LATENT BUNDLE ADJUSTMENT

Anonymous authors
Paper under double-blind review

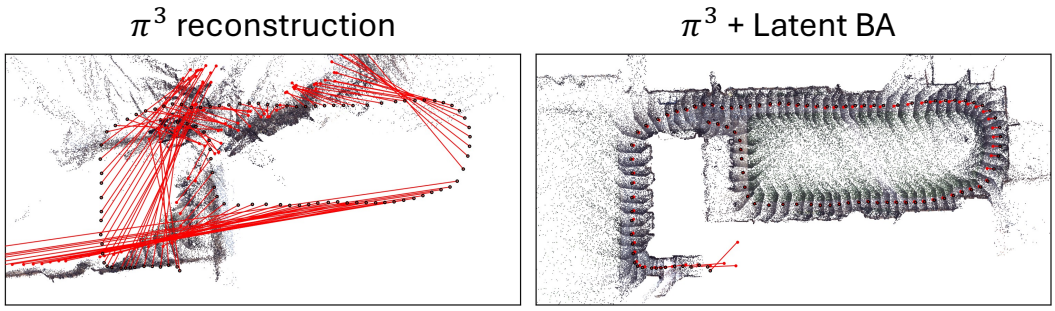


Figure 1: Latent Bundle Adjustment refines camera poses and dense geometry of feed-forward reconstruction models by optimising intermediate latent representations via reprojection error. Red lines represent the distance between the estimated and GT camera pose.

ABSTRACT

We present a method for dense bundle adjustment of 3D scenes by optimising the latent representations of feed-forward reconstruction models, such as VGGT-Omega and π^3. These learned models can be more robust than traditional Structure-from-Motion (SfM) pipelines, particularly on challenging scenes. However, combining their robustness with the accuracy of classical bundle adjustment remains challenging and is typically limited to poses and sparse points or patches. We propose a novel parametrisation of the bundle adjustment problem by optimising a set of tokens within a feed-forward reconstruction model. Specifically, we solve for latent updates that minimise reprojection error given a set of sparse correspondences between the input frames. Latents jointly encode the dense scene structure and camera poses, and the change in representation provides implicit regularisation during optimisation, allowing for refinement of dense geometry from sparse matches. We evaluate our approach on 3D reconstruction and camera pose estimation tasks across different datasets, demonstrating state-of-the-art performance. We further analyse parametrisations, both qualitatively and quantitatively.

1 INTRODUCTION

Reaching consensus on the state of the world from distinct viewpoints is fundamental to how humans perceive, navigate, and act. In computer vision, 3D reconstruction from 2D images involves understanding scene geometry from finite-resolution projections. Traditionally, structure-from-motion (SfM) has relied on multi-view geometry. After finding pixel correspondences and initialising points and camera poses, bundle adjustment (BA) iteratively optimises predicted projections against measurements. Explicit parametrisation of poses and points yields a simple projection model, but points lack direct correlation, which is evident in degenerate configurations such as unconstrained depth without translation. BA lacks a prior on the distribution of natural scenes and observers.

Generalisable priors over geometry can be learned via data-driven models. Monocular depth prediction and relative pose estimation can add robustness to BA via initialisation or a direct prior over their values. However, these priors provide a single estimate given limited information and ambiguity, which biases SfM. Predicting distributions over depth captures the space of viewer-centric geometry, but not its correlation across views or with camera parameters.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

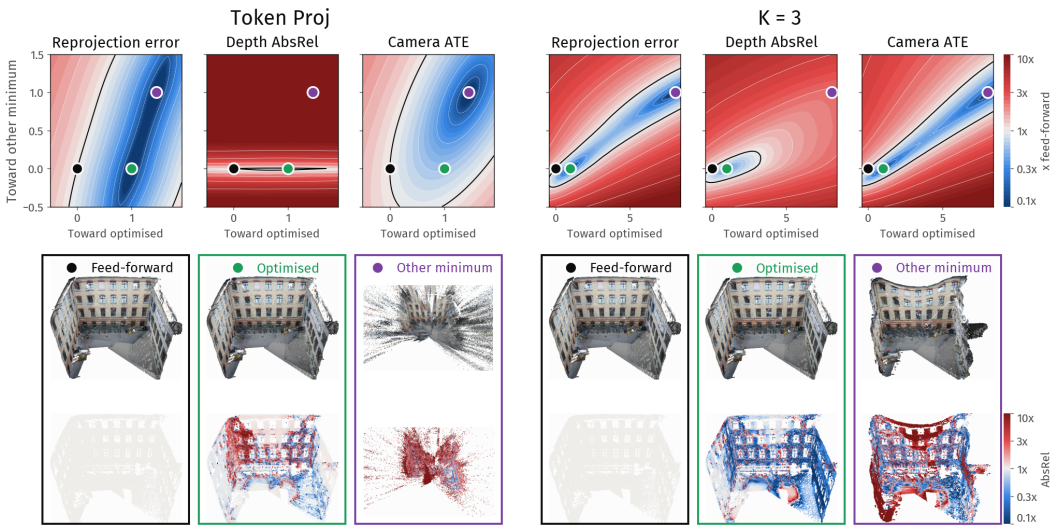


Figure 2: Reprojection error basin for latent representation optimisation of π^3. "Token proj" is the latent before projection to pixel space while "$k = 3$" adds three inter-frame attention blocks. Axes span two directions of latent space and colour corresponds to an improvement multiplier over the feed-forward result. Gradient descent finds solutions with improved accuracy despite latent representations containing unnatural solutions with lower reprojection error. Deeper layers show a wider receptive field and achieve both improved depth and pose accuracy.

Feed-forward reconstruction models predict camera poses, intrinsics, and per-pixel 3D geometry directly from images. This can be viewed as amortised inference, where initialisation, correspondence, triangulation, and refinement are implicitly handled by a neural network trained offline. While this achieves robust estimation, precision lags that of optimisation, which reasons at pixel-level rather than token-level, does not limit the number of computational steps, and can incorporate additional information such as intrinsics. The outputs of feed-forward reconstruction models directly correspond to the parametrisation of BA and are often used as initialisations or the centre of a prior. However, this single estimate discards the joint prior implicitly learned by the network. BA can be performed on sparse matches to achieve accurate poses, but without these correlations, it inherits the same inclination to unnatural geometry and requires a second stage to refine dense geometry.

Just as natural images are hypothesised to lie on a low-dimensional manifold, so may the space of geometric scenes and observers. Feed-forward reconstruction models have learned rich structure about the world, including 3D consistency, auto-calibration, and the distribution of camera poses. The network outputs are an image-conditioned point estimate of geometry and poses, while the network itself encodes the prior that produced it. To retain this knowledge, we propose a change in parametrisation and optimise the latent representations of feed-forward reconstruction models. Due to the connectivity of such networks, this provides a coupling of poses and geometry, while the weights learned on data offline provides general knowledge for guiding the optimisation landscape.

