FROM VISIBILITY TO GEOMETRY: EFFICIENT TRAINING FOR FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

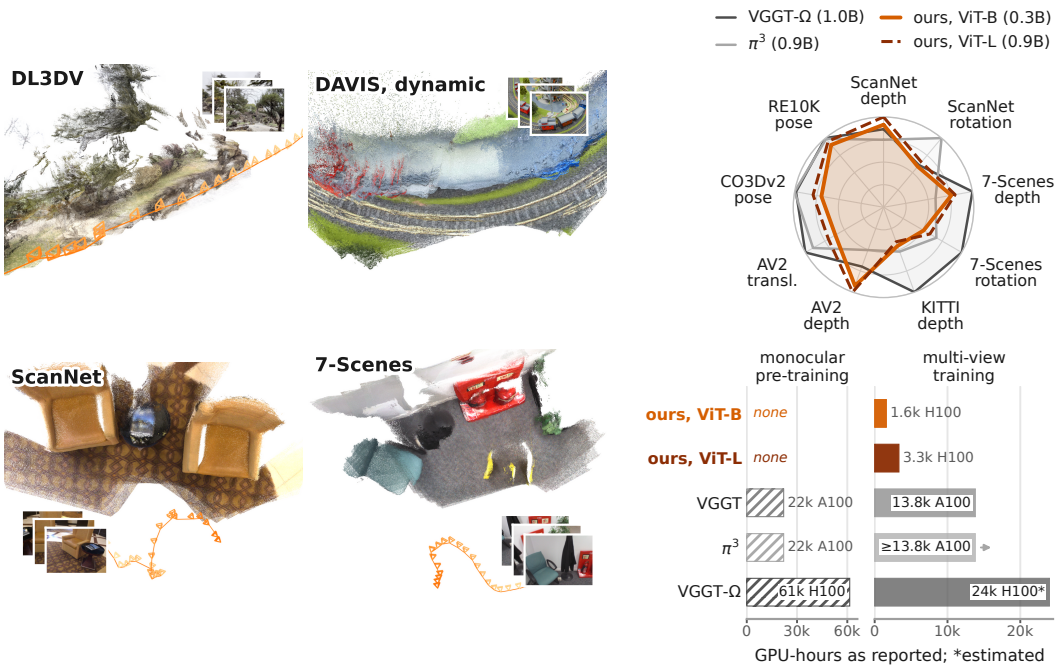


Figure 1: **Towards efficient training of feed-forward 3D reconstruction models.** Restructuring the training process around geometric relationships substantially improves the efficiency of learning multi-view geometry. We first pre-train the network from scratch to predict covisibility maps derived from the ground-truth geometry, and then fine-tune it to regress camera poses and depth maps. Our models can reconstruct both static and dynamic scenes (left: our ViT-B on held-out clips), achieve performance close to released feed-forward models on held-out benchmarks (top right: the outer ring shows the best model on each axis), and do so with a fraction of their training compute (bottom right).

ABSTRACT

Feed-forward reconstruction models, such as VGGT-Ω and π^3, have demonstrated impressive reconstruction performance, often leveraging powerful pre-trained visual features such as DINOv2 or DINOv3. In this work, we show that restructuring the training process around geometric relationships can substantially improve the efficiency of learning multi-view geometry. Our key insight is that predicting covisibility for multiple pairs of images provides a strong, unified learning signal for geometric relationships across views. Instead of directly predicting camera poses and depth maps, we first pre-train the network to predict covisibility maps derived from the ground-truth geometry, and then fine-tune it to regress camera poses and depth maps. This two-stage training strategy substantially accelerates convergence, yielding significantly better reconstruction performance under a fixed compute budget. Remarkably, our approach achieves competitive reconstruction quality when trained entirely from scratch, without relying on any pre-trained visual features. Code and models will be made available.

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.

1 INTRODUCTION

Feed-forward 3D reconstruction models have demonstrated impressive reconstruction capabilities by recovering depth and camera poses for many images in a single forward pass. This line of work was pioneered by DUSt3R and MASt3R, which regress pointmaps for pairs of images (Wang et al., 2024; Leroy et al., 2024). Subsequent methods, including VGGT, π^3, Depth Anything 3, and VGGT-Ω, extended this paradigm to large image sets (Wang et al., 2025a; 2026b; Lin et al., 2026; Wang et al., 2026a). Meanwhile, the initialization of these models has evolved considerably. DUSt3R builds on CroCo, a network pre-trained on pairs of views through cross-view completion (Weinzaepfel et al., 2022; 2023), whereas more recent models initialize their visual encoders from monocular models such as DINOv2 or DINOv3 (Oquab et al., 2024; Siméoni et al., 2026). In this paper, we propose to restructure their training process around geometric relationships to substantially improve the efficiency of learning multi-view geometry, thereby enabling training from scratch without relying on pre-trained visual features.

We argue that, even though a feed-forward reconstruction network ultimately predicts camera poses and depth maps, directly regressing these quantities provides only a weak learning signal for multi-view geometry. A depth map can be predicted remarkably well from a single image (Yang et al., 2024; Wang et al., 2025d), while a camera pose amounts to only a few scalars per image. Consequently, these objectives provide limited supervision for learning the dense geometric relationships between views.

Our key insight is that predicting covisibility for multiple pairs of images provides a strong, unified learning signal for geometric relationships across views. For a pair of views, a dense covisibility map indicates whether each pixel in the source view is visible in the target view, occluded by another surface, or outside the target view's field of view. Each pixel therefore receives a label that can only be predicted by relating the two views, requiring cross-view geometric reasoning. Supervising the network with covisibility maps for multiple image pairs thus encourages it to learn geometric relationships across diverse viewpoints.

