This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Flow3r: Factored Flow Prediction for Scalable Visual Geometry Learning

Zhongxiao Cong Qitao Zhao Minsik Jeon Shubham Tulsiani
Carnegie Mellon University

project & code: https://flow3r-project.github.io/



Figure 1. `Flow3r` leverages *unlabeled* videos (using flow supervision) alongside labeled 3D data for scalable visual geometry learning. This enables accurate multi-view 3D reconstruction in-the-wild, in particular for settings with scarce labeled data *e.g.*, interaction videos and dynamic scenes.

Abstract

Current feed-forward 3D/4D reconstruction systems rely on dense geometry and pose supervision – expensive to obtain at scale and particularly scarce for dynamic real-world scenes. We present `Flow3r`, a framework that augments visual geometry learning with dense 2D correspondences ('flow') as supervision, enabling scalable training from unlabeled monocular videos. Our key insight is that the flow prediction module should be factored*: predicting flow between two images using geometry latents from one and pose latents from the other. This factorization directly guides the learning of both scene geometry and camera motion, and naturally extends to dynamic scenes. In controlled experiments, we show that factored flow prediction outperforms alternative designs and that performance scales consistently with unlabeled data. Integrating factored flow into existing visual geometry architectures and training with* $\sim$*800K unlabeled videos, `Flow3r` achieves state-of-the-art results across eight benchmarks spanning static and dynamic scenes, with its largest gains on in-the-wild dynamic videos where labeled data is most scarce.*

1. Introduction

The task of 'visual geometry inference', *i.e.*, recovering the 3D structure of a (static or dynamic) scene from multi-view input images has undergone a paradigm shift – evolving from classical optimization-based methods [1, 15, 26, 34, 37, 43] to recent data-driven predictors [38, 40, 41, 48] that can directly output the geometry and pose corresponding to the input images. The success of such systems, however, has crucially relied on multi-view training data with dense geometry and camera pose supervision. Unfortunately, such supervision is not easily available across all settings of interest, *e.g.*, for dynamic scenes in the wild or domains like egocentric videos, and existing methods do not generalize well to such scenarios. More broadly, unlike the self-supervised objectives that have enabled scaling of LLMs and vision transformers, the reliance on dense geometry and pose labels prevents truly large-scale visual geometry learning.

We take a step towards scalable learning of multi-view models and present `Flow3r`, a framework to guide visual geometry learning from *unlabeled* videos, *i.e.*, without any explicit geometry or pose supervision. Instead, `Flow3r` leverages a readily available supervisory signal that is a cornerstone of classical (and recent) optimization-based multi-view methods – dense *correspondences across images*. Inspired by recent progress in inferring dense correspondences for generic image pairs and videos, we seek to

unlock scalable learning by incorporating such 2D flow as auxiliary supervision for 3D visual geometry models. The key technical question we address is: ***'how can flow effectively supervise visual geometry prediction?'***.

We are not the first to consider flow supervision for visual geometry learning. Indeed, VGGT [38] adds a 'tracking' module that uses local features from two images to predict flow between them as an auxiliary training objective. However, as we show later, this merely encourages visually discriminative features but does not directly aid the learning of pose or geometry. Our key insight is that, to guide geometry learning, the *flow prediction module should be asymmetric*. We observe that for static scenes, the flow between a source and a target image can be induced only via the geometry of the source image (pointmaps in a global coordinate) and the camera pose of the target. Motivated by this, we propose to incorporate a ***factored flow prediction module*** in visual geometry models. Specifically, such models typically compute 'local' patchwise features that later predict geometry as well as a global per-image token that infers camera pose. Our flow prediction module computes flow between two images using *only the global pose token for the target along with the patchwise tokens for the source*.

We find that factored flow prediction provides stronger supervision for pose and geometry than the 'tracking' design adopted by previous works. Moreover, unlike projection-based flow inference, it enables robust prediction and naturally extends to dynamic scenes. `Flow3r` integrates factored flow prediction and leverages ∼800K unlabeled videos as supervision alongside existing labeled 3D datasets. This allows `Flow3r` to achieve state-of-the-art results across diverse benchmarks spanning static and dynamic scenes, with the largest improvements on in-the-wild dynamic videos where labeled data is scarce. More broadly, by extracting the supervisory signal from unlabeled videos, `Flow3r` represents a step towards large-scale visual geometry learning without large-scale labeled data.

2. Related Work

Dense Correspondence Learning. Dense correspondence methods estimate pixel-level matches across images, providing a foundation for downstream camera motion and 3D structure recovery. These have progressed from classical optical flow [18, 27] to recent wide-baseline matchers such as DKM [12], RoMa [13], and UFM [46] that achieve robust matching across large viewpoint changes. Video-based trackers like TAPIR and CoTracker [10, 11, 17, 20] further extend correspondence estimation to long-range temporal reasoning across multiple frames. `Flow3r` builds on these models, using their predicted correspondences as training supervision to guide 3D geometry learning on data without 3D labels.

Correspondence-driven Reconstruction. Building on dense correspondence estimation, reconstruction methods recover 3D structure and camera motion from inferred matches. Classical Structure-from-Motion [34] pipelines estimate camera poses and scene structure by detecting local features, computing pairwise correspondences, and jointly optimizing camera parameters and 3D points through bundle adjustment. Visual SLAM systems [14, 28] extend this to dynamic settings, tracking features across frames to jointly estimate camera trajectories and scene geometry. Recent approaches, including Robust-CVD [21], CasualSAM [47], VIPE [19] and MegaSAM [24], further incorporate monocular depth priors to enhance robustness under motion and occlusion. However, these methods remain optimization-based, requiring per-video refinement and lacking feed-forward efficiency.

