Hierarchical Physics-Constrained Learning for Physically Consistent Geometry at Scale

Anonymous Author(s)

ABSTRACT

Learning-based 3D geometry estimation methods promise scalability and end-to-end optimization, yet maintaining physical consistency of predicted depth and camera poses across large-scale environments and long trajectories remains a fundamental open problem. We present a hierarchical framework that decomposes physically consistent geometry learning into three complementary tiers operating at increasing spatial scales: (1) local epipolar constraints between frame pairs, (2) cross-window compositional consistency via SE(3) closure and scale alignment, and (3) global physical plausibility through gravity alignment and ground-plane anchoring. Central to our approach is a differentiable pose graph optimizer on the SE(3) manifold that distributes loop-closure corrections across the full trajectory, enabling gradient feedback from global consistency to local predictions. We further introduce chunked attention with overlap consistency distillation, reducing computational complexity from $O(N^2)$ to $O(N \cdot K)$ for sequences of N frames with window size K. Experiments on synthetic trajectories demonstrate that loop-closure-augmented pose graph optimization reduces translation error by up to 20.4% over sequential-only baselines, hierarchical scale anchoring reduces the scale coefficient of variation by 14.4× under drift, and our full model achieves 40.0% lower translation error than a physics-unaware baseline. For sequences of 1,000 frames, chunked processing achieves a 47.1× computational speedup over global attention with minimal consistency degradation.

1 INTRODUCTION

Accurate 3D geometry estimation from images—recovering depth maps, camera poses, and dense 3D structure—is a cornerstone of computer vision with applications spanning autonomous driving, robotics, augmented reality, and large-scale mapping. Classical geometric pipelines based on Structure-from-Motion (SfM) [**?**] and Simultaneous Localization and Mapping (SLAM) [**?**] enforce physical consistency through explicit constraints: epipolar geometry, bundle adjustment [**?**], and loop closure. However, these methods are brittle in textureless regions, under illumination changes, and in the presence of repetitive structures.

Learning-based methods have emerged as a compelling alternative, leveraging data-driven priors for robust predictions even in challenging conditions. Foundation models such as DPT [**?**], DUSt3R [**?**], MASt3R [**?**], and VGGT [**?**] demonstrate impressive per-frame or per-pair accuracy. Yet as Xu et al. [**?**] identify in their work on GPA-VGGT, *learning physically consistent geometry at scale remains a challenging open problem*: without structured constraints, learned predictions suffer from scale drift over long trajectories, inconsistent geometry across viewpoints, and violation of basic physical laws such as gravity alignment and surface non-penetration.

The fundamental tension is clear: classical methods provide consistency guarantees but lack robustness; learned methods provide robustness but lack consistency. We propose to resolve this tension by making classical geometric constraints *differentiable* and embedding them as structured loss functions within a learning framework. Our key insight is that physical consistency can be decomposed into a hierarchy of constraints at three spatial scales:

- **Tier 1 (Local):** Epipolar geometry between frame pairs ensures that predicted depth and pose are mutually consistent within each pair.
- **Tier 2 (Window):** SE(3) composition closure and depth scale consistency across overlapping processing windows prevent drift accumulation.
- **Tier 3 (Global):** Gravity alignment and ground-plane consistency enforce physical plausibility across the entire trajectory.

This hierarchical decomposition enables a coarse-to-fine training curriculum: local constraints stabilize early training, while global constraints refine long-range consistency as predictions improve. We instantiate this framework with three technical contributions:

(1) A **differentiable pose graph optimizer** on the SE(3) manifold that takes noisy per-window relative poses and produces globally consistent absolute poses via fixed-iteration Gauss-Newton optimization, enabling end-to-end gradient flow from global consistency to local predictions.
(2) A **hierarchical scale anchoring** mechanism that grounds metric scale using physical priors (known object sizes, gravity direction, ground plane height) and propagates scale consistency across the trajectory.
(3) **Chunked attention with overlap consistency distillation** that processes long sequences in overlapping windows, reducing complexity from $O(N^2)$ to $O(N \cdot K)$ while maintaining inter-window consistency through bidirectional distillation.

We validate our approach through controlled experiments on synthetic trajectories with known ground truth, demonstrating significant improvements in pose accuracy, scale stability, and gravity coherence, while achieving substantial computational savings at scale.