Even when feed-forward initialisation is far from 3D-consistent, optimising via gradient descent can achieve globally consistent results as seen in Figure 1. Local optimisation of overparameterised latent representations can find solutions with consistent 3D geometry, despite unnatural solutions achieving lower reprojection error, as illustrated in Figure 2. The implicit prior combined with error from sparse matches alone enables joint refinement of dense geometry and poses in a single stage.

In summary, we propose a bundle adjustment framework that uses latent representations of feed-forward reconstruction models as a parametrisation to achieve an implicit, joint prior over poses and geometry. Along the way, we provide intuitions and comparisons of the implicit regularisation of feed-forward reconstruction models across architectures, representation depth, and test-time parametrisations. Lastly, we show that a straightforward system with sparse reprojection error can achieve state-of-the-art performance on pose estimation and dense geometry estimation.

2 Related Work

Structure-from-motion (SfM) recovers camera motion and 3D geometry given images. SfM pipelines solve for poses and sparse 3D points by comparing expected projections with measurements derived from pixel matching Snavely et al. (2006); Agarwal et al. (2011). Bundle adjustment (BA) Triggs et al. (1999) jointly refines these variables via iterative optimisation and remains the most precise method to estimate camera poses at scale Schönberger & Frahm (2016); Pan et al. (2024). However, its scalability rests on ignoring correlations among 3D points, so noise, non-discriminative pixels, and low parallax result in unnatural reconstructions. Data-driven priors can improve robustness on the measurement side by learning correspondence DeTone et al. (2018); Sarlin et al. (2020); Teed & Deng (2021); Lindenberger et al. (2023); Edstedt et al. (2024), while depth prediction from a single image Yin et al. (2023); Yang et al. (2024); Piccinelli et al. (2024) sidesteps ambiguity in scale and the need for parallax. Modern SfM systems leverage these pretrained correspondence and depth priors Li et al. (2025); Huang et al. (2025); Pataki et al. (2025; 2026) to provide robustness where traditional SfM systems fail, including wide baselines, pure rotation, and low texture. However, predicting geometry from a single view is ambiguous and committing to a single estimate contributes bias, so alternatives predict the space of depths via distributions Bloesch et al. (2018); Dexheimer & Davison (2023) or basis functions Tang & Tan (2019); Graham & Novotny (2020); Leroy et al. (2026). Regardless, per-view depth ignores correlations across frames, poses, and intrinsics, while joint priors may better model the space of scenes.

Feed-forward reconstruction models are neural networks that directly predict per-pixel scene geometry and camera parameters. Pioneered by DUSt3R Wang et al. (2024), two-view models implicitly solve correspondence, triangulation, relative pose estimation, and camera calibration. MASt3R Leroy et al. (2024) shows that pixel-level matching improves pose accuracy. Multi-view transformers Wang et al. (2025b; 2026b); Keetha et al. (2026); Lin et al. (2026); Wang et al. (2026a); Burzio et al. (2026) predict for all images simultaneously, yielding greater robustness. However, precision lags that of optimisation, so their outputs are used to initialise BA or provide direct priors Duisterhof et al. (2025); Wang et al. (2025b); D'Urso et al. (2026); Pan et al. (2026). For example, VGGT uses its point tracking and variable initialisation with BA to improve pose accuracy. This avoids poor local minima, but the BA parametrisation discards the joint prior and can produce unnatural geometry. In this work, we optimise the network's latent representations to retain an implicit, coupled prior.

Test-time optimisation adapts a model to a new example at inference. Efficient ways to fine-tune neural networks include prompt tuning Li & Liang (2021); Jia et al. (2022), low-rank adaptation of weights (LoRA) Hu et al. (2022), and hidden-state interventions Wu et al. (2024). Both Test3R Yuan et al. (2025) and LoRA3D Lu et al. (2025) adapt DUSt3R with prompt tuning and LoRA, respectively. TTT3R Chen et al. (2026) views recurrent updates of a fixed-size state as test-time optimisation. Self-Geometry Youn et al. (2026) finds that reprojection error alone is insufficient when applying LoRA to feed-forward reconstruction models. In contrast, we show that optimising latent representations instead of weights is advantageous, and that reprojection error alone is sufficient.

Implicit regularisation describes how optimisation, architecture, and weight initialisation bias overparameterised networks toward particular solutions Neyshabur et al. (2015); Gunasekar et al. (2018); Jacot et al. (2018); Ulyanov et al. (2018). Probing feed-forward reconstruction models reveals emergent epipolar geometry, correspondence, and sparse computation Bratulić et al. (2026); Ziliotto Filippo & Campari (2026); Wang et al. (2025a). In this work, we explore how the representation optimised by gradient descent through feed-forward reconstruction models biases the solution space, and the implications for pose and geometry estimation.

3 Method

Given a set of N unordered RGB images $\mathbf{I}_n \in \mathbb{R}^{H\times W\times 3}, n \in \{1, ..., N\}$, a scene can be reconstructed via a feed-forward reconstruction model $\mathcal{F}_{\mathcal{FFM}}(\mathbf{I})$ producing per frame camera poses $\mathbf{T}_n \in \mathrm{SE}(3)$ in a unified coordinate frame and depth maps $\mathbf{D}_n \in \mathbb{R}^{H\times W}$.

Our objective is to refine the initial reconstruction to obtain more accurate poses and dense geometry $(\mathbf{T}^*, \mathbf{D}^*)$. We propose optimising a *latent representation* $\mathbf{Z} \mapsto (\mathbf{T}, \mathbf{D})$ that encodes geometry and camera poses. Given a set of sparse 2D-2D correspondences across frames and camera intrinsics $\mathbf{K}$, we minimise reprojection error with respect to $\mathbf{Z}^*$ and obtain $(\mathbf{T}^*, \mathbf{D}^*)$. This achieves dense

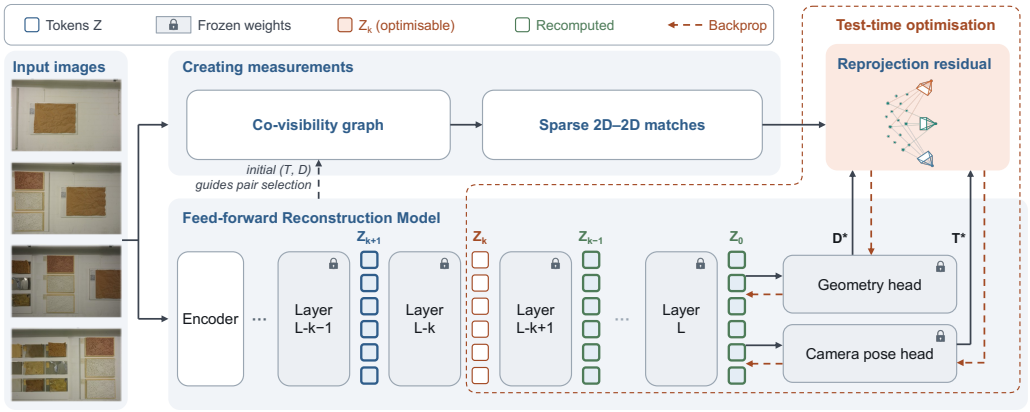


Figure 3: Given a set of images, we run a feed-forward reconstruction model to get an initial reconstruction that is used to extract co-visible pairs that are used for matching 2D-2D correspondences. We formulate a BA objective given these 2D measurements and optimise the reprojection residual using our latent parametrisation that implicitly couples dense geometry and camera poses.

refinement from sparse measurements. As shown in Figure 3, our pipeline starts by extracting a co-visibility set $\mathcal{O}$, followed by matching and pruning the co-visible pairs. Finally, we perform Latent-BA to refine dense geometry and camera poses.