Feed-forward Visual Geometry Learning. Recent efforts aim to replace traditional optimization pipelines with feed-forward networks that directly predict visual geometry from images. DUSt3R [41] first demonstrated that dense pointmaps can be estimated from image pairs within a shared coordinate system, enabling efficient two-view reconstruction. MASt3R [22] further improves this paradigm by introducing a learned matching head for better correspondence reasoning, while DiffusionSfM [48] and VGGT [38] generalize to multi-view settings, jointly estimating camera parameters and scene structure. Subsequent works such as MonST3R [45], CUT3R [40], and StreamVGGT [51] extend this formulation to dynamic scenes, learning temporally consistent geometry across video frames. However, these models rely on 3D and camera supervision, which is not easily scalable. In contrast, our method enables scalable feed-forward learning of dynamic visual geometry using flow as auxiliary supervision, allowing training on unlabeled real-world videos.

3. Approach

Our approach, `Flow3r`, scales visual geometry learning by supervising factored flow prediction, enabling the use of in-the-wild data without ground-truth geometry annotations (*e.g.*, camera poses and depth maps). We first review the standard paradigm for fully supervised visual geometry learning (Sec. 3.1), then introduce our factored flow prediction formulation (Sec. 3.2), which decodes flow by combining the camera token from one view with the patch tokens from another – allowing flow supervision to directly guide geometry learning while naturally extending to dynamic scenes. Finally, we describe the overall supervision signals and architecture (Sec. 3.3).

3.1. Preliminaries: Visual Geometry Networks

State-of-the-art visual geometry networks (*e.g.*, VGGT [38] and π^3 [42]) take as input a set of images $\{I_1, I_2, \ldots, I_N\}$

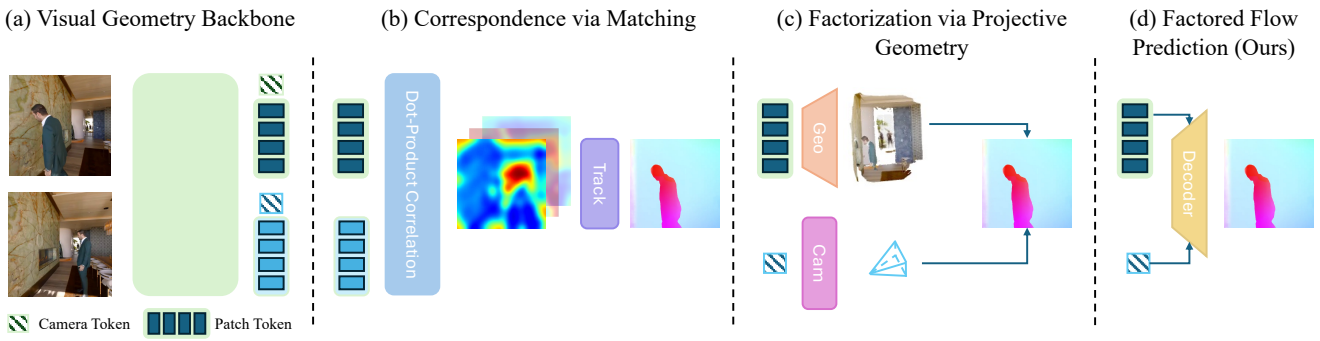


Figure 2. **Mechanisms for flow prediction.** (a) The visual geometry backbone first extracts the camera token and patch tokens of each input. (b) Existing correspondence heads [38] predict flow directly from local features via matching. (c) Flow can also be obtained by explicitly projecting predicted 3D points into another view via decoded camera parameters. However, this projection-based formulation is restricted to static scenes and sensitive to geometric errors. (d) Our factored flow approach conditions source-view geometry latents on the target-view camera latent and decodes dense correspondences directly in latent space, providing a geometry-aware and robust flow prediction mechanism that naturally extends to dynamic scenes.

and infer 3D scene geometry through a unified multi-view transformer. For instance, VGGT first encodes the input images into latent patch tokens using an off-the-shelf vision backbone [29]:

$$\mathbf{X}_i = f_{\text{enc}}(I_i), \quad \mathbf{X}_i \in \mathbb{R}^{P \times D}, \tag{1}$$

where P and D denote the number of patch tokens and their feature dimension, respectively.

A multi-view transformer then performs cross-view reasoning over these features to jointly infer camera-related features and geometry-related features:

$$\{\mathbf{c}_i, \mathbf{g}_i\}_{i=1}^{N} = f_{\text{multi-view}}\big(\{\mathbf{X}_i\}_{i=1}^{N}\big), \tag{2}$$

where $\mathbf{c}_i$ and $\mathbf{g}_i$ denote the per-view camera features and geometry features, respectively [1].

Finally, the updated latent representations are decoded into explicit geometric properties – camera poses, depths, and global pointmaps.

$$[\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i, \hat{\mathbf{K}}_i] = f_{\text{cam}}(\mathbf{c}_i), \tag{3}$$

$$\hat{\mathbf{D}}_i, \hat{\mathbf{P}}_i = f_{\text{depth}}(\mathbf{g}_i), f_{\text{point}}(\mathbf{g}_i). \tag{4}$$

These outputs are supervised using ground-truth labels, often obtained from multi-view reconstruction pipelines such as Structure-from-Motion [1, 34]. However, such supervision is expensive and limited in coverage, restricting the diversity of scenes and motions seen during training – motivating our exploration of scalable alternatives.

[1] In VGGT, $\mathbf{c}_i$ corresponds to a learnable camera token appended to patch tokens, whereas in π^3 it represents latent camera features predicted from the multi-view transformer rather than a single token.

