

SpectralSplats: Robust Differentiable Tracking via Spectral Moment Supervision

Avigail Cohen Rimon¹ , Amir Mann¹ , Mirela Ben-Chen¹ , and Or Litany^{1,2} 

¹ Technion - Israel Institute of Technology

² Nvidia

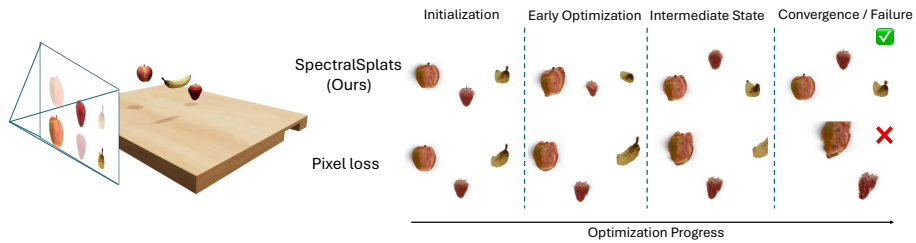


Fig. 1: SpectralSplats enables robust tracking from zero-overlap initializations. *Left:* A 3DGS asset is initialized (transparent overlay) far from a target pose image (solid image), resulting in strictly zero spatial overlap in the rendered camera view. *Right:* We compare the optimization progression. Standard photometric tracking (Pixel loss) implicitly requires spatial overlap; without it, directional gradients vanish, causing the optimizer to strand the asset and eventually collapse into spurious local minima. SpectralSplats (Ours) shifts supervision to the frequency domain via *Spectral Moments*. This establishes a global basin of attraction, allowing the Gaussians to flow across the image domain and successfully recover the extreme displacement.

Abstract. 3D Gaussian Splatting (3DGS) enables real-time, photorealistic novel view synthesis, making it a highly attractive representation for model-based video tracking. However, leveraging the differentiability of the 3DGS renderer “in the wild” remains notoriously fragile. A fundamental bottleneck lies in the compact, local support of the Gaussian primitives. Standard photometric objectives implicitly rely on spatial overlap; if severe camera misalignment places the rendered object outside the target’s local footprint, gradients strictly vanish, leaving the optimizer stranded. We introduce SpectralSplats, a robust tracking framework that resolves this “vanishing gradient” problem by shifting the optimization objective from the spatial to the frequency domain. By supervising the rendered image via a set of global complex sinusoidal features (Spectral Moments), we construct a global basin of attraction, ensuring that a valid, directional gradient toward the target exists across the entire image domain, even when pixel overlap is completely nonexistent. To harness this global basin without introducing periodic local minima associated with high frequencies, we derive a principled Frequency

Annealing schedule from first principles, gracefully transitioning the optimizer from global convexity to precise spatial alignment. We demonstrate that SpectralSplats acts as a seamless, drop-in replacement for spatial losses across diverse deformation parameterizations (from MLPs to sparse control points), successfully recovering complex deformations even from severely misaligned initializations where standard appearance-based tracking catastrophically fails.

1 Introduction

The recent advent of 3D Gaussian Splatting (3DGS) [16] has fundamentally disrupted the landscape of 3D reconstruction. By representing scenes as a collection of anisotropic 3D Gaussians, 3DGS achieves real-time rendering speeds and photorealistic quality. On top of being exceptionally capable at *static* Novel View Synthesis (NVS), its differentiable rendering property enables a critical application: the ability to take a reconstructed static asset and "enact" it by fitting it to a target video [2, 20, 26].

This task of model-based video tracking – estimating continuous geometric motion parameters to match a target observation – is foundational for applications like driving digital avatars, markerless motion capture, and editable dynamic scenes. Yet, estimating these continuous geometric displacements purely from visual observation remains an open and highly fragile challenge.

The core difficulty lies in the optimization landscape of *Analysis-by-Synthesis*. In a typical model-based tracking pipeline, we seek the motion parameters θ that minimize the photometric error between the rendered model and the observed target. This optimization relies on the differentiability of the renderer to backpropagate gradients from pixel errors to motion parameters. Crucially, this mechanism relies on *local* spatial overlap: for a primitive to receive gradient updates towards its corresponding visual structure in the target image, its rendered footprint must already intersect with that structure’s location. Since Gaussian splats are local primitives with compact support, if the estimated motion parameters are sufficiently far from the target (e.g., due to a coarse initialization or noisy pose priors), the rendered Gaussians do not overlap with their intended target pixels. As illustrated in Fig. 1, without this directional signal, the gradient component corresponding to the true target vanishes ($\nabla_{\theta} \mathcal{L}_{target} \rightarrow 0$), and the optimizer is actively steered towards arbitrary distractors or irrelevant local minima rather than the correct solution.

Fig. 2 dissects this “vanishing gradient” pathology in 1D. Under large spatial displacements, the standard spatial L_2 landscape lacks a global basin leading to the correct state, causing the tracker to fail catastrophically.

A standard workaround for this "basin of attraction" problem in dynamic 3D reconstruction is to rely on manual alignment or controlled setups to guarantee sufficient spatial overlap from the very first frame. Recent approaches like [2] found it useful to replace the standard L_2 loss with deep feature distances such as LPIPS. While the hierarchical receptive fields of these networks moderately widen the basin of attraction compared to raw pixel errors, they still

fundamentally rely on localized spatial overlap. Under severe camera misalignments or rapid motion where the rendered asset and the target are disjoint, the gradients from these deep features still vanish. Alternatively, approaches relying on *category-specific priors* [24, 40] bypass the global search problem by leveraging off-the-shelf pose estimators to provide a strong initial alignment, ensuring sufficient spatial overlap before appearance-based optimization even begins. While this reduces the photometric tracking to a simple “last-mile” refinement, it achieves robustness only by sacrificing generality, rendering them unsuitable for tracking arbitrary, “in-the-wild” objects. Consequently, there remains a critical need for a purely optimization-based tracking objective that is both global (capable of handling large, disjoint displacements) and class-agnostic.