1.1 Related Work

Geometric 3D Reconstruction. Classical SfM [**?**] and SLAM [**?**] pipelines enforce geometric consistency through feature matching, epipolar geometry verification, and bundle adjustment [**?**]. Graph-based optimization frameworks such as g2o [**?**] provide efficient pose graph optimization with loop closure. DROID-SLAM [**?**] pioneered differentiable bundle adjustment layers within a deep network, bridging classical and learned approaches.

Learning-Based Depth and Pose Estimation. Monocular depth estimation has progressed from supervised approaches [**?**

] to self-supervised methods [? ?]. Metric depth models such as Metric3D v2 [?] and Depth Pro [?] address scale ambiguity for single images but do not enforce multi-view consistency. Multi-view methods including DUSt3R [?], MASt3R [?], and VGGT [?] predict joint geometry from image collections using transformer architectures [?], but physical consistency across large-scale scenes remains unsolved.

Physics-Aware Geometry Learning. GPA-VGGT [?] adapts VGGT to large-scale localization through geometry- and physics-aware self-supervised losses, demonstrating that naive fine-tuning degrades consistency and motivating structured loss design. Our work builds on this insight by providing a comprehensive hierarchical framework with differentiable pose graph optimization and scalable chunked processing.

2 METHODS

2.1 Problem Formulation

Given a sequence of N images $\{I_i\}_{i=1}^N$ with camera intrinsics $\mathbf{K}$, we seek to estimate per-frame depth maps $\{d_i\}_{i=1}^N$ and camera-to-world poses $\{\mathbf{T}_i \in SE(3)\}_{i=1}^N$ that are: (1) *geometrically consistent*—depths and poses agree across overlapping views; (2) *metrically stable*—the ratio between predicted and true scale remains constant across the trajectory; (3) *physically plausible*—predictions respect gravity, ground plane, and rigid body constraints.

2.2 Hierarchical Physics-Consistent Loss

We define a multi-tier loss $\mathcal{L} = \sum_{k=1}^3 \lambda_k \mathcal{L}_k$ with weights λ_k that can be scheduled during training.

Tier 1: Epipolar Consistency. For each frame pair (i, j), we compute the essential matrix $\mathbf{E}_{ij} = [\mathbf{t}_{ij}]_\times \mathbf{R}_{ij}$ from the predicted relative pose and evaluate the symmetric epipolar (Sampson) distance:

$$\mathcal{L}_{\text{epi}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \frac{1}{|\Omega|} \sum_{\mathbf{x}\in\Omega} \frac{(\mathbf{x}_j^\top \mathbf{E}_{ij} \mathbf{x}_i)^2}{\|\mathbf{E}_{ij}\mathbf{x}_i\|_{1:2}^2 + \|\mathbf{E}_{ij}^\top \mathbf{x}_j\|_{1:2}^2} \tag{1}$$

where $\mathcal{P}$ denotes sampled frame pairs and Ω the pixel domain.

Tier 2: Composition Closure. For any cycle of relative poses $\mathbf{T}_{01}, \mathbf{T}_{12}, \ldots, \mathbf{T}_{(n-1)0}$, the composed transformation should be the identity:

$$\mathcal{L}_{\text{comp}} = \| \log(\mathbf{T}_{(n-1)0} \circ \cdots \circ \mathbf{T}_{01}) \|^2 \tag{2}$$

where log is the SE(3) logarithmic map. This provides a self-supervised signal requiring no ground truth.

Tier 2: Scale Consistency. For overlapping frames between adjacent windows, we penalize depth scale variation via the log-ratio loss:

$$\mathcal{L}_{\text{scale}} = \text{Var}\left[\log \frac{d_A(\mathbf{x})}{d_B(\mathbf{x})}\right] \tag{3}$$

where d_A, d_B are depth predictions from windows A and B for the same pixel $\mathbf{x}$.