Therefore, instead of directly predicting camera poses and depth maps, we first pre-train the network to predict covisibility maps derived from the ground-truth geometry, and then fine-tune it to regress camera poses and depth maps. Remarkably, our approach achieves competitive reconstruction quality when trained entirely from scratch, without relying on any pre-trained visual features. We find that this two-stage training is important for efficient learning: training from scratch to jointly predict covisibility maps, camera poses, and depth maps is substantially less efficient. We conjecture that this is because camera poses and depth maps constitute a disentangled representation of the 3D world, whereas covisibility captures geometric relationships between views in a more entangled form. Jointly learning these two representations requires the network to simultaneously discover cross-view geometric relationships and disentangle them into camera poses and depth maps. Our two-stage training therefore acts as a curriculum: the first stage learns a geometrically entangled representation through multi-view covisibility prediction, while the second stage learns to disentangle this representation into camera poses and depth maps. To further support this hypothesis, we find that fine-tuning the network while keeping the backbone frozen already produces strong results, suggesting that the features learned during the first stage already capture information useful for pose and depth estimation.

In summary, our contributions are the following:

1. We introduce multi-view covisibility prediction to transform geometric relationships into a unified learning signal for multi-view geometry, providing dense supervision that explicitly requires cross-view geometric reasoning.
2. We propose a two-stage training strategy that first learns a geometrically entangled representation through multi-view covisibility prediction and then fine-tunes it to regress camera poses and depth maps. This curriculum enables efficient training of feed-forward reconstruction models from scratch, without pre-trained visual features.
3. Under matched compute, we show that our two-stage training substantially improves camera and depth estimation over training from scratch. Moreover, our approach scales to public datasets under academic compute budgets, achieving competitive in-domain depth performance with a fraction of the training compute used by released models.

2 RELATED WORK

Feed-forward multi-view geometry. DUSt3R regresses pointmaps for pairs of images (Wang et al., 2024) and MASt3R adds dense matching (Leroy et al., 2024). Later models process image sets with a spatial memory (Wang & Agapito, 2025; Wang et al., 2025b), multi-view decoders (Cabon et al., 2025; Tang et al., 2025; Yang et al., 2025), or attention alternating within and across frames (Wang et al., 2025a; 2026b;a), and extend to dynamic scenes, structure-from-motion, relocalization and metric reconstruction (Zhang et al., 2025; Elflein et al., 2025; Dong et al., 2025; Keetha et al., 2026; Lin et al., 2026). Like monocular depth estimators (Ranftl et al., 2021; Yang et al., 2024; Piccinelli et al., 2024; Wang et al., 2025c), they start from pre-trained weights: CroCo for DUSt3R and the models built on it, DINOv2 or DINOv3 for VGGT, π^3, Depth Anything 3 and VGGT-Ω. We ask what that initialization should be.

Self-supervised and multi-view pre-training. Single-image objectives learn invariance across augmentations (Chen et al., 2020; Grill et al., 2020; Caron et al., 2021; Zhou et al., 2022), reconstruct masked patches (Bao et al., 2022; He et al., 2022; Xie et al., 2022; Bachmann et al., 2022; Yang et al., 2026), or predict masked content in feature space (Assran et al., 2023; 2025). At scale they yield the DINO encoders (Oquab et al., 2024; Siméoni et al., 2026), and video variants learn from the frames of one camera (Tong et al., 2022; Gupta et al., 2023; Venkataramanan et al., 2024). Such features encode depth within an image but are not consistent across large changes of viewpoint (El Banani et al., 2024). SeeSE3 measures how much 3D structure they expose (Chen et al., 2026). Pose and matching supervision also yield transferable features (Zamir et al., 2016; 2018). CroCo reconstructs a masked view from a second, overlapping one (Weinzaepfel et al., 2022; 2023), which transfers to stereo, optical flow and zero-shot correspondence (An et al., 2025). PMatch targets dense matching (Zhu & Liu, 2023), MuM and Muskie mask whole image sets (Nordström et al., 2026; Li et al., 2025), and RayZer and E-RayZer pre-train by self-supervised view synthesis (Jiang et al., 2025; Zhao et al., 2026). Gekko turns the gap between cross-view and single-view reconstruction errors into a covisibility signal (Loiseau et al., 2026), and Alligat0R predicts dense covisibility for image pairs (Loiseau et al., 2025), which we extend to many views.

Covisibility and correspondence. Which pixels two views share underlies structure-from-motion, SLAM (Schönberger & Frahm, 2016; Mur-Artal et al., 2015) and learned matching (Sarlin et al., 2020; Lindenberger et al., 2023; Sun et al., 2021; Edstedt et al., 2023; 2024). Other work reasons about occluded or unseen regions (Fan et al., 2023; Germain et al., 2022), regresses pose from covisible regions (Hutchcroft et al., 2022), or probes VGGT for image overlap (Ziliotto et al., 2026). With labels from a trained model, our pre-training resembles pseudo-labeling and distillation (Hinton et al., 2014; Tarvainen & Valpola, 2017; Xie et al., 2020; Sohn et al., 2020), on a discrete target derived from the model's geometry.

3 METHOD

Our goal is to improve the training efficiency of feed-forward 3D reconstruction models. To this end, we leverage multi-view covisibility prediction. Our approach is illustrated in Figure 2.

3.1 WHY MULTI-VIEW COVISIBILITY?

We start from the observation that camera poses and depth maps constitute a disentangled representation of the 3D world. Directly supervising these quantities may therefore be inefficient for learning multi-view geometry. For example, a depth map can be predicted remarkably well from a single image (Yang et al., 2024; Wang et al., 2025d), making depth regression alone a weak incentive for the network to learn multi-view geometry. We instead turn to multi-view covisibility prediction, which directly encodes geometric relationships between views in a unified manner.