3.2. Learning Visual Geometry via Factored Flow

`Flow3r` leverages *flow* (dense pixel correspondences) as an additional supervision signal for visual geometry learning, enabling the use of in-the-wild data without ground-truth camera or depth annotations. The key technical question we answer is about *how* one can effectively leverage such supervision, and we first detail different possible alternatives before describing our proposed mechanism:

Flow as Visual Correspondence. One possible approach is to infer flow using architecture relying on local feature 'matching' across image pairs (Fig. 2 (b)). Indeed, prior multi-view models [38] adopt this design and leverage flow as an auxiliary training signal. However, this design primarily ensures visually discriminative local features but, as we empirically show, does not facilitate the learning of scene geometry and camera motion.

Flow from Explicit Camera and Scene Geometry. Classical projective geometry provides an alternative mechanism to infer flow between images. Considering a static scene with predicted global pointmap $\hat{\mathbf{P}}_i$ from view i and the camera parameters $(\hat{\mathbf{R}}_j, \hat{\mathbf{t}}_j, \hat{\mathbf{K}}_j)$ for view j, the flow between these images can be analytically computed by projecting $\hat{\mathbf{P}}_i(\mathbf{u}_i)$ into camera j:

$$\hat{\mathbf{F}}_{i \to j}(\mathbf{u}_i) = \hat{\mathbf{u}}_{i \to j} = \pi\big(\hat{\mathbf{K}}_j(\hat{\mathbf{R}}_j\, \hat{\mathbf{P}}_i(\mathbf{u}_i) + \hat{\mathbf{t}}_j)\big), \tag{5}$$

where the operator $\pi(\cdot)$ denotes perspective projection from 3D points to pixel coordinates.

This explicit flow computation mechanism, grounded in projective geometry, provides an ideal factorization that jointly promotes learning of both scene geometry and camera motion, as illustrated in Fig. 2 (c). However, the pure projection in Eq. 5 cannot account for scene motion. While this can be mitigated by addition scene flow prediction as in ZeroMSF [25], we find that supervising flow from the

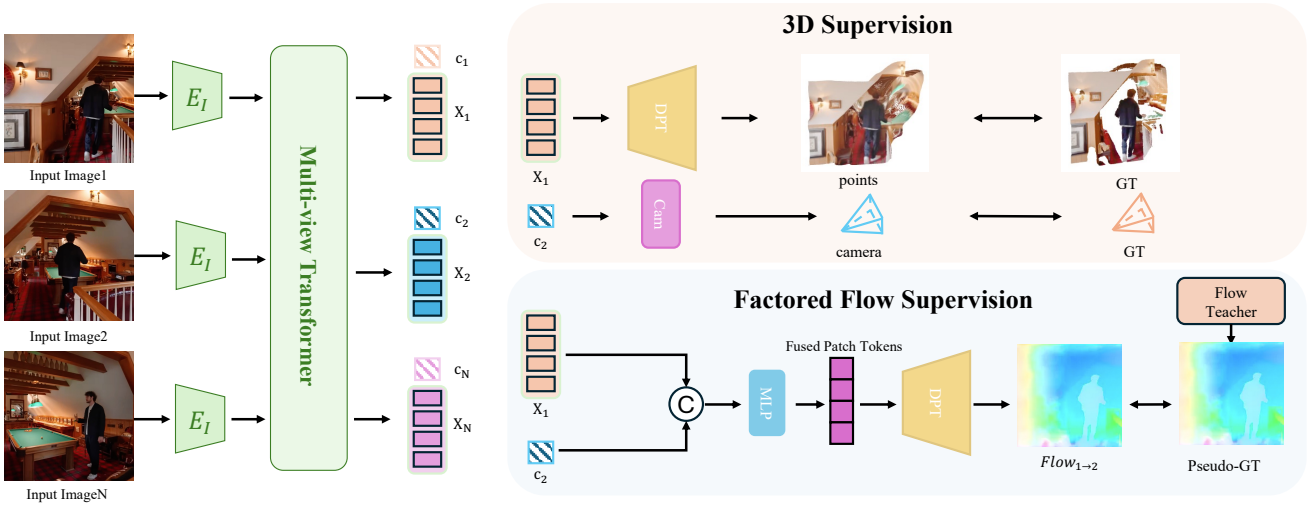


Figure 3. **Overview of `Flow3r`.** `Flow3r` predicts visual geometry using factored flow supervision, enabling scalable geometry learning from unlabeled videos. Each input image is encoded and processed by the multi-view transformer to produce camera tokens and patch tokens. For data with dense geometry and pose labels, we directly supervise the patch tokens and camera tokens with the corresponding labels. For unlabeled datasets without geometry and pose supervision, we predict flow between two frames in a factorized manner, supervised by the pseudo labels from an off-the-shelf 2D flow prediction model [46]. To obtain the factored flow, we fuse the patch features of one frame with the camera features of another, and decode the fused representation through the DPT head to produce dense flow predictions.

decoded cameras and depths can introduce instability even in static scenes, as errors in either component may lead to large inaccuracies in the resulting flow.

Factored Flow Prediction. Our approach builds on an insight from the projection-based flow computation – that asymmetric flow computation using geometry from the source and pose from the target can better guide visual geometry learning. However, rather than relying on analytical computation, we propose to leverage a learned *factored flow prediction* mechanism to infer flow from pose and geometry latent representations (see (d) in Fig. 2).

More specifically, given the geometry latents $\mathbf{g}_i$ from view i and camera latents $\mathbf{c}_j$ from view j (the outputs of the multi-view transformer in Eq. 2), `Flow3r` leverages a flow prediction head to predict the flow:

$$\hat{\mathbf{F}}_{i\to j} = \Phi_{\text{flow}}(\mathbf{g}_i, \mathbf{c}_j), \tag{6}$$

where Φ_{flow} denotes a learned flow prediction module. The camera latents from view j are used to modulate the geometry latents from view i, which are then decoded by a DPT head [31] to produce the flow field.