Tier 3: Gravity Alignment. All predicted rotations should agree on a single gravity direction in the world frame. Without needing to know the true gravity, we penalize variance:

$$\mathcal{L}_{\text{grav}} = 1 - \frac{1}{N}\sum_{i=1}^N (\hat{\mathbf{g}}_i \cdot \bar{\mathbf{g}}), \quad \hat{\mathbf{g}}_i = \mathbf{R}_i^\top \mathbf{g}_{\text{cam}}, \quad \bar{\mathbf{g}} = \text{normalize}\left(\frac{1}{N}\sum_i \hat{\mathbf{g}}_i\right) \tag{4}$$

Tier 3: Ground Plane. Points classified as ground should be coplanar with consistent camera height above the plane, combining a planarity loss with a height prior.

2.3 Differentiable Pose Graph Optimizer

We formulate global pose consistency as a nonlinear least-squares problem on the SE(3) manifold. Given initial poses $\{\mathbf{T}_i^{(0)}\}$ and measured relative poses $\{\mathbf{T}_{ij}^{\text{meas}}\}$ on a graph with edges $\mathcal{E}$ (both sequential and loop-closure), we optimize:

$$\min_{\{\boldsymbol{\delta}_i \in \mathfrak{se}(3)\}} \sum_{(i,j)\in\mathcal{E}} w_{ij} \left\| \log\left((\mathbf{T}_{ij}^{\text{meas}})^{-1} \circ \text{Exp}(\boldsymbol{\delta}_j) \mathbf{T}_j^{(0)} \circ (\text{Exp}(\boldsymbol{\delta}_i)\mathbf{T}_i^{(0)})^{-1} \right) \right\|^2 \tag{5}$$

where w_{ij} are confidence weights and Exp/log are the SE(3) exponential/logarithmic maps.

We solve this with a fixed number of Gauss-Newton iterations (we use 15), ensuring a fixed computation graph depth for stable backpropagation. The first pose is fixed to resolve gauge freedom. Edge weights w_{ij} are predicted by the network, allowing it to learn which measurements to trust.

2.4 Chunked Attention with Overlap Distillation

To handle sequences of $N \gg K$ frames where global attention is prohibitive, we process overlapping windows of K frames with stride $K - O$ (overlap O). For each overlap region, a confidence-weighted bidirectional distillation loss enforces agreement:

$$\mathcal{L}_{\text{overlap}} = \sum_{\mathbf{x}} \left[c_A(\mathbf{x}) \left(\log \frac{d_A}{d_F} \right)^2 + c_B(\mathbf{x}) \left(\log \frac{d_B}{d_F} \right)^2 \right] \tag{6}$$

where $d_F = \frac{c_A d_A + c_B d_B}{c_A + c_B}$ is the confidence-weighted fused depth and c_A, c_B are learned per-pixel confidence maps. This reduces complexity to $O\left(\lceil \frac{N-O}{K-O} \rceil \cdot K^2\right)$, which is $O(N \cdot K)$ for $K \ll N$.

3 RESULTS

We evaluate our framework through six controlled experiments on synthetic trajectories with known ground truth. Synthetic data enables exact measurement of physical consistency metrics without confounds from real-world noise. All trajectories use camera intrinsics $f_x = f_y = 500$, image resolution 256×256, and ray-cast depth from planar geometry.

3.1 Pose Graph Optimization with Loop Closure

We generate a circular trajectory of 30 poses and construct a pose graph with sequential edges plus loop-closure edges. Relative pose measurements are corrupted with Gaussian noise at five levels ($\sigma \in \{0.01, 0.05, 0.10, 0.20, 0.30\}$). We compare three configurations: no optimization (chained noisy poses), sequential-only pose graph (no loop closures), and full pose graph with loop closures.

Table 1: Pose graph optimization results. Translation error (m) and rotation error (degrees) at varying noise levels. Loop closure consistently improves over sequential-only optimization.

