a long time horizon, while the baseline produces a less coherent flow.

Comparisons on HyperNeRF dataset. Tab. 2 summarizes the quantitative performance on HyperNeRF. We obtain SOTA performance, highlighted by the strongest SSIM scores in every scene. The qualitative comparison of HyperNeRF is shown in Fig. 6. Our approach has results that have the best sharpness and visual quality amongst the other methods. The zoom-in comparison is shown in Fig. 1. While other methods have blurred or illegible text, our method preserves the fine textual details of the wordings on the machine more accurately, with more sharpness and clarity.

Table 3: Ablation studies. We evaluate on the effect of the proposed IDN, GMN and the Depth Enhanced Loss on NeRF-DS (Yan, Li, and Lee 2023) dataset.

Methods	PSNR(\uparrow)	M-SSIM(\uparrow)	LPIPS(\downarrow)
FLAG-4D w/o IMN	24.01	0.868	0.226
FLAG-4D w/o GMN	23.39	0.853	0.223
FLAG-4D w/o Mutual Learning	23.70	0.878	0.157
FLAG-4D	24.23	0.884	0.157

Ablation Studies

Removing IDN In this ablation, we evaluate a variant of our model where the IDN is removed (“FLAG-4D w/o IDN”). This setup relies solely on the GMN for the deformation prediction. GMN still receives temporal signals and fused optical flow embeddings, but the attention mechanism and final prediction lack the local deformation hypotheses and hypothesized future local trajectory $\delta\theta_{rep}$, which are provided by IDN. From Tab. 3, the performance drops even optical flow is provided to GMN, and it struggles to reconstruct these high-frequency spatial and temporal details.

Removing GMN Conversely, we evaluate another variant where the GMN is disabled (“FLAG-4D w/o GMN”). In this setup, the deformations are predicted solely by the IDN, based on its temporal signal Time(t) and its internal GRU processing its own future hypothesized local trajectory. This variant lacks the integration of external optical flow features and the global refinement provided by the GMN’s attention mechanism. Disabling GMN also results in degradation in performance, showing it struggles with long-range temporal coherence and global consistency.

Disabling Mutual Learning IDN and GMN are trained separately by its own loss. Disabling Mutual Learning results in degradation in performance, and this is attributed to the IDN and GMN can diverge in their learned representation. Mutual Learning acts as an essential regularizer, it encourages GMN to respect local details while forcing the IDN to adhere to global motion. The degradation further proves the mutual learning is vital for combining the strength of local and global networks into a unified, high-fidelity representation.

Conclusion

In this work, we introduce FLAG-4D, a dual-network framework that harmonizes the core tension between local detail and global coherence in 4D reconstruction. Our framework delegates these tasks to two synergistic specialists, IDN for local detail and GMN for global coherence. These specialists are tightly coupled by our CDA mechanism. By directly leveraging the optical flow embedding as input to guide the deformation, our method achieves state-of-the-art performance, producing high fidelity and temporal consistency. Future work can focus on reducing the dependency on pre-trained flow networks to improve robustness in scenes with significant motion blur, where flow estimation can easily fail.

Acknowledgments

We acknowledge the GRS support from the School of Information Technology, as well as the support and provision of GPU resources from the HPC/APC team and its manager, Dr. Marcus. Dr. Srinivas and Dr. Arghya acknowledge the support of the Anusandhan National Research Foundation (ANRF), Government of India (CRD/2024/000973).