This approach bypasses the need to decode explicit geometric elements for flow computation, improving robustness and enabling end-to-end training. Moreover, it implicitly handles dynamic scenes where the flow field reflects a combination of camera motion and scene motion rather than pure geometric projection.

3.3. Overall Architecture and Learning Objectives

`Flow3r` builds upon a standard supervised visual geometry learning framework (Section 3.1) – a multi-view transformer backbone with camera and geometry heads supervised using labeled 3D data. In addition, `Flow3r` learns a factored flow prediction module, allowing 2D flow as supervision on both labeled and (crucially) unlabeled data, enabling scalable visual geometry learning across both static and dynamic scenes – see Fig. 3 for an overview.

Camera and Geometry Supervision from Labeled Data. `Flow3r` predicts per-image cameras and local geometry along with confidence maps that indicate model uncertainty. When 3D labels are available, we supervise the decoded predictions using ground-truth camera poses $\mathbf{T}_i$ and geometry annotations $\mathbf{G}_i$:

$$\mathcal{L}_{sup} = \sum_{i=1}^{N} \Big(\lambda_{cam}\, \mathcal{L}_{cam}(\hat{\mathbf{T}}_i, \mathbf{T}_i) + \lambda_{geo}\, \mathcal{L}_{geo}(\hat{\mathbf{G}}_i, \mathbf{G}_i) \Big). \tag{7}$$

To disentangle visual geometry learning from predefined coordinate frames, `Flow3r` adopts the permutation-equivariant geometry prediction framework introduced by π^3 [42], which predicts geometry in local coordinates without relying on a fixed global frame. Models that predict geometry in a global frame (*e.g.*, VGGT [38]) can be modified to follow the same design by removing the 'first camera' token and using an optimal-alignment-based loss for supervising the predicted pointmaps $\hat{\mathbf{P}}$

Model Variant	Static Scenes					Dynamic Scenes				
	Data (# Seqs)	RRA@30↑	RTA@30↑	CD↓	MSE↓	Data (# Seqs)	RRA@30↑	RTA@30↑	CD↓	MSE↓
`3d-sup`	ScanNet++ (1K)	0.7500	0.6929	0.030	0.088	OmniWorld (1K)	66.01	62.37	0.105	0.637
`flow-projective`	+ ScanNet++ (1K)	0.6700	0.4572	0.033	0.088	+ SpatialVID (3K)	61.23	56.12	0.158	0.710
`flow-tracking`	+ ScanNet++ (1K)	0.7438	0.7021	0.030	0.089	+ SpatialVID (3K)	68.56	62.95	0.107	0.628
`flow-factored`	+ ScanNet++ (1K)	**0.7700**	**0.7366**	**0.026**	**0.078**	+ SpatialVID (3K)	**76.26**	**68.84**	**0.103**	**0.598**

Table 1. **Does factored flow prediction help visual geometry learning?** We evaluate the camera and geometry prediction of a model trained with only 3D labeled data and variants that use additional unlabeled data via flow supervision. We find that on both static and dynamic scenes, our proposed factored flow prediction (`flow-factored`) yields consistent gains over the model variant without additional unlabeled data (`3d-sup`) as well as other flow-supervised alternatives.

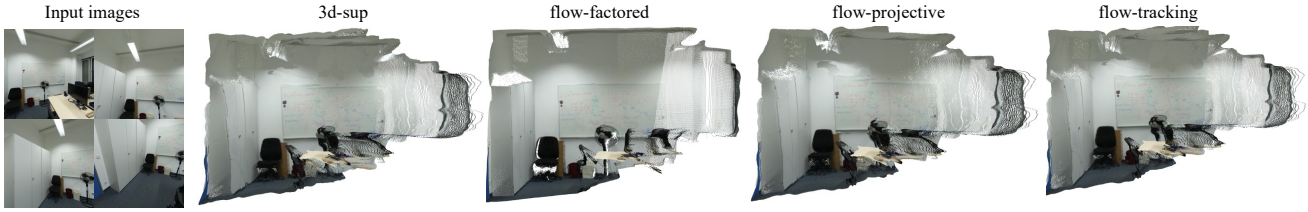


Figure 4. **Factored flow prediction aids visual geometry learning.** Compared with the baseline (`3d-sup`) and alternative formulations that use flow supervision (`flow-projective`, `flow-tracking`), `Flow3r` (`flow-factored`) yields more accurate scene reconstruction, highlighting the benefits of unlabeled data and our factored flow formulation.



Figure 5. **Scaling with unlabeled videos.** With factored flow supervision, increasing the amount of unlabeled SpatialVID data progressively improves dynamic-scene reconstruction quality, even surpassing the model with additional 3D-labeled training data.

(see appendix for details):

$$\mathcal{L}_{\text{point}} = \frac{1}{K} \min_{\mathbf{R}\in SO(3),\, \mathbf{t}\in\mathbb{R}^3} \sum_k \|\mathbf{R}\hat{\mathbf{p}}_k + \mathbf{t} - \mathbf{p}_k\|_2 \,, \quad (8)$$

where $\hat{\mathbf{p}}_k \in \hat{\mathbf{P}}$ and $\mathbf{p}_k \in \mathbf{P}$ denote corresponding 3D points in a sequence, and K is the number of valid points.