3.1 Preliminary on feed-forward reconstruction models

In this work, we are concerned with the class of models that ingest a set of images in a single forward pass. These models are parameterised by weights θ, are composed of three main components: an encoder $\mathcal{E}$, a set of transformer blocks $\mathcal{G}$, and decoding heads $\mathcal{H}$.

The image encoder processes the images individually producing a tokenized representation of input images $\mathcal{E}(\mathbf{I}) = \mathbf{Z}_L$, where $\mathbf{Z}_k \in \mathbb{R}^{N \times t \times c}$ and t is the number of tokens per frame and c is the embedding size. We denote k, with $0 \leq k \leq L$, as the number of decoder layers counted from the output of $\mathcal{G}$, which allows us to express different frontends of different lengths L in a common convention. For simplicity of notation, we assume that per-frame tokens include additional register and camera tokens. Tokens are processed through the network via a series of L attention layers that alternate between inter- and intra-frame self-attention, forming a stack of attention blocks $\mathcal{G}$.

Lastly, geometry and camera poses, optionally intrinsics Wang et al. (2026a), are obtained via the decoding heads $\mathcal{H}(\mathbf{Z}_0) = (\mathbf{T}, \mathbf{D})$. While there exist different parametrisations of geometry, all can be expressed as depth maps with an associated $\mathbf{K}$ under a pinhole camera assumption. If the network predicts point-maps Wang et al. (2026b) $\mathbf{X}_n \in \mathbb{R}^{H \times W \times 3}$, $\mathbf{K}$ is recovered using the Weiszfeld algorithm Plastria (2011), while depth indexes the z-axis component of $\mathbf{X}_n$.

3.2 Bundle Adjustment

Bundle adjustment minimises the reprojection error r_{ij}^m, or the difference between a point observed in the image plane and its modelled projection. We adopt a depth map formulation where geometry is hosted per frame. Depth $\mathbf{D}_i(x_i)$ is sampled from at a 2D point x_i, backprojected to 3D, transformed from frame i to frame j, and projected onto the image plane. The error is then measured to the matched 2D point x_j:

$$r_{ij}^m(\mathbf{T}_i, \mathbf{T}_j, \mathbf{D}_i, x_i^m, x_j^m) = \pi_{\mathbf{K}}\big(\mathbf{T}_{ij}\, \pi_{\mathbf{K}}^{-1}(\mathbf{D}_i(x_i^m), x_i^m)\big) - x_j^m. \tag{1}$$

The optimisation objective $\mathcal{R}$ is defined as the sum of all costs ρ, such as Huber or Cauchy, over all the corresponding frames in $\mathcal{O}$ and their matches $\mathcal{M}$:

$$\mathcal{R}(\mathbf{T}, \mathbf{D}) = \sum_{(i,j)\in\mathcal{O}} \sum_{m\in\mathcal{M}_{ij}} \rho\left(r_{ij}^{m}(\mathbf{T}_i, \mathbf{T}_j, \mathbf{D}_i, x_i^m, x_j^m)\right). \tag{2}$$

Alternatively, this minimisation can be viewed as *Maximum A Posteriori* (MAP) estimation,

$$(\mathbf{T}^*, \mathbf{D}^*) = \arg\max_{\mathbf{T},\mathbf{D}} \underbrace{P(\mathbf{T}, \mathbf{D})}_{\text{Prior}} \prod_{(i,j)\in\mathcal{O}} \underbrace{P(x_j \mid \mathbf{T}_i, \mathbf{T}_j, \mathbf{D}_i)}_{\text{Reprojection Likelihood}}, \tag{3}$$

where we maximise the posterior consisting of the prior and likelihood over $(\mathbf{T}, \mathbf{D})$, where the product over matches is omitted for simplicity. In explicit BA, only a set of sparse depths $\mathbf{d} \subset \mathbf{D}$ with measurements and camera poses $\mathbf{T}$ are recovered. Providing a well-specified prior is challenging beyond centring around outputs from a network, which lacks correlations across variables.

3.3 LATENT BA

To inject the learned network prior into BA, we modify the explicit BA parametrisation of explicit $\mathbf{T}$ and $\mathbf{D}$ into an implicit one such that our parametrisation $\mathbf{Z}_k$ jointly encodes the scene geometry and camera pose via $\mathbf{Z} \mapsto (\mathbf{T}, \mathbf{D})$. Using the layer notation introduced previously, this is defined as:

$$\mathbf{Z}_k = \Phi(\mathbf{I}) = \mathcal{G}_{L-k} \circ \cdots \circ \mathcal{G}_1 \circ \mathcal{E}(\mathbf{I}), \tag{4}$$

$$(\mathbf{T}, \mathbf{D}) = \Psi(\mathbf{Z}_k) = \mathcal{H} \circ \mathcal{G}_L \circ \cdots \circ \mathcal{G}_{L-k+1}(\mathbf{Z}_k), \tag{5}$$

where $\mathcal{F}_{\mathcal{F}\mathcal{F}\mathcal{M}} = \Psi \circ \Phi$. The network is split at $\mathcal{G}_{L-k}$, with k providing different inductive biases for BA across layers and models. The remainder of the network Ψ is treated as function that can decode the implicit latent into explicit geometry.