Predicting a covisibility map between a source image and a target image requires dense cross-view geometric reasoning. Every pixel in the source view receives a label that depends on the target view, indicating whether it is visible, occluded, or outside the target's field of view. Supervising the network with covisibility maps for multiple image pairs thus encourages it to learn geometric relationships across diverse viewpoints. At 192×256 resolution, a single pair provides up to $49{,}152$

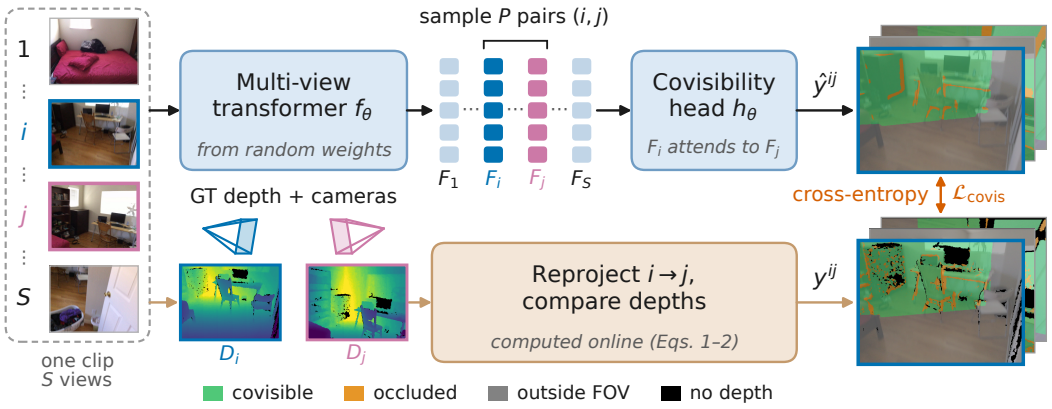


Figure 2: **Multi-view covisibility pre-training.** A multi-view transformer encodes the S views of a clip in a single forward pass. We sample P view pairs (i, j) from its output tokens and use a head in which the tokens of view i attend to those of view j to predict, for every pixel in view i, whether it is visible in view j, occluded by another surface, or outside its field of view. A target covisibility map y^{ij} is computed on the fly during training using a depth check involving ground-truth poses and depths.

pixel-level labels, while the 32 pairs we sample from a clip of 24 views provide up to 1.6 million labels, against 216 camera parameters (nine for each of the 24 views) for the same clip. All of these are obtained from a single forward pass of the encoder and optimized with a single pixel-wise cross-entropy loss. This suggests that multi-view covisibility can provide a more sample-efficient learning signal. Point maps in a common frame, as regressed by DUSt3R and VGGT (Wang et al., 2024; 2025a), are also dense and require cross-view geometric reasoning. We believe that multi-view covisibility has three advantages over point maps. First, it is a classification task, which can be more robust to noise in the ground-truth geometry. Second, it does not suffer from the scale ambiguity inherent to 3D point regression. Third, covisibility labels are obtained through a depth check, which makes them relatively insensitive to small errors in the underlying depth or camera poses, whereas regression targets directly inherit these errors.

3.2 Covisibility labels from depth and cameras

For image $\mathtt{I}_i$, let $\mathtt{D}_i$, $\mathtt{K}_i$, and $(\mathtt{R}_i, \mathbf{t}_i)$ denote its depth map, its intrinsics, and its camera-to-world pose respectively. A pixel location $\mathbf{p}$ of the source view i with valid depth can be lifted to 3D ($\mathbf{x}_i$), transformed into the coordinate system of the target view j ($\mathbf{x}_j$) and finally reprojected into the image domain ($\mathbf{q}$) as follows:

$$\mathbf{x}_i = \mathtt{D}_i(\mathbf{p})\mathtt{K}_i^{-1}\tilde{\mathbf{p}}, \qquad \mathbf{x}_j = \mathtt{R}_j^\top(\mathtt{R}_i\mathbf{x}_i + \mathbf{t}_i - \mathbf{t}_j), \qquad \mathbf{q} = \pi(\mathtt{K}_j\mathbf{x}_j), \quad z_j = (\mathbf{x}_j)_3, \tag{1}$$

where $\tilde{\mathbf{p}}$ is $\mathbf{p}$ in homogeneous coordinates and π is the projection function (i.e. divides by the third coordinate). In practice, we round $\mathbf{q}$ to its nearest neighbor $\bar{\mathbf{q}}$. The source pixel $\mathbf{p}$ is *outside the field of view* when $\bar{\mathbf{q}}$ falls outside the target image or when $\mathbf{x}_j$ is behind the camera ($z_j \leq 0$). Otherwise, if the target has valid depth at $\bar{\mathbf{q}}$, $\mathbf{p}$ is defined as *covisible* when the following relative depth check is satisfied and *occluded* when it is not:

$$y^{ij}_{\mathbf{p}} = \begin{cases} 0 & \text{if } |z_j - \mathtt{D}_j(\bar{\mathbf{q}})| \leq \max(\tau_r \mathtt{D}_j(\bar{\mathbf{q}}), \tau_a), \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Pixels without source depth, or that land on a target pixel without depth, are ignored.

Covisibility labels require only depth and cameras. With metric ground-truth geometry, we use $\tau_r = 0.05$, and $\tau_a = 0.05$ meters. Non-metric ground-truth geometry has an arbitrary scale, so the absolute floor τ_a becomes 5% of the median depth, and we use $\tau_r = 0.15$.