References

- Bae, J.; Kim, S.; Yun, Y.; Lee, H.; Bang, G.; and Uh, Y. 2024. Per-Gaussian Embedding-Based Deformation for Deformable 3D Gaussian Splatting. *arXiv preprint arXiv:2404.03613*.
- Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 130–141.
- Das, D.; Wewer, C.; Yunus, R.; Ilg, E.; and Lenssen, J. E. 2024. Neural parametric gaussians for monocular non-rigid object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10715–10725.
- Dey, R.; and Salem, F. M. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 1597–1600. IEEE.
- Dong, Q.; and Fu, Y. 2024. MemFlow: Optical Flow Estimation and Prediction with Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19068–19078.
- Dosovitskiy; Alexey; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Smagt, P. V. D.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2758–2766.
- Duisterhof, B. P.; Mandi, Z.; Yao, Y.; Liu, J.-W.; Shou, M. Z.; Song, S.; and Ichnowski, J. 2023. Md-splatting: Learning metric deformation from 4d gaussians in highly deformable scenes. *arXiv preprint arXiv:2312.00583*.
- Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.
- Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.
- Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. In *ECCV*.
- Gao, Q.; Xu, Q.; Cao, Z.; Mildenhall, B.; Ma, W.; Chen, L.; Tang, D.; and Neumann, U. 2024. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*.
- Huang, Y.-H.; Sun, Y.-T.; Yang, Z.; Lyu, X.; Cao, Y.-P.; and Qi, X. 2024. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4220–4230.
- Huang, Z.; Shi, X.; Zhang, C.; Wang, Q.; Cheung, K. C.; Qin, H.; Dai, J.; and Li, H. 2022. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, 668–685. Springer.
- Katsumata, K.; Vo, D. M.; and Nakayama, H. 2023. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv preprint arXiv:2311.12897*.
- Katsumata, K.; Vo, D. M.; and Nakayama, H. 2025. A compact dynamic 3d gaussian representation for real-time dynamic view synthesis. In *European Conference on Computer Vision*, 394–412. Springer.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.
- Kim, S.; Bae, J.; Yun, Y.; Son, H.; Lee, H.; Bang, G.; and Uh, Y. 2024. Optimizing Dynamic NeRF and 3DGS with No Video Synchronization. In *ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*.
- Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kratimenos, A.; Lei, J.; and Daniilidis, K. 2025. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In *European Conference on Computer Vision*, 252–269. Springer.
- Kwak, S.; Kim, J.; Jeong, J. Y.; Cheong, W.-S.; Oh, J.; and Kim, M. 2025. MoDec-GS: Global-to-Local Motion Decomposition and Temporal Interval Adjustment for Compact Dynamic 3D Gaussian Splatting. *arXiv:2501.03714*.
- Li, T.; Slavcheva, M.; Zollhoefer, M.; Green, S.; Lassner, C.; Kim, C.; Schmidt, T.; Lovegrove, S.; Goesele, M.; Newcombe, R.; et al. 2022. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5521–5531.
- Li, Z.; Chen, Z.; Li, Z.; and Xu, Y. 2024. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8508–8520.
- Liang, Y.; Khan, N.; Li, Z.; Nguyen-Phuoc, T.; Lanman, D.; Tompkin, J.; and Xiao, L. 2023. Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. *arXiv preprint arXiv:2312.11458*.
- Lin, Y.; Dai, Z.; Zhu, S.; and Yao, Y. 2024. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21136–21145.
- Liu, B.; and Banerjee, S. 2024. SwinGS: Sliding Window Gaussian Splatting for Volumetric Video Streaming with Arbitrary Length. *arXiv preprint arXiv:2409.07759*.
- Liu, Y.-L.; Gao, C.; Meuleman, A.; Tseng, H.-Y.; Saraf, A.; Kim, C.; Chuang, Y.-Y.; Kopf, J.; and Huang, J.-B.

2023. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13–23.
- Lu, Z.; Guo, X.; Hui, L.; Chen, T.; Yang, M.; Tang, X.; Zhu, F.; and Dai, Y. 2024. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8900–8910.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021a. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5865–5874.
- Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021b. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.
- Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.
- Rota Bulò, S.; Porzi, L.; and Kotschieder, P. 2024. Revising densification in gaussian splatting. In *European Conference on Computer Vision*, 347–362. Springer.
- Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.
- Stearns, C.; Harley, A.; Uy, M.; Dubost, F.; Tombari, F.; Wetzstein, G.; and Guibas, L. 2024. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*.
- Stephen, L.; Simon, T.; Schwartz, G.; Zollhoefer, M.; Sheikh, Y.; and Saragih, J. 2021. Mixture of volumetric primitives for efficient neural rendering. volume 40, 1–13. ACM New York, NY, USA.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz., J. 2018. Pwcnet: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8934–8943.
- Sun, J.; Jiao, H.; Li, G.; Zhang, Z.; Zhao, L.; and Xing, W. 2024. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20675–20685.
- Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 402–419.
- Waczyńska, J.; Borycki, P.; Kaleta, J.; Tadeja, S.; and Spurek, P. 2024. D-MiSo: Editing Dynamic 3D Scenes using Multi-Gaussians Soup. *arXiv preprint arXiv:2405.14276*.
- Wang, F.; Tan, S.; Li, X.; Tian, Z.; Song, Y.; and Liu, H. 2023a. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19706–19716.
- Wang, J.; Zhong, Y.; Dai, Y.; Birchfield, S.; Zhang, K.; Smolyanskiy, N.; and Li, H. 2021. Deep Two-View Structure-From-Motion Revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8953–8962.
- Wang, L.; Hu, Q.; He, Q.; Wang, Z.; Yu, J.; Tuytelaars, T.; Xu, L.; and Wu, M. 2023b. Neural residual radiance fields for streamably free-viewpoint videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 76–87.
- Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20310–20320.
- Yan, Z.; Li, C.; and Lee, G. H. 2023. NeRF-DS: Neural Radiance Fields for Dynamic Specular Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8285–8295.
- Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2024. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20331–20341.
- Yang, Z.; Yang, H.; Pan, Z.; and Zhang, L. 2023. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*.
- Yu, H.; Julin, J.; Milacski, Z. Á.; Niinuma, K.; and Jeni, L. A. 2024. Cogs: Controllable gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21624–21633.
- Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.
- Zheng, H.; Lin, Z.; Lu, J.; Cohen, S.; Shechtman, E.; Barnes, C.; Zhang, J.; Liu, Q.; Amirghodsi, S.; Zhou, Y.; et al. 2024. Structure-guided image completion with image-level and object-level semantic discriminators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7669–7681.
- Zhu, R.; Liang, Y.; Chang, H.; Deng, J.; Lu, J.; Yang, W.; Zhang, T.; and Zhang, Y. 2024. MotionGS: Exploring Explicit Motion Guidance for Deformable 3D Gaussian Splatting. *arXiv preprint arXiv:2410.07707*.