The general Latent-BA MAP problem is described in Equation (6), which can be contrasted with the explicit BA MAP in Equation (3). In our method we the networks weights are frozen and not optimised, hence the prior term $P(\mathbf{Z}_k \mid \mathbf{Z}_k^{(0)})$, where $\mathbf{Z}_k^{(0)} = \Phi(\mathbf{I})$, only provides an initialisation for for $\mathbf{Z}_k$. Since the geometry and camera poses are recoverable via the latent, we define our likelihood of the x_j to be given $\mathbf{Z}_k$ and the implicit regularisation provided by the gradients flowing through the frozen suffix Ψ. In our experiments we find the small gradient steps and the smoothness of the optimisation manifold defined by Ψ provide sufficient regularisation for our MAP problem.

$$\mathbf{Z}_k^* = \arg\max_{\mathbf{Z}_k} \; P(\mathbf{Z}_k) \prod_{(i,j)\in\mathcal{O}} P(x_j \mid \Psi(\mathbf{Z}_k)), \tag{6}$$

$$(\mathbf{T}^*, \mathbf{d}^*) = \arg\min_{\mathbf{T},\mathbf{d}} \mathcal{R}(\mathbf{T}, \mathbf{d}) \quad \longrightarrow \quad \mathbf{Z}_k^* = \arg\min_{\mathbf{Z}_k} \mathcal{R}(\Psi(\mathbf{Z}_k)). \tag{7}$$

Given the Latent-BA MAP formulation, we reformulate the explicit BA reprojection residuals over poses and geometry to a function of latents in Equation (7). Unlike the explicit formulation, poses and depth are coupled via the parametrisation, and dense dense depth $\mathbf{D}^*$ can be refined from sparse measurements. Given the initialisation from the feed-forward pass, gradient descent can be used to update $\mathbf{Z}_k$, which couples geometry and poses, and find a local minima $\mathbf{Z}_k^*$, and the refined poses and geometry $(\mathbf{T}^*, \mathbf{D}^*) = \Psi(\mathbf{Z}_k^*)$. Compared to the explicit BA parametrisation, where camera poses and sparse geometry are coupled only through the reprojection residuals, Latent-BA explicitly couples poses and geometry through various levels: tokenisers connect pixels at the patch level, frame-wise attention communicates at the image level, and global self-attention provides a receptive field across all frames. We visualise how the choice of representation k in Figure 4 impacts coupling of log-depth across frames, showing larger k provides a more global view, while retaining local granularity. More k experiments are presented in Figure 6.

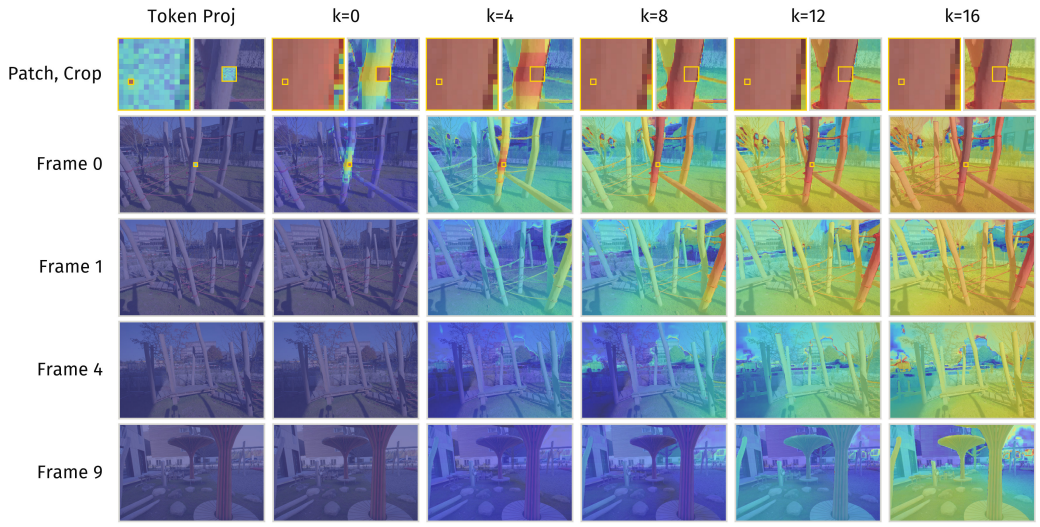


Figure 4: Visualisation of linearised coupling between a change in the selected pixel's log-depth and all log-depths $\mathbf{y}$ for different latent representation layers of π^3 at the feed-forward initialisation. Formally, this is one row of the kernel $\mathbf{K} = \mathbf{J}\mathbf{J}^T$ of gradient similarities, with Jacobian $\mathbf{J} = \partial \mathbf{y} / \partial \mathbf{Z}_k$, normalised within a column and shown on a square-root scale. For larger k, the receptive field becomes more global while still retaining well-defined responses locally.

3.4 Co-visibility graph construction and Matching

Similar to EPO D'Urso et al. (2026), we leverage our reconstruction prior to build a co-visibility graph between image pairs. Given an initial reconstruction of the scene, we exhaustively check if each frame's geometry projects inside other frame's view frustum. Pairs are chosen when they symmetrically satisfy a minimum thresholds for number of points in front and inside the 2D image plane. This creates an initial set of candidate pairs, with the co-visibility set refined after matching. Under the assumption that feed-forward reconstruction model can produce locally robust camera pose and geometry estimates, we obtain a sufficient set of candidates to match.

To minimise our geometric objective function, we require a set of correspondences between images. We opt for sparse feature matches since salient points provide more reliable measurements for bundle adjustment. Given the initial graph pairs, we run an matcher $\mathcal{F}_{\mathcal{M}}$ over candidate pairs $(i, j) \in \mathcal{O}$, which produces a set of sparse matches $\mathcal{M}$. Pairs are then pruned via fitting a relative pose between the pair using a RANSAC loop. Given the matched 2D points and the 3D points from the initial reconstruction, a pose is fitted with Umeyama algorithm Umeyama (1991) between candidate pairs. An inlier ratio threshold is then used to prune pairs with low consensus on the 3D point error between them. This keeps pairs that agree on their corresponding geometry.

4 Experiments

We provide qualitative and quantitative evaluations of our method on camera pose estimation and depth estimation. In addition, we provide an analysis of our parametrisation, demonstrating the effects of different design decisions. In our experiments, we use RoMav2 Edstedt et al. (2026) for image matching. We optimise for 100 iterations with Adam Kingma & Ba (2015), with a cosine annealed learning rate schedule. Unless stated otherwise we run our experiments as well as our baseline with the dataset GT provided intrinsics. We use a Cauchy loss for our robust estimator ρ. In the appendix, we provide additional results of our method in a calibration free setting. Unless mentioned otherwise, we use $k = 14$ for our experiments.