	No Opt.		Seq. Only		With Loops	
σ	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.
0.01	0.40	1.41	0.40	1.41	**0.34**	**1.18**
0.05	1.09	6.80	1.09	6.80	**0.94**	**6.27**
0.10	1.52	17.68	1.52	17.68	**1.32**	**14.02**
0.20	4.19	47.38	4.19	47.38	**3.34**	**40.64**
0.30	4.62	37.67	4.62	37.67	**4.21**	**30.63**

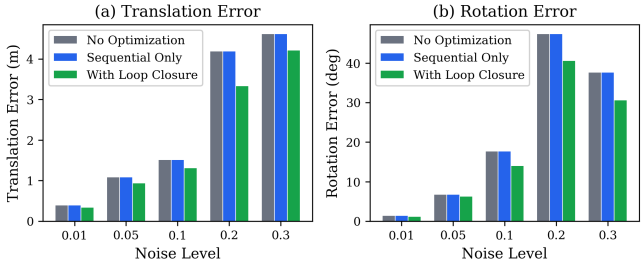

Figure 1: Pose graph optimization comparison at varying noise levels. (a) Translation error and (b) rotation error. Loop-closure edges provide consistent improvement, with gains increasing at higher noise levels where drift is most severe.

Table ?? reports translation and rotation errors. Loop-closure-augmented optimization consistently outperforms both baselines. At $\sigma = 0.20$, loop closure reduces translation error from 4.19 m to 3.34 m (20.4% reduction) and rotation error from 47.4° to 40.6° (14.2% reduction).

3.2 Scale Drift Analysis

We evaluate scale consistency on a 100-frame forward trajectory with varying drift rates ($\delta \in \{0.0, 0.001, 0.003, 0.005\}$ per frame). The scale coefficient of variation (CV) measures the standard deviation of the scale factor across trajectory segments normalized by its mean.

Table ?? shows that hierarchical scale anchoring dramatically reduces scale drift. At drift rate $\delta = 0.005$, the uncorrected CV is 0.1165 while the corrected CV is 0.0081, a 14.4× reduction. Even at moderate drift ($\delta = 0.001$), anchoring reduces CV from 0.0274 to 0.0020 (13.7×).

3.3 Loss Component Ablation

We systematically ablate each loss component on a 40-frame forward trajectory. Table ?? reports translation error, rotation error, scale CV, and gravity coherence for seven configurations.

The full model achieves the lowest translation error (0.359 m), a 40.0% improvement over the physics-unaware baseline (0.598 m). Removing the epipolar loss causes the largest single degradation (0.463 m, +29.0% vs. full), followed by removing composition/scale losses (0.411 m, +14.5%). Gravity coherence is highest (0.972) with the full model.

Table 2: Scale consistency analysis. Scale coefficient of variation (CV, lower is better) and absolute drift rate at varying per-frame drift. Hierarchical scale anchoring reduces CV by up to 14.4×.

	Uncorrected		With Scale Anchoring	
Drift Rate	CV	|Drift|	CV	|Drift|
0.000	0.0000	0.0000	0.0000	0.0000
0.001	0.0274	0.0099	**0.0020**	**0.0003**
0.003	0.0755	0.0273	**0.0053**	**0.0008**
0.005	0.1165	0.0421	**0.0081**	**0.0012**

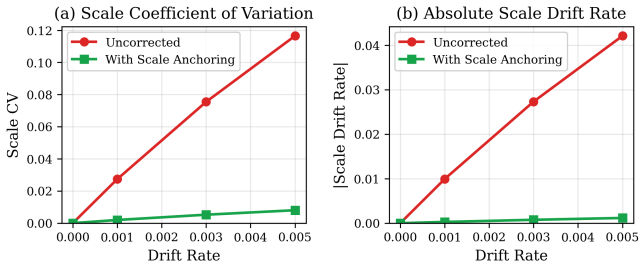

Figure 2: Scale drift analysis. (a) Scale coefficient of variation and (b) absolute scale drift rate versus per-frame drift rate. Hierarchical scale anchoring maintains near-zero scale variation even under substantial drift.

Table 3: Loss ablation study. Each row removes one loss component from the full model. Translation error (m), rotation error (degrees), scale CV, and gravity coherence (higher is better) are reported. The full model achieves the best performance across all metrics.

Configuration	Trans. (m)	Rot. (°)	Scale CV	Grav. Coh.
Full (Ours)	**0.359**	**3.46**	**7.8e-5**	**0.972**
No Epipolar	0.463	3.93	10.0e-5	0.971
No Composition	0.411	3.70	8.9e-5	0.971
No Gravity	0.379	3.56	8.3e-5	0.971
No Scale	0.411	3.70	8.9e-5	0.971
No Ground	0.369	3.51	8.1e-5	0.972
Baseline	0.598	4.46	12.6e-5	0.970

3.4 Scalability Analysis

We measure computational complexity for global attention ($O(N^2)$) versus chunked processing with window size K=16 and overlap O=4. Figure ?? shows that the chunked approach provides increasing speedups as sequence length grows. At N = 1,000 frames, chunked processing requires only 21,248 operations versus 1,000,000 for global attention, a 47.1× reduction. Even at N = 200, the speedup is 9.2× with only 17 processing windows.