To bypass this initialization dependency, we introduce *SpectralSplats*, a robust tracking framework that solves the vanishing gradient problem through Spectral Moment supervision. Our key insight is to shift the optimization objective from the spatial domain to the frequency domain. Unlike pixels or rendered splats, which are local, sinusoidal basis functions are global. By projecting the rendered image onto a set of complex Fourier features, we compute a “spectral signature” of the current pose. A spatial displacement of the object corresponds to a phase shift in these frequencies, providing a strong, non-zero gradient signal even when the object and its target are spatially disjoint.

To successfully harness this global basin, we employ a rigorous **coarse-to-fine Frequency Annealing strategy**. We establish that while low-frequency moments provide the long-range attraction necessary for global tracking, they lack fine grained precision. By dynamically adjusting the active frequency bandwidth—systematically transitioning from coarse boundaries to precise structural alignments—we guide the underlying tracker into an accurate final pose. Our spectral loss serves as a general-purpose objective function that is agnostic to the underlying deformation model. We demonstrate its efficacy on two prevalent non-rigid parameterizations: sparse control points driven continuously by neural MLPs [36], and control points optimized directly via explicit displacements [12]. By integrating our global supervision into these distinct architectures, we show that it can guide the underlying tracker from extreme initial displacements – which cause standard photometric losses to fail – towards a highly accurate final pose, without requiring modifications to the deformation models themselves.

Our contributions are:

- **Spectral Moment Loss:** A novel, global objective function for 3DGS that provides non-vanishing directional gradients, effectively eliminating the “vanishing gradient” problem inherent to localized photometric losses under large spatial misalignments.
- **Principled Frequency Annealing:** A systematic optimization schedule derived from a first-principles analysis of phase wrapping. By progressively expanding the active frequency bandwidth from coarse to fine, we effectively smooth the high-frequency ambiguities of the spatial loss landscape. This significantly broadens the basin of attraction, bridging large spatial misalignments before refining high-frequency structural details.

- **Initialization-Robust Tracking:** We demonstrate the versatility of our global formulation across both synthetic and real-world datasets. By seamlessly integrating our spectral loss with diverse deformation representations (MLPs and sparse control points) and standard local objectives (L_2 and LPIPS), we consistently improve tracking stability. Our method successfully recovers complex deformations and survives severe camera misalignments, where standard appearance-based objectives fail.

2 Related Work

The development of SpectralSplats intersects with two primary research trajectories: the parameterization of Dynamic 3D Scene Reconstruction, and the shaping of Frequency-Guided Optimization Landscapes.

2.1 Dynamic and Deformable 3D Gaussian Splatting

Following the seminal work on static 3DGS [16], splat-based representations were rapidly extended to *dynamic* scenes [4, 6, 18, 21, 26, 29, 30, 32, 34–36, 38, 39]. The core challenge is to model the temporal evolution of Gaussian parameters while preserving temporal coherence. A dominant paradigm is *canonicalization*, which pairs a static canonical set of Gaussians with a time-varying deformation model. Such systems are typically trained either end-to-end from video [4, 38, 39] or via a two-stage pipeline that first initializes a canonical representation and then tracks per-frame deformations [6, 32]. Our setting aligns with the latter: we focus on deformation-based matching across frames, assuming a reliable initialization of the canonical scene.

Tracking dynamic scenes is inherently under-constrained and prone to geometric artifacts. To make tracking tractable and enforce temporal coherence, prior work commonly injects structural priors into the deformation model. Coordinate-based MLPs are frequently used to learn continuous displacement fields, prioritizing smoothness and coherence [21, 30, 36]. To accelerate training and inference speeds, approaches utilize structured grid encodings [3, 7, 32]. To further regularize these fields, recent methods have moved toward explicit geometric constraints like sparse control points [2, 12], while DynMF [18] utilizes low-dimensional neural motion factorization. Recent advancements in online tracking have further pushed the boundaries of this paradigm; [15] utilizes incremental 2D Gaussian Splatting [10] for efficient online 6-DoF object pose estimation, while FeatureSLAM [31] integrates foundation model features into the 3DGS rasterization pipeline for real-time semantic tracking.

While these structural design choices improve temporal consistency and rendering quality, they fundamentally assume that gradients from a photometric objective remain informative. Consequently, they do not resolve the optimization failure that occurs when the rendered object is spatially disjoint from its true image location. Our SpectralSplats framework is complementary to these

motion models; it provides a global supervisory signal that can guide any of the aforementioned parameterizations toward alignment from poor initializations.

To bypass this global search problem, domain-specific parameterizations heavily restrict the solution space. Human-centric methods such as HUGS [17] leverage SMPL [24] to optimize body pose deformations. Similarly, GART [20] proposes a canonical articulated template, extending the rigidity of bone-transformations to 3DGS primitives. While these articulated priors yield strong performance when the category assumption holds, they are brittle to initialization errors that place the template outside the local photometric basin. Our SpectralSplats framework is complementary to these motion models; it provides a global supervisory signal that can guide any of the aforementioned parameterizations toward alignment from poor initializations.

2.2 Frequency Analysis and Annealing in Neural Rendering

The interplay between spectral analysis and neural optimization has been a focal point of recent research, particularly regarding the "spectral bias" of neural networks. While high-frequency components are essential for capturing fine-grained detail, they often induce a rugged loss landscape, complicating the optimization of geometric parameters.

Frequency for Representation Quality. To mitigate these instabilities, several works have proposed managing spectral bandwidth to improve reconstruction fidelity. In the implicit domain, SAPE [9] modulates the frequency of positional encodings spatially, preventing noise-induced minima in smooth regions. With the shift to explicit Gaussian primitives, similar principles have been applied to regularize structure: FreGS [37] employs progressive frequency regularization to mitigate densification artifacts, while PGDGS [11] adopts progressive Gaussian densification for sparse-view reconstruction. Lavi et al. [19] structure the scene into hierarchical Laplacian pyramid subbands to decouple low-frequency geometry from high-frequency residuals. Crucially, these methods leverage frequency decomposition primarily for *level-of-detail* control and static representation quality.