3.3 SHARED ENCODING AND PAIRWISE PREDICTION

The encoder $f_{\boldsymbol{\theta}}$ is a multi-view vision transformer (Dosovitskiy et al., 2021) with rotary position embeddings (Su et al., 2024) and alternating frame-wise and inter-frame attention, following the VGGT family (Wang et al., 2025a; 2026a). Unlike these models, it starts from random weights and tokenizes each view with a convolutional patch embedding rather than a pre-trained DINO encoder. We use ViT-S in our ablation study and ViT-B and ViT-L when scaling (see Appendix A for implementation details).

The covisibility head $h_{\boldsymbol{\phi}}$ takes the output tokens of a source view i and a target view j. Two cross-attention blocks let each source token attend to the target tokens, and a linear layer maps each source token to $16 \times 16 \times 3$ logits, which are unpatchified into a three-class prediction for every pixel. At each training step, we sample P pairs without replacement from the $S(S-1)$ pairs of the clip. All pairs share one encoder forward pass, so additional pairs only cost head computation. Let $\mathcal{V}$ contain all labeled pixels over the batch and the sampled pairs. The loss is the cross-entropy averaged over this set:

$$\mathcal{L}_{\text{covis}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{|\mathcal{V}|} \sum_{(b,i,j,\mathbf{p}) \in \mathcal{V}} \text{CE}\big(h_{\boldsymbol{\phi}}(\mathtt{F}_b, i, j)_{\mathbf{p}}, y^{ij}_{b,\mathbf{p}}\big), \qquad \mathtt{F}_b = f_{\boldsymbol{\theta}}(\mathtt{I}_{b,1:S}). \tag{3}$$

Following VGGT (Wang et al., 2025a), we draw the number of views S uniformly between 2 and 24 for each batch and adjust the number of clips so that every step processes the same number of images. The number of pairs scales with the number of views as $P = \lceil 32S/24 \rceil$, so that each image receives the same amount of supervision: a clip of 24 views contributes 32 of its 552 pairs.

Joint objective. Covisibility labels are undefined where the ground truth has no depth, so they leave holes in the supervision. To also supervise these regions, we can add multi-view masked image modeling, the objective of MuM (Nordström et al., 2026): 75% of the patches of every view are masked, and a four-block decoder regresses their normalized pixels with a mean squared error. The two objectives are complementary, since reconstructing a view from the others also depends on which pixels they share, and it is shown in Gekko (Loiseau et al., 2026) that covisibility can be inferred from masked image modeling. In this *joint* objective, one masked forward pass feeds both the pixel decoder and the covisibility head, which predicts from the mask tokens the labels of the complete image. We experimentally observe that this extra loss allows to slightly improve the performance of our models.

3.4 TRANSFER AND COMPUTE BUDGET

After pre-training, we discard the pretext head and attach new DPT depth (Ranftl et al., 2021) and camera heads, which we train on ground-truth depth and cameras with the same loss as in VGGT-Ω (Wang et al., 2026a). The encoder is either frozen, which probes the representation, or fine-tuned with the head, which measures its value as an initialization. Fine-tuning draws the number of views as in pre-training, so that the model accepts any number of input images.

In the ablation study on ScanNet (Section 4.2), every model is trained for 120k optimizer steps at 1,536 images per step (64 clips at $S = 24$) of 192×256 pixels, so that pre-training is compared with supervised training from scratch at equal cost. Pre-trained models split these into 60k pre-training and 60k transfer steps, whereas the scratch baseline spends all 120k on depth and cameras. The scaled models follow longer schedules on more data and are not part of this matched comparison. Appendix A provides further details.

4 EXPERIMENTS

We first detail our experimental setup (§4.1). Thereafter, we perform an ablation study at a moderate scale (§4.2), then compare our scaled models with released feed-forward models (§4.3), and finally examine what their frozen encoders have learned (§4.4). Section 4.5 discusses limitations.

4.1 Experimental Setup

Ablation setting. A ViT-S is pre-trained from random weights on ScanNet (Dai et al., 2017a), from which 120 scans are held out for evaluation, and depth and camera heads are then trained on ScanNet with the encoder either frozen or fine-tuned. Pre-trained models spend 60k steps on the pretext task and 60k on depth and cameras, while models trained from scratch spends all 120k steps on depth and cameras. Covisibility labels are computed from ground-truth depth and cameras unless stated otherwise.

Scaled setting. We train a ViT-B and a ViT-L from random weights and fine-tune both on eighteen datasets with ground-truth depth and cameras, which cover indoor scans, synthetic scenes, object captures and driving. The last fine-tuning steps are done at 384×512 pixels. Every run processes $1{,}536$ images from clips per step on eight H100 GPUs, and the complete training of the ViT-B and of the ViT-L takes 1.6k and 3.3k H100-hours. Appendix A lists the datasets and gives the schedules.

Benchmarks and metrics. We evaluate clips of 24 views at 192×256 pixels: 960 clips from the held-out ScanNet scans, and 7-Scenes (Shotton et al., 2013), DL3DV (cameras only) and KITTI (Geiger et al., 2013), which no model of the ablation setting sees in training. The scaled models of §4.3 are fine-tuned on DL3DV and on driving datasets that include KITTI-360 and Virtual KITTI 2, a synthetic re-rendering of KITTI sequences, so for them only 7-Scenes is out of domain. In Table 3, every model runs at its own training resolution, ours at 384×512, on the same clips upsampled. Depth is scored by AbsRel after one median scale alignment per clip,

$$s = \mathrm{median}_{p \in V_+} \frac{D(p)}{\hat{D}(p)}, \qquad E_{\mathrm{AbsRel}} = \frac{1}{|V|} \sum_{p \in V} \frac{|s\hat{D}(p) - D(p)|}{D(p)}, \tag{4}$$

where D and $\hat{D}$ are the ground-truth and predicted depth, V contains the pixels with valid ground truth in all views, and $V_+ \subseteq V$ those with a positive prediction. For cameras, we report the mean rotation error and the mean translation-direction error relative to the first view, in degrees. On CO3Dv2 (Reizenstein et al., 2021) and RealEstate10K (Zhou et al., 2018), we follow the protocol of VGGT (Wang et al., 2025a) and report the pose AUC at $30°$. Correspondence is scored on held-out ScanNet, ScanNet-1500 (Sarlin et al., 2020) and NAVI (Jampani et al., 2023). Appendix B gives the details. We call a difference between two models significant when the 95% confidence interval of their paired difference, obtained by bootstrap resampling of the test scenes (Efron, 1979), excludes zero. Two training runs of the same recipe differ by 1 to 3% in ScanNet AbsRel (Appendix E).