Baselines We evaluate Latent-BA with different feed-forward reconstruction models including π^3, VGGT-Ω, Déjà View and Depth Anything 3. Our core evaluation steams around how much refinement can Latent-BA achieve over the initial reconstructions. We also compare against an explicit

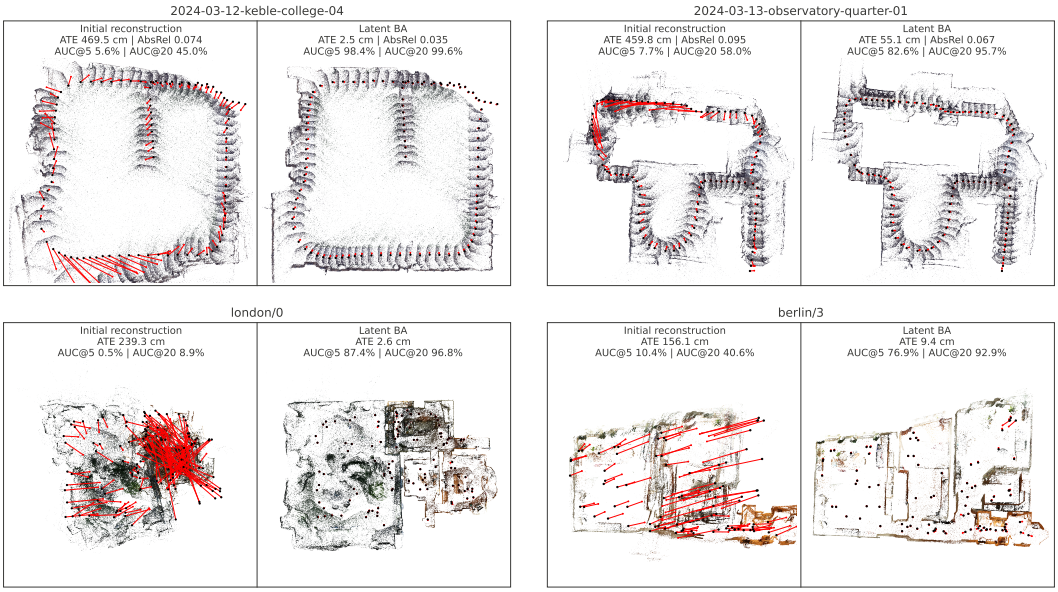


Figure 5: Top-down views of Oxford Spires (top row) and SMERF (bottom row). Red lines indicate the distance between the estimated and GT camera poses.

BA baseline that we implement with GTSAM Dellaert & Contributors (2022). For fair comparison, we give it the same initial reconstruction and match set that we use in Latent-BA. We also compare against EPO, BLASt3R and MP-SfM, and create an EPO* variant that we feed our matches to setup a reprojection loss, while we run MP-SfM with SuperPoint and RoMav2.

Datasets We run our evaluations on the full ETH3D MVS Schops et al. (2017) dataset, as well as Oxford Spires Tao et al. (2025) on 18 scenes with COLMAP GT on our test-split which we created to have long sequences without significant overlap, capping the sequence lengths at 110. In addition, we run on the synthetic NRGBD dataset Azinović et al. (2022) on the π^3 test split. Finally, we evaluate on the SMERF Duckworth et al. (2024) test splits from MP-SfM.

4.1 Camera Pose and Geometry Evaluation

Our camera pose estimation results are provided in Table 1 and video depth evaluation is provided in Table 2. We highlight four main comparisons from these results.

Latent-BA vs feed-forward reconstruction. We show a significant boost over the initial reconstruction output of VGGT-Ω and π^3 in both camera pose and dense depth estimation. This demonstrates the effectiveness of our proposed method in jointly refining the outputs of these models. We also show improvements across all other frontends in Figure 6. Qualitative results are provided on challenging scenes in Figure 5. Calibration free pose evaluations can be found in Figure 10.

Latent-BA vs explicit BA. Given the same initialisation but expressed in different forms, latent vs explicit, ours provide more accurate poses. These results showcase the advantage of our regularised optimisation and tight coupling of geometry and camera poses in the representation space.

Latent-BA vs other implicit BA. We outperform EPO and EPO* on pose estimation accuracy, highlighting the effectiveness of our latent representation compared to their decoupled sparse geometry and MLP refined camera poses.

Latent-BA vs SfM with learned priors. We also outperform BLASt3R on both evaluations, showing that leveraging large off-the-shelf models provide a stronger learned prior when accuracy is key requirement. However, BLASt3R is less demanding computationally compared to ours. Finally our comparison with MP-SfM demonstrates the competitiveness of our proposed approach compared to a full SFM pipeline. we outperform MP-SfM on dense geometry refinement. Our simple approach to establish frame pairs outperforms MP-SfM on ETH3D, NRGBD and SMERF. However on Ox-

Method	Init	ETH3D	NRGBD	Oxford Spires	SMERF
		AUC@1/5/20 ↑	AUC@1/5/20 ↑	AUC@1/5/20 ↑	AUC@1/5/20 ↑
Feed-forward	π^3	11.7/50.3/79.4	25.5/78.4/94.2	0.0/8.7/53.4	0.2/14.1/49.3
Feed-forward	VGGT-Ω	28.6/75.9/93.2	39.6/85.0/96.1	0.0/7.2/54.3	1.3/36.6/78.5
EPO	π^3	45.6/78.5/91.9	45.1/85.5/95.4	9.7/36.1/62.8	0.5/15.4/49.9
BLASt3R	–	65.0/89.9/97.1	66.9/91.1/96.9	9.5/30.6/54.7	4.2/29.4/51.5
EPO*	π^3	76.0/91.8/97.0	60.9/90.6/97.6	37.8/60.1/74.9	16.4/36.7/60.1
Classic BA	π^3	82.9/95.4/98.8	64.7/88.7/95.5	46.0/66.2/79.8	36.5/63.4/75.0
MP-SfM	–	84.0/93.7/96.1	62.8/90.1/97.2	**69.3/83.9**/89.8	22.6/58.8/77.5
Latent BA (ours)	VGGT-Ω	**85.2/96.7/99.2**	72.2/92.6/98.0	30.5/60.4/80.7	31.8/74.0/**92.8**
Latent BA (ours)	π^3	84.5/96.2/99.0	**73.1/93.1/98.1**	55.6/82.7/**91.3**	**46.3/76.2**/87.1

Table 1: Camera pose AUC at $1/5/20°$ (%).

Method	ETH3D		NRGBD		Oxford Spires	
	AbsRel ↓	$\delta_{1.01}$ ↑	AbsRel ↓	$\delta_{1.01}$ ↑	AbsRel ↓	$\delta_{1.01}$ ↑
BLASt3R	0.0377	0.370	0.0311	0.478	0.9316	0.035
MP-SfM	0.0326	0.372	0.0203	0.528	0.1148	0.130
π^3	0.0221	0.408	0.0135	0.673	0.0986	0.108
VGGT-Ω	0.0112	0.660	0.0096	0.778	0.1101	0.097
Latent BA + VGGT-Ω (ours)	**0.0059**	**0.899**	**0.0086**	0.815	0.0954	0.148
Latent BA + π^3 (ours)	0.0065	0.883	0.0113	**0.820**	**0.0831**	**0.233**

Table 2: Video depth estimation results. AbsRel with scene-median depth scale $\delta_{1.01}$.

ford Spires, the two methods are comparable on AUC at 5/20, but ours slightly lag behind on the tighter AUC@1 tolerance, in effect due to failure of the reconstruction models on initialisation to provide adequate set of pairs to match.