Frequency for Geometric Optimization. Beyond representation, frequency analysis offers a powerful tool for shaping the optimization landscape. In the context of NeRF [27], BARF [22] utilized spectral annealing of positional encoding to widen the basin of attraction for camera registration, while MomentsNeRF [1] leveraged moment constraints for few-shot supervision. We transpose these insights to the domain of dynamic 3DGS. However, rather than annealing positional encodings, we propose Spectral Moment Supervision directly on the rendered output. This effectively bypasses the vanishing gradient problem inherent in spatial losses, creating a global basin of attraction that guides Gaussians even from zero-overlap initializations. Crucially, to avoid the phase-wrapping traps inherent in high frequencies, we introduce a principled Frequency Annealing schedule. While prior methods motivated linearly scaling frequency schedules heuristically through Neural Tangent Kernel [13] theory or signal bandwidth blurring [22, 28] we formally derive our annealing schedule from first principles.

3 Method

We present SpectralSplats, a framework for robust dynamic tracking that replaces standard spatial photometric errors with a spectral objective. We first formalize the “vanishing gradient” failure mode inherent to 3DGS tracking, establish the spectral-spatial duality of our objective, and then introduce our principled Spectral Moment Supervision and Frequency Annealing schedule.

3.1 Differentiable Gaussian Tracking and the Vanishing Gradient

A 3D Gaussian Splatting scene is parameterized by a set of primitives $\mathcal{G} = \{G_i\}$, each defined by a 3D mean μ_i , covariance Σ_i , opacity α_i , and spherical harmonics coefficients c_i . The rasterization function $\mathcal{R}$ projects these 3D primitives onto the 2D image plane to produce a rendering $\mathbf{I}_{\text{rend}}$.

In a tracking context, we assume a static canonical model $\mathcal{G}_{\text{ref}}$ is given. We seek a set of motion parameters Θ (e.g., representing a rigid transformation $\mathbf{T} \in SE(3)$ or neural deformation weights) that parameterize a deformation function $\mathcal{D}$. This function acts on the canonical model to produce a displaced scene: $\mathcal{G}_{\text{def}} = \mathcal{D}(\mathcal{G}_{\text{ref}}; \Theta)$. The rasterization function $\mathcal{R}$ then projects these deformed 3D primitives onto the 2D image plane to produce the rendering $\mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) = \mathcal{R}(\mathcal{D}(\mathcal{G}_{\text{ref}}; \Theta))$, which we aim to align with an observed target image $\mathbf{I}_{\text{gt}}$. To formally analyze the optimization landscape, we treat the image domain continuously and define the standard objective as minimizing the photometric difference over all 2D spatial coordinates $\mathbf{p}$:

$$\mathcal{L}_{\text{photo}}(\Theta) = \frac{1}{2} \int |\mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) - \mathbf{I}_{\text{gt}}(\mathbf{p})|_2^2 d\mathbf{p} \quad (1)$$

The Vanishing Gradient Problem. Before diving into the formal analysis, the core intuition behind this failure mode is remarkably simple: standard photometric tracking compares pixels locally. Because a Gaussian primitive only influences a compact spatial footprint, it must physically overlap with the target structure to receive a meaningful update. If the initial displacement is large enough that there is strictly zero overlap, moving the Gaussian slightly in any direction does not alter the total image loss. Because a small local translation yields absolutely zero change in the photometric error, the gradient evaluates exactly to zero. The loss is high, but as simulated in Fig. 2 (Col 1), the local optimization landscape is entirely flat, leaving the optimizer stranded.

To rigorously derive this “locality trap”, let us isolate the optimization of a rendered Gaussian and its corresponding true target signal in the image. By expanding the derivative of the squared error for this source-target pair, we can decompose its gradient contribution into two distinct components:

$$\nabla_{\Theta} \mathcal{L}_{\text{photo}}^{\text{target}} = \underbrace{\int \mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) \nabla_{\Theta} \mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) d\mathbf{p}}_{\text{Self-Term}} - \underbrace{\int \mathbf{I}_{\text{gt}}(\mathbf{p}) \nabla_{\Theta} \mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) d\mathbf{p}}_{\text{Target Supervision}} \quad (2)$$

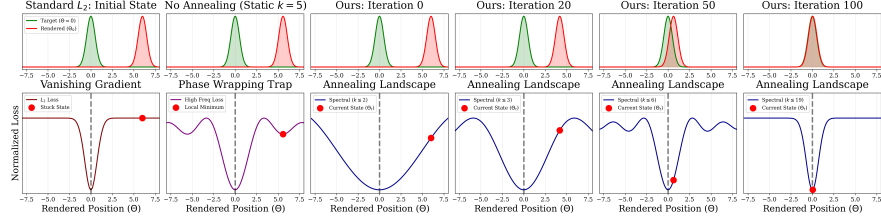


Fig. 2: Breaking the Locality Trap: A 1D Optimization Analysis. We simulate the optimization landscape (bottom) for aligning a rendered 1D Gaussian pulse (top, red) to a target (top, green) under a large initial spatial displacement ($\theta_0 = 6$). **Standard L_2 (Col 1):** Photometric objectives implicitly rely on spatial overlap; without it, the gradient strictly vanishes, leaving the optimizer stranded. **No Annealing (Col 2):** Projecting the loss onto a static, high-frequency spectral basis ($k = 5$) ensures the gradient no longer vanishes globally, but introduces severe phase-wrapping that traps the optimizer in false local minima. **Ours (Cols 3-6):** Spectral Moment Supervision with Frequency Annealing. By restricting initial supervision to low frequencies, we construct a globally convex basin of attraction that provides a valid, directional gradient from any initialization. As the spatial error strictly decreases, our principled annealing schedule safely expands the active bandwidth, seamlessly transitioning the landscape to achieve high-frequency spatial precision without phase-wrapping.