4.2 Ablation Study

Table 1 compares pre-training with training from scratch under the same budget. Training from scratch is not limited by its schedule: doubling it from 60k to 120k steps lowers ScanNet AbsRel only from 0.0960 to 0.0863. A pre-trained encoder does better even when it is frozen, so that the depth and camera heads can only read out what pre-training has put into the representation: with covisibility, ScanNet depth error is 32% lower and rotation error 44% lower, and every benchmark except KITTI improves significantly. Fine-tuned, covisibility pre-training reduces ScanNet depth error by 55%, from 0.0863 to 0.0385, and translation-direction error by 71%, and it also improves the three out-of-domain benchmarks, by 15 to 37%: every difference is significant.

Joint objective and single-stage training. Adding masked image modeling, which also supervises the regions without covisibility labels (§3.3), improves the frozen encoder on all six columns and the fine-tuned one on five, lowering its ScanNet depth error to 0.0316, 63% below training from scratch. The two fine-tuned models are level on KITTI depth. Splitting training into two stages is also important: optimizing the same four losses at once for 120k steps reaches only 0.0516 ScanNet AbsRel and is significantly less accurate than the two-stage model on every benchmark. With the same budget, the two-stage joint ViT-S also matches a frozen DINOv3 ViT-L (Siméoni et al., 2026) on ScanNet depth and is significantly more accurate on cameras (Appendix D.1), although DINOv3 was pre-trained on far more data.

Robustness to label noise. Covisibility labels need only depth and cameras, so they can also be computed from a model's predictions, which provides a test with very noisy labels. We train

	ScanNet			7-Scenes	DL3DV	KITTI
pretraining	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓
from scratch						
none, 60k steps	0.0960	8.60	37.9	0.1656	36.34	0.2565
none, 120k steps	0.0863	7.70	37.2	0.1632	35.26	0.2543
none, 120k steps, all losses at once	0.0516	4.01	14.5	0.1345	23.46	0.2534
pretrained, encoder frozen						
covisibility	0.0586	4.34	15.5	0.1231	25.22	0.2560
covisibility + masked modeling	0.0537	3.72	12.7	0.1126	**19.68**	0.2233
pretrained, encoder fine-tuned						
covisibility	0.0385	3.13	10.7	0.1112	22.15	0.2167
covisibility + masked modeling	**0.0316**	**2.73**	**9.3**	**0.0946**	20.38	**0.2161**

Table 1: **Pretraining against training from scratch** (ViT-S, ScanNet). The first two models spend all of their steps on depth and cameras, and the third optimizes the covisibility, masked modeling, depth and camera losses at once. Pretrained encoders spend 60k steps on the pretext task, and depth and camera heads are then trained for 60k steps with the encoder frozen or fine-tuned, so every model except the first trains for 120k steps. Rotation and translation-direction errors are in degrees. Best per column in **bold**.

	ScanNet			7-Scenes	DL3DV	KITTI
pretraining	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓
none (from scratch, 120k steps)	0.0863	7.70	37.2	0.1632	35.26	0.2543
covisibility, labels from that ViT-S	0.0474	4.15	13.7	0.1320	26.81	0.2264
covisibility, ground-truth labels	**0.0385**	**3.13**	**10.7**	**0.1112**	**22.15**	**0.2167**

Table 2: **Robustness to label noise** (ViT-S, ScanNet, fine-tuned). The supervised ViT-S is trained for 120k steps to regress depth and cameras. Covisibility labels computed from its predictions are used to pretrain the model below it. Rotation and translation-direction errors are in degrees.

a ViT-S to regress depth and cameras on ScanNet for 120k steps and compute covisibility labels from its predictions (Table 2). These labels disagree with the ground-truth ones on about one pixel in five. Surprisingly, a second ViT-S, pre-trained on them and then fine-tuned on the same data, reaches about half the depth error of the network that produced its labels (0.047 against 0.086 ScanNet AbsRel), and it is significantly more accurate than training from scratch on every benchmark. Ground-truth labels remain better, but the second network receives no information beyond what was available to the first, so this result supports our hypothesis that covisibility is a stronger learning signal than direct regression (§3.1). Appendix D.3 computes the labels from VGGT-Ω instead.

4.3 Comparison with released models

Table 3 compares our scaled ViT-B and ViT-L with VGGT-Ω, π^3 and the monocular MoGe-2 (Wang et al., 2026a;b; 2025d). On ScanNet, the ViT-B matches VGGT-Ω on all three metrics, and the ViT-L is significantly more accurate on all three. Both estimate depth significantly more accurately than π^3 and MoGe-2 on ScanNet and 7-Scenes. The ViT-B has a third of the parameters of VGGT-Ω and π^3, and its entire training took 1.6k H100-hours, about fifteen times fewer than our estimate of 24k for the multi-view training of VGGT-Ω alone, which moreover starts from DINOv3 (see Figure 1). The ViT-L took about seven times fewer. π^3 remains more accurate in camera poses on ScanNet, and both released models on the other benchmarks, especially on CO3Dv2, whose frames surround an object, whereas most of our training clips are nearby video frames. VGGT-Ω is also more accurate on 7-Scenes depth, and VGGT-Ω and MoGe-2 on KITTI depth.