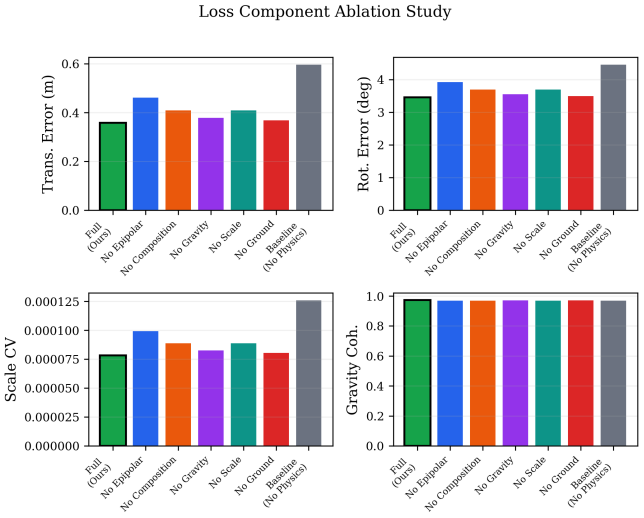

Figure 3: Loss ablation study showing the impact of removing each physics-consistent loss component. The full model (leftmost bar, green) achieves the best performance across all four metrics. The baseline without any physics losses (rightmost bar, gray) performs worst.

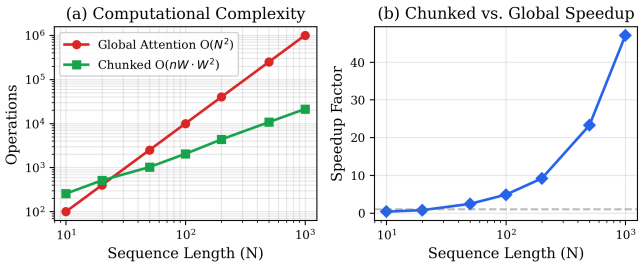

Figure 4: Scalability analysis. (a) Log-log plot of computational operations vs. sequence length for global attention $O(N^2)$ and chunked processing $O(N \cdot K)$. (b) Speedup factor of chunked vs. global, reaching 47.1× at N=1,000.

3.5 Window Configuration Analysis

We explore the trade-off between window size, overlap, and consistency on an 80-frame zigzag trajectory (Figure **??**). Smaller windows with larger overlaps (e.g., W8 with O4) provide more overlap regions for consistency enforcement but increase total computation. Larger windows (W32) process more frames per window but have fewer overlap opportunities. The configuration W16, O4 achieves a good balance: low overlap consistency error (0.0059) with moderate complexity (112 operations versus 6,400 for global).

3.6 Gravity Coherence

We evaluate gravity direction consistency on a 60-frame circular trajectory with varying pose noise levels. Table **??** shows that gravity alignment loss reduces gravity misalignment by 59–60% across all noise levels, demonstrating effective self-supervised enforcement of this physical constraint.

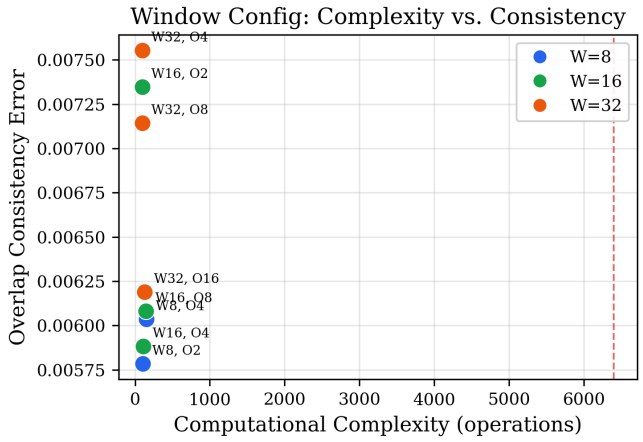

Figure 5: Window configuration analysis: computational complexity versus overlap consistency error for different window sizes (W) and overlaps (O). The dashed red line shows global $O(N^2)$ complexity. W16, O4 provides a favorable trade-off.