This decomposition highlights the fundamental flaw in tracking with highly localized spatial functions. The *Self-Term* can be rewritten via the chain rule as $\frac{1}{2} \nabla_{\theta} \int \mathbf{I}_{\text{rend}}^2 d\mathbf{p}$. For translations parallel to the image plane, this operation preserves the total footprint mass of the rendered object, making the integral strictly invariant to the motion parameter θ and its derivative exactly zero. While depth translations do yield a non-zero derivative due to perspective projection, in the absence of target overlap, this gradient merely acts to minimize the rendering footprint, driving the object to shrink by moving far away from the camera. In neither case does the self-term provide a directional signal toward the true target. The tracking signal therefore relies entirely on the *Target Supervision* cross-term. However, if θ positions the rendered Gaussian such that it is spatially disjoint from its true target location in $\mathbf{I}_{\text{gt}}$, the product of $\mathbf{I}_{\text{gt}}(\mathbf{p})$ and the spatial boundaries of the rendered object $\nabla_{\theta} \mathbf{I}_{\text{rend}}(\mathbf{p}; \theta)$ is zero everywhere. Consequently, the gradient contribution pulling the object to its true destination vanishes completely. This mathematical trap is further enforced by the 3DGS architecture itself: to maintain real-time performance, the rasterizer splits the screen into 16×16 tiles and culls primitives using a 99% confidence interval [16], forcefully zeroing out gradients for targets outside the immediate tile vicinity.

Crucially, this vanishing gradient means that even though the photometric error is large, the loss cannot decrease because the local gradient landscape is flat. Worse, when viewing the entire loss $\mathcal{L}_{\text{photo}}$ over a complete scene, the overall gradient $\nabla_{\theta} \mathcal{L}_{\text{photo}}$ does not evaluate to zero. The misaligned Gaussian inevitably overlaps with incorrect content in $\mathbf{I}_{\text{gt}}$ (e.g., background clutter). Because the true gradient has vanished, the optimizer receives only corrupted gradients driven by

this spurious overlap. Rather than pulling the Gaussian toward its target, these gradients anchor the object to the background.

3.2 Image Moments and Spectral Duality

To resolve the strict locality of the spatial loss, we shift our objective from direct pixel-to-pixel comparisons to the alignment of *image moments*. Intuitively, computing a moment is equivalent to multiplying the image by an auxiliary static field $F(\mathbf{p})$ and integrating the result. If we choose a field that varies continuously across the entire spatial domain – such as a sinusoidal wave or a polynomial function – this projection acts as a global coordinate system.

This global integration breaks the locality trap. Let us define a simple moment-matching objective between the rendered image and the target:

$$\mathcal{L}_{\text{moment}}(\Theta) = \frac{1}{2} (M_{\text{rend}}(\Theta) - M_{\text{gt}})^2, \quad (3)$$

where $M = \int \mathbf{I}(\mathbf{p})F(\mathbf{p})d\mathbf{p}$ denotes the projection of an image $\mathbf{I}$ onto the field.

The gradient of this objective with respect to the motion parameters is:

$$\nabla_{\Theta} \mathcal{L}_{\text{moment}}^{\text{target}} = (M_{\text{rend}}(\Theta) - M_{\text{gt}}) \nabla_{\Theta} M_{\text{rend}}(\Theta). \quad (4)$$

Unlike the spatial cross-term that vanished, this gradient consists of two reliably non-zero components. First, provided the global field $F(\mathbf{p})$ does not repeat values across the spatial domain, the scalar projections of the disjoint rendered and target objects will differ, ensuring a valid error magnitude: $(M_{\text{rend}}(\Theta) - M_{\text{gt}}) \neq 0$. (As we will discuss next, guaranteeing this non-repeating property is a central challenge when employing periodic spectral bases). Second, assuming simple translation, the directional vector – the gradient of the rendered moment itself – evaluates to:

$$\nabla_{\Theta} M_{\text{rend}}(\Theta) = \int \nabla_{\Theta} \mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) F(\mathbf{p}) d\mathbf{p} = \int \mathbf{I}_{\text{rend}}(\mathbf{p}; \Theta) \nabla_{\mathbf{p}} F(\mathbf{p}) d\mathbf{p}, \quad (5)$$

where the final equality follows by first applying the chain rule for spatial translation ($\nabla_{\Theta} \mathbf{I}_{\text{rend}} = -\nabla_{\mathbf{p}} \mathbf{I}_{\text{rend}}$) and subsequently performing integration by parts. By ensuring the spatial derivative of the field $\nabla_{\mathbf{p}} F(\mathbf{p})$ is non-zero in the region of interest, this integral provides a valid directional signal. Therefore, even if the rendered object and the target are completely disjoint, the optimizer “feels” the slope of the field at the object’s current location. The scalar difference provides the magnitude of the pull, while the field gradient dictates the direction, enabling robust registration without explicit feature correspondences. While various global kernels exist (e.g., the standard geometric and orthogonal polynomial moments utilized in classic correspondence-free shape alignment [5]), we propose using *Spectral Moments* governed by complex sinusoidal functions, as they provide geometrically meaningful phase shifts under translation. We define a spectral moment for a discrete 2D spatial frequency vector ω_{k_x, k_y} (where k_x and k_y are the horizontal and vertical frequency indices) as:

$$\mathcal{M}(k_x, k_y; \mathbf{I}) = \sum_{\mathbf{p}} \mathbf{I}(\mathbf{p}) \cdot \exp(-j\omega_{k_x, k_y}^T \mathbf{p}). \quad (6)$$

Unlike the standard spatial L_2 loss, this operation pointwise multiplies the image by a complex sinusoid and integrates it over the entire domain.³

Spectral duality. An appealing property of choosing this specific spectral basis is the direct mathematical link it provides back to our original spatial objective. By Parseval’s theorem, the sum of squared errors evaluated across a *complete* orthogonal frequency basis is strictly equivalent to the spatial L_2 loss. However, this equivalence presents a fundamental paradox: if we were to optimize the full spectral basis simultaneously, the objective would perfectly reconstruct the spatial loss landscape, thereby inheriting the exact same vanishing gradient and local minima traps we set out to avoid. As demonstrated in Fig. 2 (Col 2), high-frequency components introduce severe phase-wrapping that fragments the global basin of attraction, trapping the optimizer in false local minima. Therefore, to harness the non-vanishing global gradients of the spectral domain while ultimately achieving the strict equivalence and precision of the spatial loss, we cannot use the full basis statically; we must dynamically control the active frequency bandwidth during optimization.