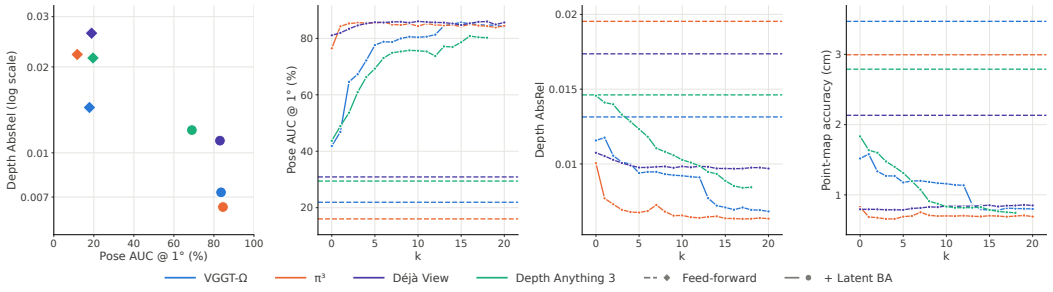


Figure 6: On the left, geometry and pose accuracy of different feed-forward methods on ETH3D for $k = 14$. We analyse the effect of using different k with different feed-forward reconstruction models, showing the impact of using deeper layers for the refinement of poses and geometry, run on subsampled set of frames of ETH3D.

4.2 Does coupling really help?

We run an ablation to evaluate the claim that coupling between geometry and poses under a joint representation is better than a decoupled representation. We modify Latent-BA, such that the geometry is parametrised by depth maps that is only initialised by the initial geometry and is independent of the $\mathbf{Z}_k$. During optimisation they only interact with with the poses, that are a function of $\mathbf{Z}_k$ via the reprojection residuals. Under this setup, $\mathbf{Z}_k$ only gets gradient updates via the camera pose decoder head only, unlike our full setup where gradient also flows back via the geometry head. As we show in Table 3, the pose estimation accuracy drops when using a decoupled parametrisation, and shows that an implicitly coupled representation matters for accuracy.

	π^3 init	Decoupled: free depth maps	Coupled: latent BA
AUC@1/5/20 ↑	16.3/62.0/87.5	76.2/94.7/98.7	**84.1/96.6/99.1**
ATE (cm) ↓	9.02	4.03	**0.97**
Depth AbsRel ↓	0.0196	0.0187	**0.0066**

Table 3: Coupling ablation on a subset of the ETH3D MVS dataset.

4.3 Different Parametrisations

We compare different parametrisations of π^3's prior: weight space (LoRA), prompt tuning with per-layer tokens, and a latent shared across frames. For LoRA, we fine-tune all linear layers, all except frame heads, or all QKV projections as in Self-Geometry Youn et al. (2026). Latent BA achieves both the best pose and depth performance in Section 4.3. The shared latent is competitive with LoRA methods, illustrating the benefit of a local parametrisation over global weight space. This local property is visible in both the lower depth changes during optimisation and lower errors near matches in Section 4.3. This aligns with the observation in Figure 4 that nearby pixels show stronger coupling. While LoRA shows improved depth at larger distances, these pixels are infrequent.

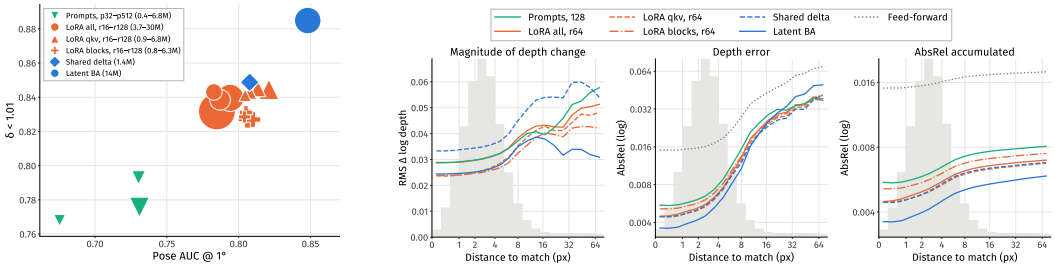


Figure 7: Comparison of parametrisations at k=3 using π^3 on 10-frame splits of ETH3D MVS using an ℓ_1 loss. The scatter plot compares methods under pose and depth thresholds. The remaining plots show depth change, depth error, and accumulated depth error for a given distance of pixels to the nearest match. The background histogram is the frequency of pixels for a given distance.

5 Limitations

Backpropagation through feed-forward reconstruction models requires significant memory which limits scalability to large scenes, as analysed in Figure 9. In the future, learning linear subspaces alongside the output of feed-forward reconstruction models, similar to depth estimation literature, may provide a balance between expressiveness and efficiency. Compared to full SfM pipelines, our straightforward frontend can be improved with more robust graph construction and chaining correspondence across views. Lastly, the use of an external matcher provides pixel-level precision, but leveraging the implicit priors from feed-forward reconstruction models to extract matches Nordström et al. (2026) may unify the whole system around a single model.

6 Conclusion and Future Work

We presented Latent Bundle Adjustment, a method for dense refinement of 3D scenes by optimising the latent representation of feed-forward reconstruction models. The change in representation provides coupling of poses and dense geometry, such that sparse reprojection error achieves both accurate poses and globally-consistent dense geometry. As this improves precision across feed-forward reconstruction models, it may be valuable for challenging scenarios with sparse data, as well as for self-supervision. Beyond performance, we studied effects across architectures, layers, and parametrisations. Architectures show qualitatively different characteristics based on connectivity and head design, and we hope this insight inspires designing future architectures of models that are more conducive to scalable and efficient test-time optimisation.

REFERENCES

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building Rome in a day. *Communications of the ACM*, 2011.

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3D reconstruction models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Alessandro Burzio, Tobias Fischer, Sven Elflein, Qunjie Zhou, Riccardo de Lutio, Jiawei Ren, Jiahui Huang, Shengyu Huang, Marc Pollefeys, Laura Leal-Taixé, Zan Gojcic, and Haithem Turki. Déjà view: Looping transformers for multi-view 3D reconstruction. *arXiv preprint arXiv:2605.30215*, 2026.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3r: 3D reconstruction as test-time training. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Frank Dellaert and GTSAM Contributors. Gtsam, 2022. URL `https://doi.org/10.5281/zenodo.5794541`.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

Eric Dexheimer and Andrew J. Davison. Learning a depth covariance function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T Barron. SMERF: Streamable memory efficient radiance fields for real-time large-scene exploration. *ACM Transactions on Graphics*, 2024.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2025.