4.4 Frozen encoders and correspondence

Table 4 measures cross-view correspondence directly in frozen features, by matching each query pixel to the most similar patch of another view (Appendix B). On held-out ScanNet, our covisibility

		training GPU-hours		ScanNet			7-Scenes		KITTI	CO3Dv2	RE10K
model	params	DINO	multi-view	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓	AUC@30 ↑	AUC@30 ↑
VGGT-Ω	1.0B	61k H100	24k H100*	0.0301	2.46	9.3	**0.0541**	**1.58**	**0.0337**	87.5	**90.3**
π^3	0.9B	22k A100	≥13.8k A100	0.0348	**1.50**	**6.6**	0.0924	2.34	0.0645	**88.9**	89.3
MoGe-2 (monocular)	0.3B	22k A100	–	0.0731	–	–	0.0959	–	0.0505	–	–
ours: ViT-B	0.3B	none	1.6k H100	0.0285	2.55	8.8	0.0709	3.09	0.0741	62.1	81.4
ours: ViT-L	0.9B	none	3.3k H100	**0.0260**	2.30	8.0	0.0672	2.67	0.0829	70.4	86.1

Table 3: **Comparison with released feed-forward models**, each at its own training resolution on the same held-out clips. Training GPU-hours are given as reported, for the DINO encoder a model starts from and for its own multi-view training, which for π^3 is at least that of VGGT, from which it starts, and for VGGT-Ω is our estimate (*, Appendix A). Rotation and translation-direction errors are in degrees; CO3Dv2 and RE10K report pose AUC@30. Best per column in **bold**.

	8 views, PCK@25		ScanNet-1500, PCK@16	
encoder	ScanNet	NAVI	frozen	linear probe
released				
DINOv3 ViT-L/16	0.778	0.762	0.482	0.633
MuM ViT-L/16	0.768	0.653	0.422	0.679
Muskie ViT-L	0.854	0.739	0.497	0.654
VGGT-Ω	0.737	0.790	0.457	0.804
π^3	0.756	0.766	0.485	0.579
DA3-L	0.810	**0.797**	0.522	0.762
Ours, ViT-B	**0.984**	0.649	**0.757**	**0.819**

Table 4: **Correspondence from frozen features**: the fraction of query pixels whose most similar patch in another view lies within 25 or 16 pixels of the true match. The linear probe trains one linear map on ScanNet training pairs, following Nordström et al. (2026). Best per column in **bold**.

encoder places 98% of the queries within 25 pixels of their true match, against at most 85% for six released networks. Figure 3 shows such matches for three views of one scan. On ScanNet-1500 (Sarlin et al., 2020), whose scenes lie outside of training data, it also leads, both without training and with the linear probe of MuM (0.757 and 0.819). This ability is what covisibility implicitly supervises, so the result is expected, but it shows that the encoder learns correspondence rather than leaving it to the pretext head. On the object captures of NAVI (Jampani et al., 2023), whose views surround each object, the released models score higher than ours, which suggests, as for the camera poses of CO3Dv2, that the gap comes from our pre-training clips, mostly nearby video frames.

Figure 4 shows the principal components of these frozen features. For our encoder, the first two components vary smoothly across the scene: a linear function of the 3D position of each patch explains 90% and 89% of their variance, more than its position in the image (88% and 76%), so the colors follow surfaces as the camera moves. For VGGT-Ω, 3D position explains 83% of the second component but only 4% of the first, and its features at this layer are dominated by high-frequency patterns. DINOv3, MuM and Muskie instead separate regions such as the carpet and the armchairs (Appendix D.4). Appendix D.2 evaluates these frozen encoders on depth and cameras.

4.5 Limitations

Our models trail released models on camera estimation outside ScanNet and on out-of-domain depth, and our correspondence advantage does not extend to object captures (Tables 3 and 4). Covisibility labels also require dense depth, since a pixel is labeled only when both it and its reprojection have depth, which limits the use of datasets with sparse LiDAR or structure-from-motion depth. Computing the labels from a model's geometric predictions would provide dense supervision where LiDAR or structure-from-motion depth is sparse, and would allow pre-training to scale to unannotated images and videos. Our experiments with labels computed from a weaker model (Table 2) and from VGGT-Ω (Appendix D.3) suggest that this is a promising direction for future work.

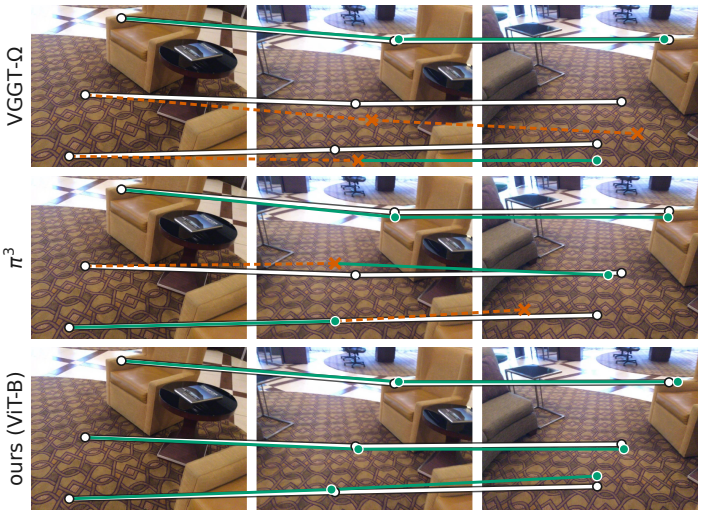


Figure 3: **Matches chained across three views** of a held-out ScanNet clip, from the frozen features of Table 4. Three random query pixels of the first view are matched to the most similar patch of the second view, and each match to the third view. White: ground truth. Green: a match within 25 pixels of it. Vermilion and dashed: a match further away.