3.3 Deriving the Frequency Annealing Schedule

To navigate this trade-off between global convergence and spatial precision, we introduce a coarse-to-fine Frequency Annealing schedule. While isolated low frequencies create a global basin of attraction, they inherently lack precision. Because the spatial gradient of a spectral loss scales with the frequency magnitude ($\nabla \mathcal{L} \propto \omega \sin(\omega d)$), the gradient signal of low-frequency moments diminishes as the spatial error d approaches zero. High frequencies are strictly required for fine-grained alignment, but as established, activating them prematurely induces phase-wrapping that traps the optimizer (Fig. 2(Col 2)). To achieve global convergence, we must systematically transition from coarse to fine frequencies. As shown in Fig. 2(Cols 3-6), this principled progression seamlessly transforms the optimization landscape: it leverages a globally convex basin to rescue the stranded Gaussians from their initial zero-overlap state, and progressively sharpens into a high-precision spatial target without introducing false minima.

We formalize this Frequency Annealing schedule from first principles. For a spatial misalignment vector $\mathbf{d}_t$ at optimization step t , the spectral loss at frequency ω is convex only if the induced phase shift does not wrap, i.e., $|\omega^T \mathbf{d}_t| < \pi$ (see Supp. Mat. for a detailed derivation). This defines a dynamic stability condition: the maximum active frequency magnitude $\|\omega_{\max}(t)\|$ must be inversely proportional to the magnitude of the spatial error $\|\mathbf{d}_t\|$. When this condition is met and $\omega^T \mathbf{d}_t$ is small, the spectral loss landscape $E(\mathbf{d}) = 2 - 2 \cos(\omega^T \mathbf{d})$ is well-approximated by a Taylor expansion as a quadratic bowl, $E(\mathbf{d}) \approx (\omega^T \mathbf{d})^2$.

³ While a naive evaluation over a dense frequency grid is computationally prohibitive, this formulation natively supports highly efficient computation via 2D FFT.

In this strongly convex regime, the gradient is directly proportional to the spatial displacement ($\nabla E \propto \mathbf{d}$). That is, gradient descent naturally takes update steps that scale with the remaining distance to the target. This guarantees the spatial estimation error decays exponentially: $\|\mathbf{d}_t\| \leq \|\mathbf{d}_0\|\gamma^t$ for a convergence factor $\gamma \in (0, 1)$. To maintain the phase-wrapping constraint $\|\omega_{\max}(t)\| \propto 1/\|\mathbf{d}_t\|$, the active frequency magnitude must therefore expand exponentially as γ^{-t} . Because standard spectral grids organize frequencies logarithmically, such that $\|\omega_k\| \propto 2^k$, an exponential growth in frequency magnitude necessitates a strictly *linear* expansion of the active frequency index $k(t)$ over time.

Crucially, this derivation provides a rigorous, first-principles optimization foundation for the empirically successful annealing schedules introduced in Nerfies [28] and later utilized in BARF [22]. While prior works motivated a linearly scaling frequency index heuristically – through the lens of Neural Tangent Kernel (NTK) theory [28] or signal bandwidth blurring [22] – our dynamic phase-wrapping analysis shows exactly *why* it works: a linearly scaling index on a logarithmic grid represents the theoretical upper bound for safe frequency expansion during spatial alignment.

To implement this expansion without injecting discontinuous shocks into the loss landscape, we adopt the smooth cosine weighting function from these works, applying a time-dependent weight $w_k(t)$ to each spectral moment:

$$w_k(t) = \frac{1 - \cos(\pi \cdot \text{clamp}(\alpha(t) - k, 0, 1))}{2} \quad (7)$$

where $\alpha(t)$ scales linearly from 0 to K over the optimization iterations to govern the active bandwidth, and $k \in \{0, \dots, K-1\}$ is the index of the specific frequency band. This formulation allows each successively higher frequency to gracefully fade into the objective and remain active once its transition window is complete.

Conservative Frequency Expansion. While our derivation establishes that $\alpha(t)$ can safely scale linearly over a logarithmic grid (yielding exponential growth of ω) under ideal linear convergence, real-world tracking scenarios are rarely ideal. In practice, background clutter, occlusions, and complex deformations often disrupt ideal exponential error decay. To account for these unpredictable optimization dynamics, we implement a two-fold conservative scheduling strategy. First, following the empirical practices introduced in BARF [22], we enforce a strictly low-frequency "warm-up" phase where $\alpha(t)$ is held constant for the initial optimization iterations. This ensures the

optimizer has time to exploit the widest global basin and resolve the severe initial spatial misalignments before high-frequency complexities are introduced. Note that if the warm-up period is too long, the high frequency detail may not be recovered correctly (Fig. 3). Second, once expansion begins, we linearly scale



Fig. 3: A long low-frequency "warm-up" phase (right) leads to loss of high frequency details (tail), compared to a shorter "warm-up" phase (left).

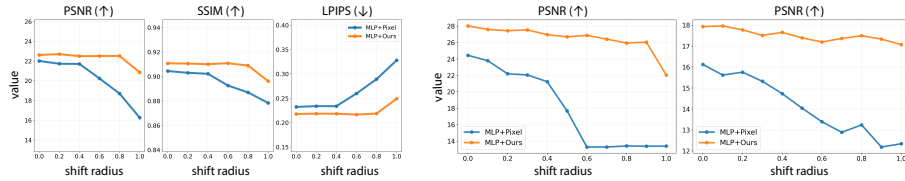


Fig. 4: (Left) Effect of spatial shift in GART, showing average PSNR, SSIM, and LPIPS versus shift radius; pixel supervision degrades rapidly, while our method remains stable. (Right) PSNR results on SC4D [14, 33] for training and novel views; pixel loss deteriorates under misalignment, whereas our method maintains stable performance.

the frequencies themselves, rather than linearly scaling across their logarithmic indices. Because linear growth is bounded well below exponential growth, this practical relaxation guarantees that we stay beneath the phase-wrapping threshold ($|\omega^T \mathbf{d}_t| < \pi$) throughout the optimization process. This delayed, sub-exponential expansion enhances tracking robustness. We note that our framework operates on foreground Gaussians only. Foreground masks can be obtained once per scene using standard 3D segmentation tools [8], enabling foreground optimization and background recombination during rendering.