Flow Supervision from Labeled and Unlabeled Data. We supervise the proposed flow head with a robust regression loss [13] between the predicted coordinates $\hat{\mathbf{u}}_{i\to j}$ and the ground-truth correspondences $\mathbf{u}_{i\to j}$:

$$\mathcal{L}_{\text{flow}} = \frac{1}{\sum_{p\in\Omega} \mathbf{C}[p]} \sum_{p\in\Omega} \mathbf{C}[p]\, \ell_{\text{robust}}(\|\hat{\mathbf{u}}_{i\to j}[p] - \mathbf{u}_{i\to j}[p]\|_2) \,,$$

where $p \in \Omega$ indexes pixels, $\mathbf{C}$ denotes the covisibility mask between view i and j, and ℓ_{robust} is a generalized Charbonnier loss emphasizing inliers. We use this loss on both labeled and unlabeled data, relying on pseudo-ground-truth 2D correspondences from a pre-trained teacher (UFM [46]) when ground-truth 2D flow is unavailable.

4. Experiments

We evaluate `Flow3r` in two stages. First, in controlled experiments, we show that factored flow prediction outperforms alternative flow-based designs and that performance scales consistently with the amount of unlabeled data (Sec. 4.1). We then integrate factored flow prediction into state-of-the-art visual geometry networks and demonstrate improvements across eight benchmarks by leveraging ∼800K unlabeled videos (Sec. 4.2).

4.1. Flow Supervision Improves Visual Geometry

We first evaluate whether factored flow prediction improves visual geometry learning in a controlled setting. We compare against training with only labeled 3D data, and against alternative flow prediction designs. We also analyze the scaling behavior by increasing the unlabeled data.

Experimental Designs. We denote the 'base model' trained with only 3D supervision as `3d-sup`. Building upon this 'no-flow' baseline, we study three variants that incorporate *additional* unlabeled data via flow supervision using different for-

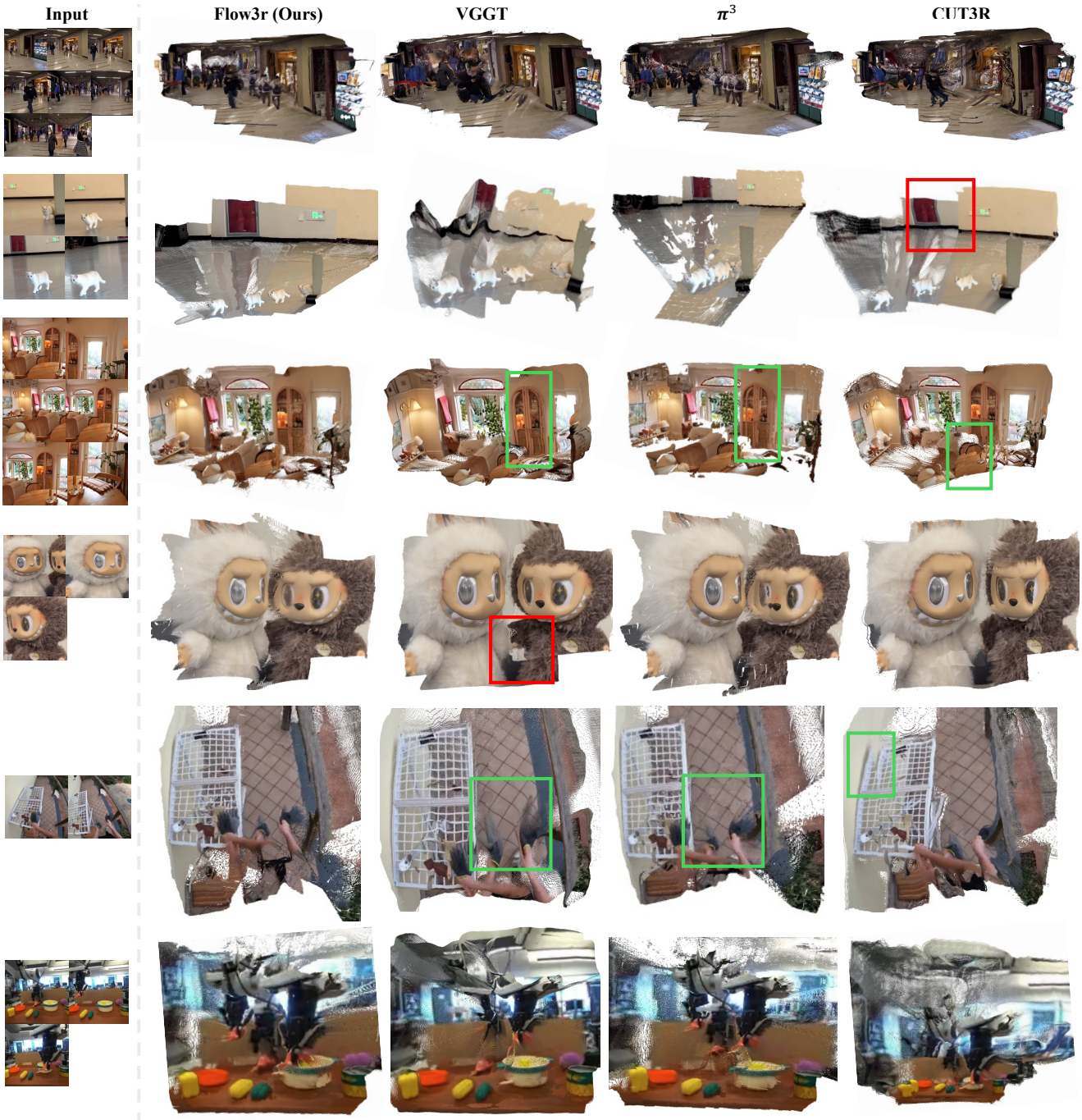


Figure 6. **Qualitative results on in-the-wild data.** We compare the reconstructions from `Flow3r` with representative feedforward reconstruction methods on in-the-wild multi-view images from static scenes, dynamic scenes, and interaction videos. We observe that `Flow3r` infers cleaner and more accurate scene structure than baselines that often yield misaligned outputs or infer inaccurate motion.

mulations: (1) `flow-projective`, computes flow explicitly from predicted camera poses and pointmaps via projective geometry; (2) `flow-tracking`, adopts a VGGT-style [38] tracking head based on pairwise patch features; (3) `flow-factored`, applies our proposed factored flow prediction formulation. All models share a (small) VGGT-like architecture, and are trained *from scratch* (except for the image encoder initialized from DINO [29]) – see the appendix for details.