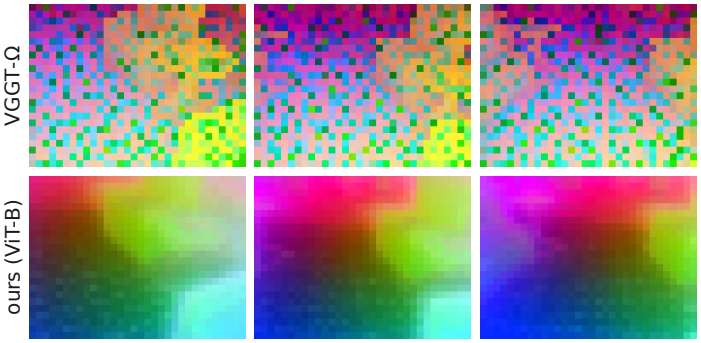


Figure 4: **Principal components of frozen features** on the three views of Figure 3, in the same columns, at 384×512 pixels. For each encoder, one PCA basis is fitted over the eight views of the clip and its first three components are shown as colors, which can be compared across the views of a row. Each encoder is read at the layer where its features match best in Table 4.

5 Conclusion

We have shown that restructuring the training of feed-forward reconstruction models around geometric relationships makes learning multi-view geometry substantially more efficient. Predicting covisibility for many view pairs provides a dense learning signal that requires relating the views. At a fixed number of training steps, pre-training a network from random weights on covisibility before fine-tuning it to regress camera poses and depth maps reduces held-out ScanNet depth error by 55% compared with training from scratch, and it is significantly more accurate than optimizing the same losses in a single stage. Even frozen, the pre-trained encoder is informative, and it matches pixels across views more reliably than those of released models on held-out indoor scenes. Trained from scratch on public data in 1.6k H100-hours, our ViT-B matches VGGT-Ω on ScanNet depth, although our scaled models still trail released models on camera rotation and on out-of-domain depth. Finally, pre-training on noisy labels computed from the predictions of a weaker model yields a network that surpasses it. Such pseudo-labels open a path to self-improvement, in which each model's predictions provide the labels that pre-train its successor, including on unannotated video.

AI USE STATEMENT

Generative AI coding assistants assisted with method implementation, experiment and methodology design, and manuscript preparation. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

Our experiments use public datasets and released models under their licenses. Several of these datasets, including the Aria datasets, Waymo Open, BEDLAM and SpatialVID, restrict their use to non-commercial research. Geometric prediction may expose spatial details of captured environments and inherits biases and errors from its training data.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the architectures, the construction of the labels and the training recipes, and Appendix B the benchmarks, metrics and significance tests. We will release the code and the trained models.

REFERENCES

Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1103–1115, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. SceneScript: Reconstructing scenes with an autoregressive structured language model. In *European Conference on Computer Vision (ECCV)*, 2024.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision (ECCV)*, pp. 348–367, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2. *arXiv preprint arXiv:2001.10773*, 2020.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1060, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Caroline Chen, Sayna Ebrahimi, Fedor Kitashov, Ming-Hsuan Yang, Leonidas Guibas, Viorica Pătrăucean, and Maks Ovsjanikov. SeeSE3: Emergence of 3D space in vision features. *arXiv preprint arXiv:2607.14228*, 2026.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017a.

Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, 36(3), 2017b.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16739–16752, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17765–17775, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3R-SfM: Towards feed-forward structure-from-motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16774–16784, 2025.

Miao Fan, Mingrui Chen, Chen Hu, and Shuchang Zhou. Occ^{2}Net: Robust image matching based on 3D occupancy estimation for occluded regions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9652–9662, 2023.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination. In *International Conference on Learning Representations (ICLR)*, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014.

Will Hutchcroft, Yuguang Li, Ivaylo Boyadzhiev, Zhiqiang Wan, Haiyan Wang, and Sing Bing Kang. CoVisPose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360° indoor panoramas. In *European Conference on Computer Vision (ECCV)*, pp. 615–633, 2022.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. RayZer: A self-supervised large view synthesis model. In *International Conference on Computer Vision (ICCV)*, 2025. Best Student Paper Honorable Mention.

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 *International Conference on 3D Vision (3DV)*, pp. 499–509, 2026.

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

Wenyu Li, Sidun Liu, Peng Qiao, Yong Dou, and Tongrui Hu. Muskie: Multi-view masked image modeling for 3d vision pre-training. *arXiv preprint arXiv:2511.18115*, 2025.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026. Oral.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17627–17638, 2023.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024.

Thibaut Loiseau, Guillaume Bourmaud, and Vincent Lepetit. Alligat0R: Pre-training through co-visibility segmentation for relative camera pose regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight.

Thibaut Loiseau, Guillaume Bourmaud, and Vincent Lepetit. Revisiting cross-view completion: Self-supervised pre-training via reconstruction error comparison. *arXiv preprint arXiv:2609.01530*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

David Nordström, Johan Edstedt, Fredrik Kahl, and Georg Bökman. MuM: Multi-view masked image modeling for 3d vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3D machine perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segù, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *International Conference on Computer Vision (ICCV)*, 2021.

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 *International Conference on Computer Vision (ICCV)*, pp. 10901–10911, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV)*, 2021.

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

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2937, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. DINOv3. *Transactions on Machine Learning Research*, 2026. Featured Certification.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, 2021.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shashanka Venkataramanan, Mamshad Nayeem Rizve, João Carreira, Yuki M. Asano, and Yannis Avrithis. Is ImageNet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *International Conference on Learning Representations (ICLR)*, 2024.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, pp. 78–89, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. Best Paper Award.