4 Experiments

Dataset. We evaluate our method on two datasets: 4D Animations generated by SC4D [33] using assets from Consistent4D [14], and the Dog dataset from GART [20]. SC4D provides a controlled setting, where clean and high-quality dynamic 3DGS models are used to render the supervision videos from known views, and provide the ground truth resting 3DGS [16]. As a result, the appearance of the 3DGS initialization is well aligned with the supervision. To assess performance in a more realistic scenario, we experiment on the GART Dog dataset. Monocular videos, collected from the 2022 National Dog Show and Adobe Stock, are used by GART to reconstruct a unified rest-pose 3DGS model per asset, together the estimated 3DGS and real videos provide our source gaussians and target supervision. This real-world setup includes lighting inconsistencies and unknown camera views, leading to noticeable deviations between the supervision video and the input 3DGS model in pose and appearance. As an additional real-world validation, we evaluate our method on a thrown basketball sequence from [26], demonstrating robustness in a challenging fast-motion scenario.

Deformation Parameterization. Across both datasets, we optimize a deformation model that predicts per-frame displacements of Gaussian *control points*, selected using standard procedures [12]. To evaluate our method across different tracking architectures, we test two variants for moving these control points: (1) *MLP Parameterization*, where a TimeNet [12] network predicts the time-dependent deformation of each control point, and (2) *Direct Morph Field*, where we optimize the positional offsets and rotations of the control points directly.

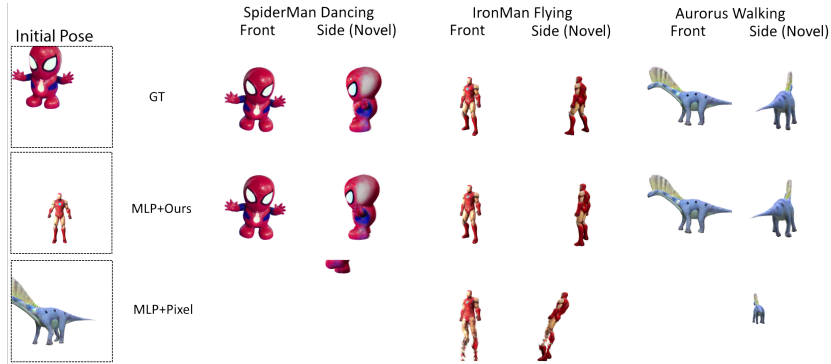


Fig. 5: Qualitative comparison on the SC4D data under initial spatial shift (radius = 0.5). For three characters and animations, we show the initial pose, GT, *MLP+Ours* and *MLP+Pixel*, both without LPIPS. While pixel-only optimization fails to recover correct pose and may drift the object outside the frame, our method achieves better alignment and more coherent structure in both training and novel views.

Training Objective & Baselines. We implement the Frequency Annealing schedule to resolve initial misalignments. As established in Section 3.2, by Parseval’s theorem, optimizing over the full spectral basis is mathematically equivalent to the spatial L_2 loss. Therefore, for computational efficiency, once the annealed spectral moments secure local spatial overlap, we naturally transition to standard spatial losses for high-frequency refinement. Specifically, we utilize a pixel loss across both datasets, and LPIPS for the synthetic SC4D dataset. To isolate the contribution of the spectral phase, we compare against baselines relying solely on spatial objectives. To further evaluate our method, we implemented the global-loss baselines Pyramid of Gaussians (PoG) [25] and Euclidean Distance Transform (DT) [23], and compare their performance on both GART (Table 2) and the teaser experiment (Supp. Mat.). To ensure a strictly fair comparison, other loss components and regularization terms follow GSGD [2] and are applied identically to both our method and the baselines (see Supp. Mat.).

4.1 SC4D Experiment

To simulate spatial misalignment arising from occlusions, drift, or fast motion, we shift the initial 3DGS model in a random direction with increasing offsets, artificially reducing its overlap with the supervision. We evaluate both pixel-only supervision (*MLP+Pixel*) and our spectral scheme (*MLP+Ours*), and report results for both the training and novel views. The right panel of Figure 4 shows the mean PSNR as a function of shift radius. As the misalignment increases, the gap between pixel supervision and our method widens. Pixel-based PSNR rapidly decreases, especially under larger shifts, while our method remains considerably more stable. Importantly, this trend holds for both the training and novel views, indicating improved generalization. In the Appendix we provide

Table 1: Metric evaluations on data generated by SC4D [33] under spatial shift (radius = 0.5). Our method shows large improvement in alignment to the source view across parameterizations and pixel losses, with a decent improvement in novel view quality.

	Loss	LPIPS ↓	PSNR ↑	SSIM ↑	NV-LPIPS ↓	NV-PSNR ↑	NV-SSIM ↑
MLP w. LPIPS	Pixel	0.0852	23.6051	0.9409	0.1153	18.3244	0.9304
	Ours	0.0489	27.1453	0.9546	0.0948	19.1977	0.9331
MLP w/o LPIPS	Pixel	0.1806	17.6748	0.9108	0.2023	14.0424	0.9107
	Ours	0.0516	26.6992	0.9507	0.1331	17.3960	0.9159
Direct w. LPIPS	Pixel	0.3133	11.6619	0.8297	0.2443	12.3815	0.8727
	Ours	0.2000	15.4626	0.8701	0.2501	12.7916	0.8675
Direct w/o LPIPS	Pixel	0.2289	16.1268	0.8562	0.2774	12.0662	0.8491
	Ours	0.1868	17.8558	0.8789	0.2640	12.5106	0.8598

additional plots evaluating SSIM and LPIPS as a function of shift radius, as well as for supervising with *multiple views*. Crucially, even in the case of multi-view supervision the pixel loss collapses under spatial misalignment, while our SpectralSplats remains robust.

Qualitative results are shown in Figure 5 for a representative shift radius of 0.5. We observe that our method remains consistent, whereas pixel-only optimization often fails to recover correct pose and structure, and in some cases even pushes the object outside the frame. We further report quantitative results in Table 1 for this shift radius, evaluated across different deformation parameterizations (MLP and direct morph field). We additionally test replacing the pixel-loss phase with LPIPS supervision. Across almost all configurations, our method consistently improves PSNR, SSIM, and LPIPS in both training and novel views. Overall, these results demonstrate the robustness of our method across parameterizations, spatial loss choices, and evaluation views.