Static Scenes. We train and evaluate on ScanNet++ [44] using $1K$ scenes as sources of labeled 3D data and another $1K$ scenes as 'unlabeled'. The model variants with flow supervision apply flow loss only to the additional unlabeled training sequences, ensuring

Methods	Kinetics700				EPIC-KITCHENS				Sintel				Bonn			
	RPE trans↓	RPE rot↓	MSE↓	f-score@5↑	RPE trans↓	RPE rot↓	MSE↓	f-score@5↑	RPE trans↓	RPE rot↓	MSE↓	f-score@5↑	RPE trans↓	RPE rot↓	MSE↓	f-score@5↑
DUSt3R	0.063	9.343	0.366	0.533	0.110	8.492	0.312	0.528	0.179	15.166	0.622	0.271	0.113	6.384	0.116	0.800
CUT3R	0.027	1.988	0.303	0.573	0.081	4.709	0.338	0.493	0.128	1.998	0.676	0.217	0.095	6.349	0.088	0.899
VGGT	0.038	1.392	0.347	0.479	0.049	3.025	0.220	0.617	0.086	1.220	0.595	0.311	0.096	**6.021**	0.082	0.884
π^3	0.023	1.006	0.267	0.585	0.043	3.025	0.200	0.620	0.066	1.122	0.523	0.317	0.095	6.240	0.076	0.905
Flow3r	**0.018**	**0.830**	**0.256**	**0.599**	**0.037**	**2.729**	**0.199**	**0.622**	**0.058**	**0.920**	**0.426**	**0.404**	**0.094**	6.262	**0.052**	**0.954**

Table 2. **Comparison on dynamic datasets.** We evaluate the camera (RPE rot and trans) and geometry prediction (MSE and f-score) for dynamic scene reconstruction across different datasets. Best, second and third results are highlighted in light red, orange, and yellow, respectively. `Flow3r` consistently outperforms state-of-the-art methods in both camera pose estimation and scene reconstruction, demonstrating the effectiveness of leveraging large-scale unlabeled videos for visual geometry learning via factored-flow supervision.

Methods	Co3Dv2				Scannet				NRGBD				7-scenes			
	RTA@30↑	AUC@30↑	MSE↓	f-score@5↑	RTA@30↑	AUC@30↑	MSE↓	f-score@5↑	RTA@30↑	AUC@30↑	MSE↓	f-score@5↑	RTA@30↑	AUC@30↑	MSE↓	f-score@5↑
DUSt3R	90.68	68.11	0.223	0.783	57.14	31.56	0.092	0.831	93.25	76.04	0.063	0.865	76.59	55.07	0.170	0.714
CUT3R	90.68	68.11	0.278	0.737	71.39	42.30	0.132	0.740	95.63	76.90	0.075	0.808	81.15	59.72	0.183	0.695
VGGT	97.27	87.62	0.151	0.874	**93.71**	**71.37**	**0.053**	0.931	99.21	93.13	0.032	0.964	86.51	69.77	0.182	0.665
π^3	97.49	**90.53**	0.151	0.874	91.14	69.39	0.058	0.930	99.20	90.40	0.021	0.983	87.69	71.80	0.169	0.737
Flow3r	**97.62**	90.41	**0.150**	**0.876**	92.89	71.00	**0.053**	**0.943**	**99.60**	**94.40**	**0.018**	**0.992**	**91.66**	**75.76**	**0.102**	**0.807**

Table 3. **Comparison on static datasets.** We evaluate the camera (RTA and AUC) and geometry prediction (MSE and f-score) for static scene reconstruction from the multi-view input. Best, second and third results are highlighted in light red, orange, and yellow, respectively. `Flow3r` outperforms state-of-the-art methods, indicating that the gains from incorporating unlabeled dynamic videos for visual geometry learning also transfer to static scenes.

Data		Metrics			
Labeled (# Seqs)	Unlabeled (# Seqs)	RRA@30↑	RTA@30↑	CD↓	MSE↓
OmniWorld (1K)	-	66.01	62.37	0.105	0.637
OmniWorld (1K)	SpatialVID (3K)	76.26	68.84	0.103	0.598
OmniWorld (1K)	SpatialVID (10K)	78.45	68.82	0.077	0.560
OmniWorld (1K)	SpatialVID (20K)	**81.12**	**71.21**	**0.075**	**0.532**
OmniWorld (4K)	-	78.68	70.26	0.080	0.565

Table 4. **Scaling with unlabeled videos for dynamic visual geometry learning.** We train model variants on OmniWorld [50] (labeled 3D supervision) together with increasing amounts of SpatialVID [39] (unlabeled videos with flow supervision). Performance improves consistently as we scale up the unlabeled data.