Mattia D'Urso, Christian Sormann, Mattia Rossi, and Friedrich Fraundorfer. Boosting 3D foundation models with edge-based pose optimization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Johan Edstedt, David Nordström, Yushan Zhang annd Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. RoMa v2: Harder better faster denser feature matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Benjamin Graham and David Novotny. RidgeSfM: Structure from motion via robust pairwise matching under depth uncertainty. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. In *NVIDIA Research Whitepapers*, 2025.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Vincent Leroy, Philippe Weinzaepfel, Lojze Zust, Yohann Cabon, and Jerome Revaud. BLASt3R: Bundle adjustment of any image set with multi-view matching and monocular priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. LoRA3D: Low-rank self-calibration of 3D geometric foundation models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

David Nordström, Xinyue Zhang, Thibaut Loiseau, Vincent Lepetit, and Fredrik Kahl. RoMa-Ω: What feed-forward 3D models know about image matching. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2026.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Linfei Pan, Johannes Lutz Schönberger, and Marc Pollefeys. Global structure-from-motion meets feedforward reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zador Pataki, Paul-Edouard Sarlin, Johannes L. Schönberger, and Marc Pollefeys. MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zador Pataki, Paul-Edouard Sarlin, and Marc Pollefeys. VidMap: Exploiting temporal structure for video-based structure-from-motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Frank Plastria. The Weiszfeld algorithm: proof, amendments, and extensions. In *Foundations of location analysis*. Springer, 2011.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. In *Proceedings of SIGGRAPH*, 2006.

Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The Oxford spires dataset: Benchmarking large-scale LiDAR-visual localisation, reconstruction and radiance field methods. *International Journal of Robotics Research (IJRR)*, 2025.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Neural Information Processing Systems (NeurIPS)*, 2021.

B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment — a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms, in association with ICCV*, 1999.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1991.

Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster VGGT with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026b.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Neural Information Processing Systems (NeurIPS)*, 2024.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Seokhyun Youn, Dahyeon Kye, Sung-Ho Bae, and Jihyong Oh. Self-Geometry: GT-free and plug-and-play test-time adaptation for geometrically consistent 3D vision foundation models. *arXiv preprint arXiv:2608.10708*, 2026.

Yuheng Yuan, Qiuhong Shen, Shizun Wang, Xingyi Yang, and Xinchao Wang. Test3R: Learning to reconstruct 3D at test time. In *Neural Information Processing Systems (NeurIPS)*, 2025.

Lamberto Ballan Ziliotto Filippo, Luciano Serafini and Tommaso Campari. What VGGT knows about overlap: Probing geometric foundation models for co-visibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

AI USE STATEMENT

In this work, we used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation. Additionally, we used generative AI tools for creating or modifying scientific figures or images, suggesting experimental parameters, creating or editing software code, summarizing or analysing existing literature, and brainstorming. We have reviewed all AI-assisted work, including verification of code for correctness by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A Appendix

A.1 Convergence Basin Example With Depth Supervision

In addition to the basin visualisation in Figure 2, we show an example beyond reprojection error in Figure 8. Given a ground-truth map, we can optimise that frame's depth alone via Latent BA and observe the response. This situation can find many unnatural configurations, as only the one depth map is constrained, while the other depths and all poses are not. Once again, while there exist other solutions with lower depth loss that achieve degraded depth and or poses, our method finds solutions that generalise. This demonstrates the implicit regularisation when using the feed-forward pass as initialisation combined with gradient descent at test-time. The difference in solutions based on the latent representation used is also visible, with the representation before the projection to pixel space only showing local changes at the supervised depth map.

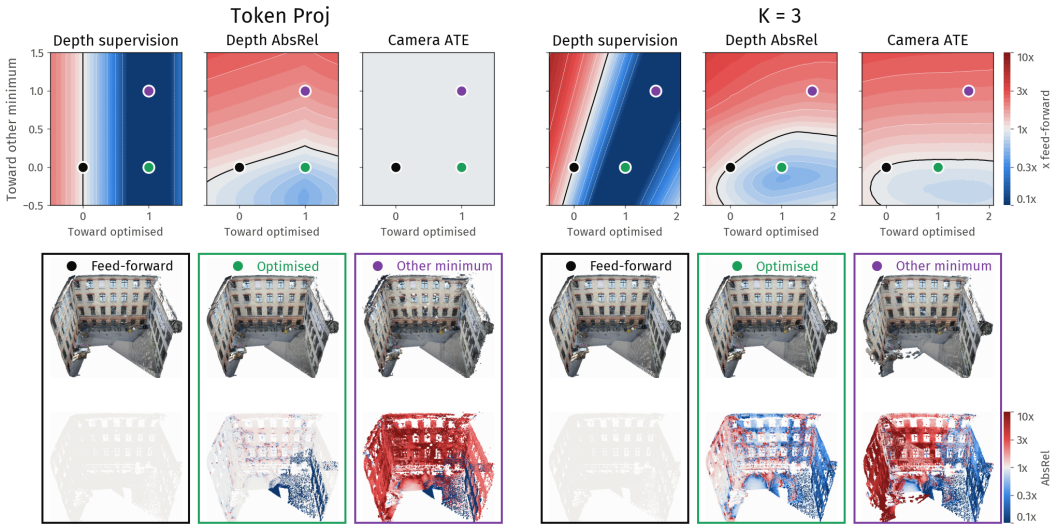


Figure 8: Convergence basin when optimizing latent representations of π^3 before the projection to pixel space and before the final three inter-frame attention blocks against depth supervision to a single frame. Colormaps show the improvement multiplier across the error and evaluation metrics. The axes show two directions of high-dimensional latent space.

A.2 Computation Analysis

We show an analysis of iteration time and peak GPU memory across k of π^3 in Figure 9. Iteration time shows a quadratic due to the nature of global self-attention, while memory is linear as the latent representations scale with the number of patches. Memory can further be reduced at the expense of iteration time via gradient checkpointing.

A.3 Calibration Free Results

As feed-forward reconstruction models also predict camera intrinsics per-frame, either implicitly or explicitly, we analyse the effect of using estimated intrinsics. For a given sequence, we assume camera intrinsics are shared across frames by averaging the predicted estimates output either from the camera head directly or after recovery from pointmaps. Results for ETH3D are shown in Figure 10. We show even this one-step estimation of intrinsics followed by optimisation still significantly improves over the feed-forward estimate, while using ground-truth intrinsics adds additional accuracy, particularly at the most precise thresholds.