4.2 GART Experiment

We follow a similar spatial misalignment experiment to SC4D, shifting the initial 3DGS model in a random direction with increasing radii. We train using either pixel-only supervision (*MLP+Pixel*) or our spectral scheme (*MLP+Ours*). The left panel of Figure 4 shows the metric means as a function of the shift radius. As the misalignment increases, pixel-only training degrades significantly, while our method remains considerably more stable, highlighting the sensitivity of pixel supervision to poor initialization. In the Appendix we provide per sample plots evaluating PSNR, SSIM and LPIPS as a function of shift radius.

For a representative shift of 0.6, quantitative results are reported in Table 2 and qualitative comparisons in Figure 6. Our SpectralSplats outperforms the pixel baseline on almost all dogs and metrics, improving the mean PSNR (22.05 vs. 20.15), SSIM (0.907 vs. 0.891), and LPIPS (0.216 vs. 0.258). Visual results for *French* and *Shiba* further show better pose recovery and sharper structure with

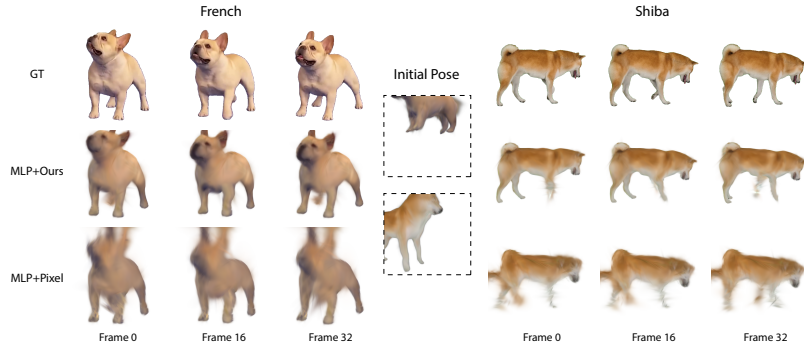


Fig. 6: Qualitative comparison under initial spatial misalignment (radius = 0.6) on three frames of two representative dogs from GART. Pixel-only optimization exhibits blur and incorrect alignment, while our method better recovers pose and structure.

Table 2: GART comparison under a spatial shift radius of 0.6, showing consistent improvements over pixel-only and global-loss baselines (best values in bold).

Dog	LPIPS ↓				PSNR ↑				SSIM ↑			
	Pixel	Ours	PoG	DT	Pixel	Ours	PoG	DT	Pixel	Ours	PoG	DT
Alaskan	0.2875	0.2664	0.3120	0.2847	20.0056	20.6333	16.5123	19.1629	0.8793	0.8845	0.8501	0.8745
Shiba	0.2749	0.1788	0.1805	0.1807	20.8241	25.3568	23.9165	24.7880	0.9069	0.9344	0.9319	0.9301
Hound	0.3406	0.2514	0.3361	0.2781	16.2762	19.4494	13.5665	17.8886	0.8372	0.8769	0.8202	0.8645
Corgi	0.1164	0.1100	0.1137	0.1082	25.4472	26.5250	24.8425	25.8148	0.9497	0.9561	0.9508	0.9533
French	0.3038	0.2339	0.2820	0.2436	17.6107	20.8778	17.0710	20.2074	0.8888	0.9106	0.8835	0.9064
English	0.2367	0.2418	0.4253	0.2408	21.2707	21.3255	8.9685	20.4019	0.8939	0.8938	0.7494	0.8871
Pitbull	0.2505	0.2340	0.2881	0.2440	19.6348	20.2401	14.7755	18.1753	0.8851	0.8937	0.8514	0.8811
Mean	0.2586	0.2166	0.2768	0.2257	20.1528	22.0583	17.0933	20.9199	0.8915	0.9071	0.8625	0.8996

MLP+Ours, while *MLP+Pixel* exhibits poor alignment. Overall, both quantitative and qualitative results demonstrate that our method consistently improves robustness to initial spatial misalignment in realistic 3DGS reconstructions.

4.3 Fast Motion Optimization

Beyond the initial misalignment setting, we demonstrate our method’s performance on a fast-motion scenario where the first frame is already aligned with the rendered object. We use a real-world video from [26], optimizing with monocular supervision. We further increase the difficulty by temporally subsampling input frames. To apply our method, we separate foreground and background Gaussians, optimize only the foreground Gaussians, and recombine them at rendering time.

Figure 7 shows the ground-truth frames and corresponding masks, together with the results obtained using our method and pixel-based supervision. We note that the dataset-provided masks contain artifacts. Despite the correct first-frame alignment, the thrown basketball undergoes large motion between frames, resulting in a zero-overlap regime. In this scenario, pixel-based supervision fails, whereas our method maintains accurate alignment despite the imperfect masks.

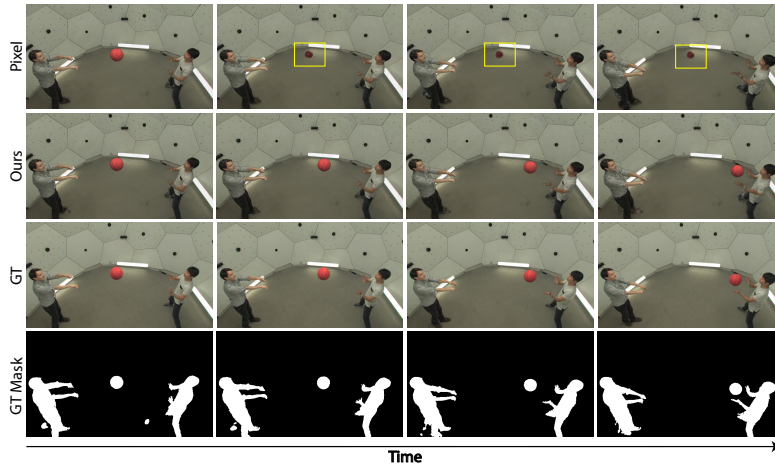


Fig. 7: Qualitative comparison on a thrown basketball sequence from [26]. Large inter-frame motion creates a zero-overlap regime where pixel-based supervision fails, while our method recovers the correct alignment despite imperfect dataset masks (bottom).