Base	3D Sup.	Flow Loss		EPIC-KITCHENS				7-scenes			
		Labeled	Unlabeled	trans↓	rot↓	MSE↓	f-score↑	RTA↑	AUC↑	MSE↓	f-score↑
VGGT	N/A (base model)			0.049	3.025	0.220	0.617	86.51	69.77	0.182	0.665
	✓	✗	✗	0.050	3.005	0.218	0.530	88.10	68.23	0.167	**0.750**
	✓	✓	✗	0.045	3.005	0.203	0.530	89.17	68.22	0.167	0.749
	✓	✓	✓	**0.039**	**2.808**	**0.202**	**0.599**	**90.67**	**71.75**	**0.166**	**0.750**
π^3	N/A (base model)			0.043	3.025	0.200	0.620	87.69	71.80	0.169	0.737
	✓	✗	✗	0.041	3.005	0.218	0.620	87.27	72.66	0.169	0.735
	✓	✓	✗	0.040	3.005	0.203	0.619	87.99	72.47	0.149	0.783
	✓	✓	✓	**0.037**	**2.729**	**0.199**	**0.622**	**91.66**	**75.76**	**0.102**	**0.807**

Table 5. **Effect of adding unlabeled videos via flow supervision.** For each backbone (VGGT [38] and π^3 [42]), we compare a 3D supervision only baseline, a labeled multi-task model (geometry + flow), and the full model incorporating additional unlabeled data via flow supervision. This demonstrates the effectiveness of unlabeled data via flow supervision for visual geometry learning.

that all model variants (except `3d-sup`) are trained with the same total number of sequences.

Dynamic Scenes. We use the OmniWorld [50] dataset for 3D supervision and evaluation. As a scalable source of unlabeled dynamic data, we incorporate videos from SpatialVID [39] for flow supervision, using UFM [46] to extract the pseudo-ground-truth 2D flow labels.

We evaluate both camera pose and geometry predictions. Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) measure the accuracy of predicted camera rotations and translations, while Chamfer Distance (CD) and Mean-Square Error (MSE) assess the geometric correctness of the reconstructed 3D structure.

Comparisons of Flow Prediction Mechanisms. The results in Tab. 1 show that, compared to the base model `3d-sup`, our `flow-factored` model yields significant improvements in camera and geometry prediction. We find that `flow-factored` also outperforms both flow-prediction alternatives (`flow-projective` and `flow-tracking`), achieving higher camera pose accuracy and geometric quality in both static and dynamic scenes. Notably, even though `flow-tracking` learns to predict accurate flow (see the appendix), it provides almost no improvement in pose accuracy and geometric quality, suggesting that supervising flow prediction from pairwise patch features may help the network to learn discriminative features, but does not meaningfully benefit visual geometry learning. Moreover, `flow-projective` even degrades performance, indicating that supervising flow computed from explicit camera and geometry predictions may suffer from instability and thus harm learning. Qualitatively, Fig. 4 shows that `flow-factored` yields cleaner reconstructions than `3d-sup` while also improving over other flow mechanisms. Although our factored flow prediction is *suboptimal* for standalone flow estimation – since it enforces an information bottleneck by conditioning on the target-view camera token rather than patch features that contain richer visual cues (see appendix for evaluation and analysis) – these results demonstrate that it provides a more effective supervisory signal for visual geometry learning.

Scaling with Unlabeled Data. To investigate the scaling behavior of the proposed factored flow prediction, we progressively increase the number of unlabeled dynamic videos used for flow supervision on SpatialVID. Specifically, we keep the 3D-labeled OmniWorld set fixed (1K sequences) and scale the number of SpatialVID sequences used to apply the flow loss (3K, 10K, and 20K sequences). For reference, we also train a model using only ad-

ditional labeled data by increasing OmniWorld to 4K sequences without any unlabeled videos. The results in Tab. 4 show that scaling the total number of training sequences to a larger regime (*e.g.*, 10× or 20× the amount used in the `3d-sup` no-flow baseline) yields consistent improvements. Notably, using 20K unlabeled videos together with 1K labeled sequences outperforms training with 4K labeled sequences alone. Qualitatively, Fig. 5 exhibits the same trend: increasing the amount of unlabeled flow supervision leads to visibly cleaner and more complete reconstructions, reinforcing the quantitative gains. These results confirm that factored flow prediction enables visual geometry learning to scale effectively with large amounts of unlabeled video data.

4.2. Improving SoTA Models with Unlabeled Data

We demonstrated in Sec. 4.1 that our factored flow supervision consistently improves visual geometry learning, motivating us to explore its efficacy in improving state-of-the-art models (*e.g.*, π^3 [42]) by incorporating a flow prediction module and leveraging large-scale unlabeled dynamic video data as auxiliary supervision.

Training Overview. Our model training proceeds in two stages. First, we initialize our model from pretrained visual geometry model checkpoint and append our factored flow prediction head to the model. In this stage, we keep the backbone frozen and use all labeled data to train only the newly added flow head (*i.e.*, the gradients from flow supervision do not update the visual geometry backbone). In the second stage, we add additional unlabeled video datasets for training, and unfreeze the whole model, performing end-to-end finetuning using both labeled and unlabeled data.

Datasets. We finetune the base model on a diverse mixture of labeled 3D/4D datasets and large-scale unlabeled video data. Our labeled supervision set consists of eleven widely-used multi-view reconstruction datasets – CO3Dv2 [32], Habitat [33], ARKitScenes [3], ScanNet [8], ScanNet++ [44], MegaDepth [23], BlendedMVS [23], StaticThings3D [35], Omniworld [50], PointOdyssey [49], and VKITTI [16] – which provide ground-truth camera poses and geometry for supervised learning, including approximately 34K sequences in total. To further scale correspondence supervision, we incorporate three unlabeled 4D video datasets: Kinetics-700 [6], SpatialVID [39] (high-quality dynamic sequences), and EPIC-Kitchens [9]. These datasets together contribute approximately 800K video sequences, offering substantial appearance and motion diversity for learning from unlabeled videos.