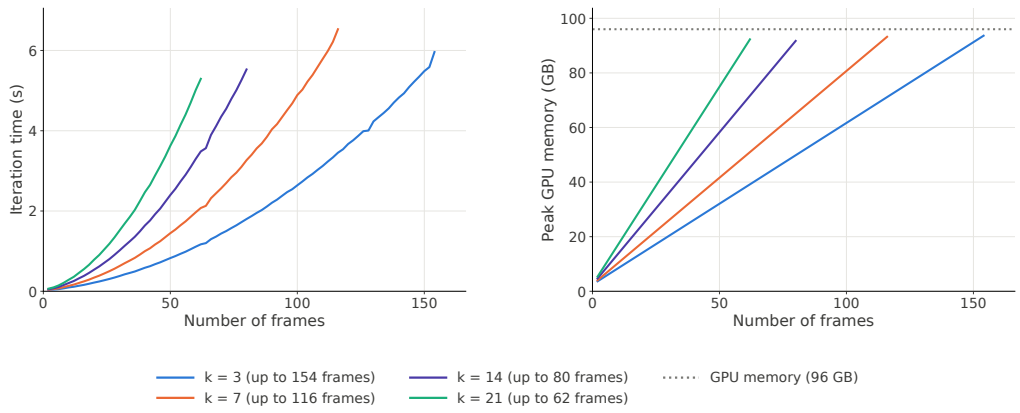


Figure 9: Compute and memory across number of frames on Nvidia GH200 GPU with 96 GB.

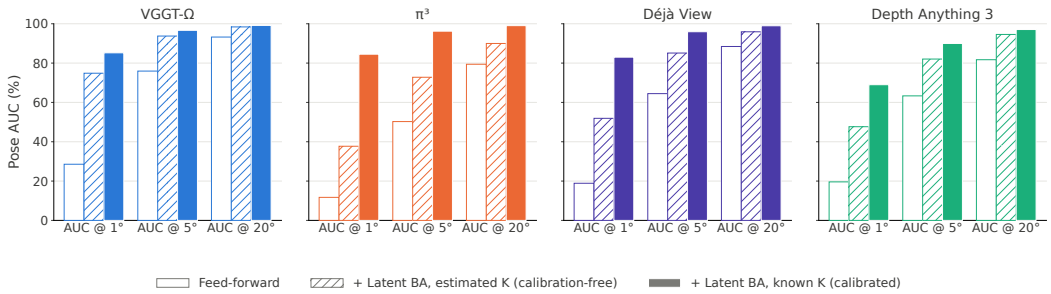
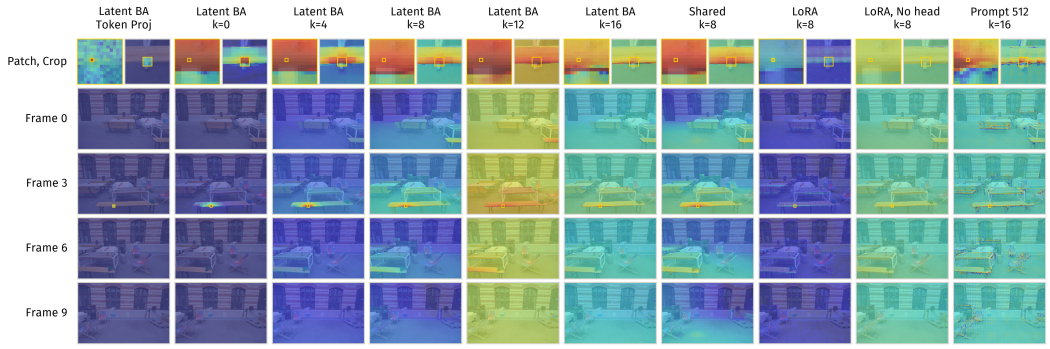


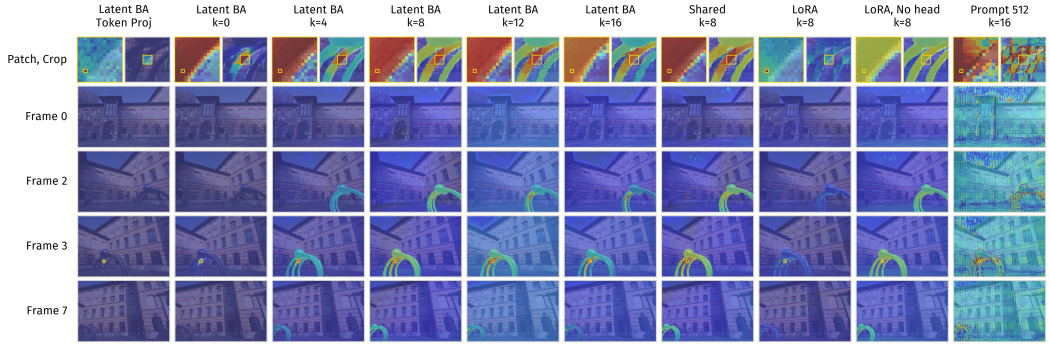
Figure 10: Pose estimation results on ETH3D MVS, comparing our method in a calibrated and a uncalibrated settings.

A.4 Additional Kernel Visualisations

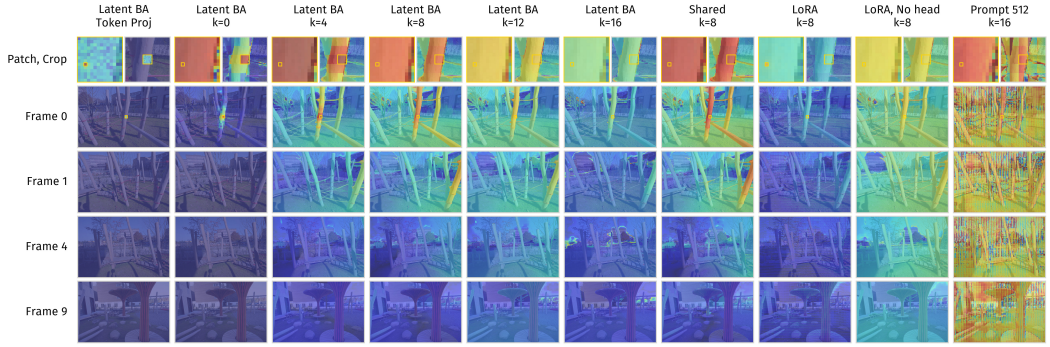
We show additional visualisations of the log-depth kernel, inspired by Jacot et al. (2018), presented in Figure 4 for π^3. As noted previously, latent representations $\mathbf{Z}_k$ further from output space demonstrate greater global receptive field. Here, we also contrast with other representations, such as LoRA which parametrises variables on weight space. In Figure 11, we visualise the response maps but after optimisation instead of at the feed-forward initialisation. The colour normalisation is performed from the initial response prior at a selected pixel prior to optimisation. A similar trend of receptive field is seen for Latent BA, while LoRA shows the inability to retain sharp responses on edges due to the global parametrisation of weight space. Patch artifacts can be removed by omitting adaption of the heads, but the degradation of local response in Figure 11a is visible.



(a) Courtyard.



(b) Facade.



(c) Playground.

Figure 11: Visualisation of linearised coupling between a change in the selected pixel's log-depth and all log-depths for different parametrisations on the ETH3D dataset. Linearisation is performed after 100 iterations of gradient descent using Adam.