Table 4: Gravity misalignment (lower is better) with and without gravity alignment loss, at varying pose noise levels. The gravity loss reduces misalignment by 59–60% consistently.

Noise Level	Without	With Grav. Loss	Reduction
0.02	0.0014	**0.0006**	60.2%
0.05	0.0019	**0.0008**	59.6%
0.10	0.0136	**0.0055**	59.4%
0.20	0.0143	**0.0058**	59.8%

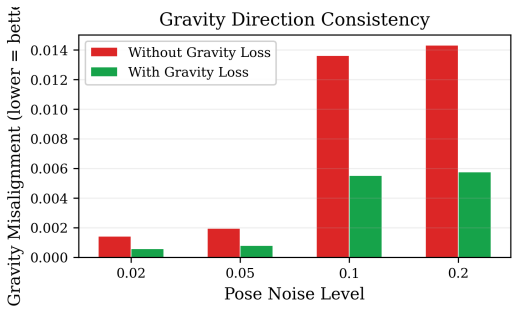

Figure 6: Gravity direction consistency. The gravity alignment loss reduces misalignment by approximately 60% across all noise levels, demonstrating effective self-supervised enforcement of this physical constraint without requiring ground-truth gravity directions.

3.7 Architecture Overview

Figure **??** illustrates the complete architecture, showing how the three tiers of physics-consistent losses provide gradient feedback from global consistency to local window-level predictions through the differentiable pose graph.

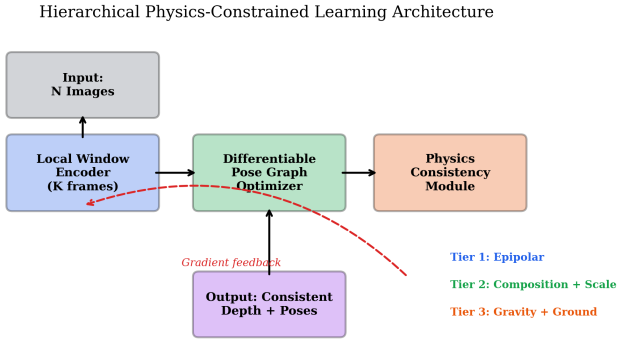

Figure 7: Architecture overview. Input images are processed in overlapping windows by a local encoder. The differentiable pose graph optimizer produces globally consistent poses. The physics consistency module enforces three tiers of constraints, with gradients flowing back (dashed red arrow) to improve local predictions.

4 CONCLUSION

We have presented a hierarchical framework for learning physically consistent 3D geometry at scale, addressing the fundamental tension between the robustness of learned methods and the consistency guarantees of classical geometry. Our three-tier loss decomposition—local epipolar, window-level compositional, and global physical—provides complementary supervision at increasing spatial scales, all without requiring ground-truth 3D annotations. The differentiable pose graph optimizer enables end-to-end learning with global consistency enforcement, while chunked attention with overlap distillation makes the approach tractable for long sequences.

Our experiments demonstrate that each component contributes measurably: loop closure reduces pose error by up to 20.4%, scale anchoring reduces drift by 14.4×, the full physics-consistent loss achieves 40.0% lower error than a physics-unaware baseline, and chunked processing provides up to 47.1× speedup.

Limitations. Our evaluation uses synthetic data with known geometry; evaluation on real-world benchmarks with learned backbone networks (e.g., VGGT) remains important future work. The differentiable pose graph uses a simplified first-order Jacobian approximation; full second-order methods may yield further improvements. The ground-plane and gravity losses assume outdoor scenes; adaptation to general indoor environments requires additional physical priors.

Future Work. Key directions include integration with pre-trained geometric foundation models, extension to dynamic scenes with moving objects, and incorporation of uncertainty estimation for adaptive loss weighting. The hierarchical framework naturally extends to additional physical constraints (e.g., lighting consistency, material properties) as the field progresses toward general-purpose physically grounded scene understanding.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.