Baselines and Metrics. Our primary model, denoted as `Flow3r`, is built upon the π^3 [42] backbone. We compare `Flow3r` with representative feed-forward visual geometry baselines: CUT3R [40], VGGT [38], and π^3 [42]. We evaluate performance using pose accuracy and reconstruction metrics on four dynamic datasets using sparsely-sampled video frames as input: Kinetics700 [6], Epic-Kitchens [9], Sintel [5], and Bonn [30]. For Kinetics700 and Epic-Kitchens, we compute pseudo ground truth from dense videos using MegaSAM [24]. We also evaluate on four static datasets: CO3Dv2 [32], ScanNet [8], NRGBD [2], and 7-Scenes [36]. Following prior work [7, 45], we report Relative Pose Error, including RPE (trans) and RPE (rot) for dynamic datasets, and RTA and AUC for static datasets. We assess 3D geometry using mean squared error (MSE) between the predicted and ground-truth points, which captures overall geometric fidelity, and F-score to evaluate the accuracy-completeness trade-off.

Results. We report quantitative results in Tab. 2 and Tab. 3, comparing `Flow3r` against state-of-the-art feed-forward visual geometry prediction models. For both pose estimation and scene reconstruction, `Flow3r` consistently outperforms baselines. We see particularly large improvements for in-the-wild dynamic videos, where our unlabeled data provides additional supervisory signals, but also note an improvement on static datasets, indicating better generalization via large-scale supervision.

We also include qualitative results on in-the-wild multi-view images of static scenes, interaction videos and dynamic scenes in Fig. 6. We observe that `Flow3r` infers cleaner and more accurate scene structure than baselines, which often produce misaligned outputs (*e.g.*, duplicated static components in room reconstruction) and fail to capture the motion of dynamic components (*e.g.*, incorrect movement of the cat).

Ablating the Effects of Unlabeled Data. Our main results show clear improvements on the pre-trained model using our framework, but these conflate additional finetuning and multi-task training on labeled data with the incorporation of unlabeled data. We perform additional experiments to disentangle these effects. First, we finetune the base model using only labeled data without the flow head, establishing a camera-and-geometry-only baseline. Second, we attach the flow head and train with labeled supervision only, forming a multi-task training setup that jointly optimizes camera, geometry and flow. Comparing these two settings measures the impact of multi-task learning itself, while comparing the labeled-only multi-task model with the full `Flow3r` reveals the benefit of incorporating unlabeled videos. To verify that this observation is not architecture-specific, we additionally repeat the same training procedure using the VGGT [38] backbone. All settings use the same total number of training iterations for a fair comparison. Tab. 5 shows that while additional finetuning and multi-task training on labeled data improves performance, the primary performance gains stem from incorporating unlabeled videos via flow supervision.

5. Discussion

In this work, we present `Flow3r` and demonstrate that it effectively leverages in-the-wild unlabeled data by introducing factored flow prediction, advancing visual geometry learning beyond existing fully supervised methods. While our approach opens up new possibilities, several challenges remain. First, `Flow3r` relies on off-the-shelf models to provide pseudo-ground-truth flow supervision, and there can be domains where such 2D prediction fails. Second, although our factored flow formulation elegantly handles dynamic scenes and enables flow supervision to improve the learning of both camera motion and scene geometry, `Flow3r` may struggle under complex scenes with multiple independently moving components. Finally, our current experiments operate at a moderate scale (∼800K video sequences for flow supervision), and scaling to truly large-scale settings (∼10-100M videos) presents an exciting but unexplored direction, and we envision `Flow3r`'s formulation serving as a building block for future large-scale learning methods.

Acknowledgements. We thank the members of the Physical Perception Lab at CMU for their valuable discussions. This work was supported by an NVIDIA academic grant. This work used Bridges-2 [4] at Pittsburgh Supercomputing Center through allocation CIS250061 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work was supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] 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. 1, 3

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *CVPR*, 2022. 8

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS D&B*, 2021. 8

[4] Shawn T Brown, Paola Buitrago, Edward Hanna, Sergiu Sanielevici, Robin Scibek, and Nicholas A Nystrom. Bridges-2: A platform for rapidly-evolving and data intensive research. In *Practice and experience in advanced research computing*, 2021. 9

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 8

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2019. arXiv preprint arXiv:1907.06987. 8

[7] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *CVPR*, 2024. 8

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 8

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 8

[10] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS*, 2022. 2

[11] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *CVPR*, 2023. 2

[12] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023. 2

[13] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, 2024. 2, 5

[14] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. In *IEEE TPAMI*, 2018. 2

[15] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *ECCV*, 2010. 1

[16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 8

[17] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 2

[18] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 1981. 2

[19] 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 arXiv:2508.10934*, 2025. 2

[20] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, 2024. 2

[21] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 2

[22] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2

[23] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 8

[24] 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 *CVPR*, 2025. 2, 8

[25] Yiqing Liang, Abhishek Badki, Hang Su, James Tompkin, and Orazio Gallo. Zero-shot monocular scene flow estimation in the wild. In *CVPR*, 2025. 3

[26] Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *ECCV*, 2024. 1

[27] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981. 2

[28] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Trans. Robot.*, 2015. 2

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2024. 3, 6

[30] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *IROS*, 2019. 8

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 4

[32] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 8

[33] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 8

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 3

[35] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *3DV*, 2022. 8

[36] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 8

[37] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM SIGGRAPH*, 2006. 1

[38] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 3, 4, 6, 7, 8

[39] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations, 2025. arXiv preprint arXiv:2509.09676. 7, 8

[40] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 1, 2, 8

[41] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1, 2

[42] 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 *ICLR*, 2026. 2, 4, 7, 8

[43] Changchang Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013. 1

[44] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 6, 8

[45] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025. 2, 8

[46] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. In *NeurIPS*, 2025. 2, 4, 5, 7

[47] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *ECCV*, 2022. 2

[48] Qitao Zhao, Amy Lin, Jeff Tan, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion. In *CVPR*, 2025. 1, 2

[49] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 8

[50] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling, 2025. arXiv preprint arXiv:2509.12201. 7, 8

[51] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer, 2025. arXiv preprint arXiv:2507.11539. 2