5 Conclusion

We presented SpectralSplats, a robust, model-agnostic framework that resolves the vanishing gradient problem inherent to dynamic 3DGS tracking. By replacing localized spatial losses with Spectral Moment supervision and a principled frequency annealing schedule, we allow tracking pipelines to recover from extreme spatial misalignments at initialization, bypassing the need for manual alignment or restrictive category-specific priors.

While SpectralSplats significantly expands the basin of attraction, its current formulation assumes access to a pre-initialized canonical asset, restricting its scope to model-based tracking. A compelling future direction is extending this frequency-guided optimization beyond tracking to full dynamic scene reconstruction, where canonical geometry and motion are jointly optimized from uncalibrated video. Additionally, exploring alternative moment types to capture highly complex dynamics remains an exciting avenue for future research.

Acknowledgments

We thank Gal Harari and Ido Sobol for their help with code and visualization, and Matan Atzmon for insightful discussions. Avigail Cohen Rimmon acknowledges support of the Miriam and Aaron Gutwirth Memorial Fellowship. Mirela Ben Chen acknowledges support from the Israel Science Foundation (grant No. 1073/21). Or Litany acknowledges support from the Israel Science Foundation (grant No. 624/25) and the Azrieli Foundation Early Career Faculty Fellowship. This research was supported by the Council for Higher Education in Israel under the Moonshot Project.

Bibliography

- [1] AlMughrabi, A., Marques, R., Radeva, P.: Momentsnerf: Leveraging orthogonal moments for few-shot neural rendering. arXiv preprint arXiv:2407.02668 (2024)
- [2] Bekor, Y., Harari, G.M., Perel, O., Litany, O.: Gaussian see, gaussian do: Semantic 3d motion transfer from multiview video. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)
- [3] Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
- [4] Chen, Z., Yang, J., Huang, J., de Lutio, R., Esturo, J.M., Ivanovic, B., Litany, O., Gojcic, Z., Fidler, S., Pavone, M., Song, L., Wang, Y.: Omnire: Omni urban scene reconstruction. pp. 1486–1505. 13th International Conference on Learning Representations, ICLR 2025 (2025)
- [5] Domokos, C., Nemeth, J., Kato, Z.: Nonlinear shape registration without correspondences. IEEE Transactions on pattern analysis and machine intelligence **34**(5), 943–958 (2011)
- [6] Duisterhof, B.P., Mandi, Z., Yao, Y., Liu, J.W., Seidenschwarz, J., Shou, M.Z., Deva, R., Song, S., Birchfield, S., Wen, B., Ichnowski, J.: DeformGS: Scene flow in highly deformable scenes for deformable object manipulation. WAFR (2024)
- [7] Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12479–12488 (2023)
- [8] Fuji Tsang, C., Hu, A., Perel, O., Kolve, C., Shugrina, M.: Artisans: Interactive tools for gaussian splat selection with ai and human in the loop. arXiv:2602.10173 (2026)
- [9] Hertz, A., Perel, O., Giryes, R., Sorkine-Hornung, O., Cohen-Or, D.: Sape: Spatially-adaptive progressive encoding for neural optimization. Advances in Neural Information Processing Systems **34**, 8820–8832 (2021)
- [10] Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
- [11] Huang, H., Zhang, Z., Wu, G., Wang, R.: Pgdgs: Improving few-shot 3d gaussian splatting with progressive gaussian densification. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
- [12] Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4220–4230 (2024)

- [13] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems* **31** (2018)
- [14] Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360° dynamic object generation from monocular video. In: *The Twelfth International Conference on Learning Representations* (2024), <https://openreview.net/forum?id=sPurdFGepF>
- [15] Jin, Y., Prasad, V., Jauhri, S., Franzius, M., Chalvatzaki, G.: 6dope-gs: Online 6d object pose estimation using gaussian splatting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8032–8043 (2025)
- [16] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
- [17] Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 505–515 (2024)
- [18] Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. *ECCV* (2024)
- [19] Lavi, Y., Segre, L., Avidan, S.: Frequency-aware gaussian splatting decomposition. *arXiv preprint arXiv:2503.21226* (2025)
- [20] Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19876–19887 (2024)
- [21] Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8508–8520 (2024)
- [22] Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5741–5751 (2021)
- [23] Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 7708–7717 (2019)
- [24] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866 (2023)
- [25] Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. In: *European Conference on Computer Vision*. pp. 154–169. Springer (2014)
- [26] Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: *3DV 2024*
- [27] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
- [28] Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: *Proceed-*

- ings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021)
- [29] Seidenschwarz, J., Zhou, Q., Duisterhof, B.P., Ramanan, D., Leal-Taixé, L.: Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. In: 2025 International Conference on 3D Vision (3DV). pp. 1012–1021. IEEE (2025)
 - [30] Sun, J., Jiao, H., Li, G., Zhang, Z., Zhao, L., Xing, W.: 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. pp. 20675–20685 (2024)
 - [31] Thirgood, C., Mendez, O., Ling, E., Storey, J., Hadfield, S.: Featureslam: Feature-enriched 3d gaussian splatting slam in real time. arXiv preprint arXiv:2601.05738 (2026)
 - [32] Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20310–20320 (June 2024)
 - [33] Wu, Z., Yu, C., Jiang, Y., Cao, C., Wang, F., Bai, X.: Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In: European Conference on Computer Vision. pp. 361–379. Springer (2024)
 - [34] Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S.: Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In: European Conference on Computer Vision. pp. 156–173. Springer (2024)
 - [35] Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
 - [36] Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
 - [37] Zhang, J., Zhan, F., Xu, M., Lu, S., Xing, E.: Fregs: 3d gaussian splatting with progressive frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21424–21433 (2024)
 - [38] Zhou, H., Shao, J., Xu, L., Bai, D., Qiu, W., Liu, B., Wang, Y., Geiger, A., Liao, Y.: Hugs: Holistic urban 3d scene understanding via gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21336–21345 (2024)
 - [39] Zhou, X., Lin, Z., Shan, X., Wang, Y., Sun, D., Yang, M.H.: Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21634–21643 (2024)
 - [40] Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6365–6373 (2017)