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916, 2020.

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 *International Conference on Learning Representations (ICLR)*, 2026b.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *International Conference on Computer Vision (ICCV)*, 2023.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. RGBD objects in the wild: Scaling real-world 3D object learning from RGB-D videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22378–22389, 2024.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves ImageNet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9653–9663, 2022.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 21875–21911, 2024.

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Saining Xie, Hengshuang Zhao, Kaiming He, and Hu Xu. In pursuit of pixel supervision for visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31974–31984, 2026.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Amir R. Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3D representation via pose estimation and matching. In *European Conference on Computer Vision (ECCV)*, pp. 535–553, 2016.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722, 2018.

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 *International Conference on Learning Representations (ICLR)*, 2025.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-RayZer: Self-supervised 3D reconstruction as spatial visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. Highlight.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 37(4):65:1–65:12, 2018.

Shengjie Zhu and Xiaoming Liu. PMatch: Paired masked image modeling for dense geometric matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21909–21918, 2023.

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

APPENDIX

A IMPLEMENTATION DETAILS

Architecture. The encoder follows VGGT-Ω (Wang et al., 2026a): its blocks alternate attention within each view and attention across all views, and in a few of the cross-view blocks only the cameras and register tokens of each view attend across views. Images are tokenized by a 16×16 convolutional patch embedding. The ViT-S, ViT-B and ViT-L have widths 384, 768 and 1024 and 12, 12 and 24 block pairs. With their depth and camera heads, they have 0.1B, 0.3B and 0.9B parameters. The covisibility head has two cross-attention blocks, and the masked modeling decoder four blocks of width 384.

Covisibility labels. Pixel centers are reprojected in 32-bit floating point and rounded to the nearest target pixel without a half pixel correction, which can shift a label by one pixel at depth edges.

Optimization. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 3×10^{-4}, weight decay of 0.05, gradient clipping at 1.0 and a cosine schedule with $1{,}000$ warm-up steps, restarted for each stage. A step processes $1{,}536$ images, and Figure 5 shows how the controlled setting divides its 120k steps.

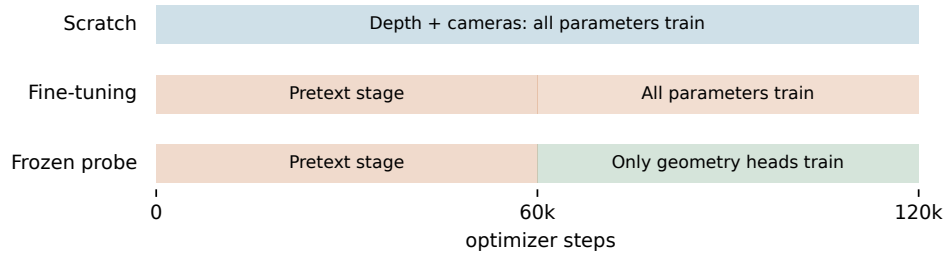


Figure 5: **Training protocols of the controlled setting.** Training from scratch spends all 120k steps on depth and cameras. Pre-trained models spend 60k steps on the pretext task and then train depth and camera heads for 60k steps, with the encoder either fine-tuned or frozen.

Scaled models. The eighteen fine-tuning datasets are ScanNet (Dai et al., 2017a), ARKitScenes (Baruch et al., 2021), WildRGB-D (Xia et al., 2024), TartanAir (Wang et al., 2020), Hypersim (Roberts et al., 2021), BundleFusion (Dai et al., 2017b), Aria Synthetic Environments (Avetisyan et al., 2024), Aria Digital Twin (Pan et al., 2023), Virtual KITTI 2 (Cabon et al., 2020), BEDLAM (Black et al., 2023), BlendedMVS (Yao et al., 2020), MegaDepth (Li & Snavely, 2018), Argoverse 2 (Wilson et al., 2021), DDAD (Guizilini et al., 2020), KITTI-360 (Liao et al., 2023), Waymo Open (Sun et al., 2020), DL3DV (Ling et al., 2024) and CO3Dv2 (Reizenstein et al., 2021). The ViT-B and ViT-L encoders are pre-trained for 500k steps, with covisibility labels from their ground-truth geometry. Fine-tuning then trains on all datasets at 192×256 for 50k steps, and continues at 384×512. The training time of VGGT-Ω is not published. We estimate its multi-view training at about 24k H100-hours from its recipe (240k iterations on 128 H100 GPUs), by counting its training FLOPs at 35% GPU utilization.

B EVALUATION PROTOCOL

Benchmarks. ScanNet uses the 120 held-out scans with eight clips each (960 clips), sampled at frame strides between 3 and 10. 7-Scenes uses its 46 sequences (368 clips, strides 3 to 10), DL3DV uses 456 clips from 57 held-out sequences with cameras from COLMAP (strides 2 to 5), and KITTI uses 1,128 clips from 141 driving sequences (strides 1 to 3) with sparse LiDAR depth. For the comparison with released models, each model receives the same clips to its training resolution, 384×512 for ours and VGGT-Ω and 392×518 for π^3 and MoGe-2.

Metrics. AbsRel follows Equation 4. On KITTI it is computed on the LiDAR pixels only. A benchmark score is the mean of the per-clip scores. Cameras are expressed relative to the first view. The rotation error is the geodesic angle averaged over all views of the clip, and the translation direction error is the angle between the predicted and ground truth translations, averaged over the views with a non-zero ground truth translation. On CO3Dv2 and RealEstate10K, the pose AUC at 30° follows VGGT (Wang et al., 2025a): for every pair of views, the larger of the relative rotation and translation direction errors is thresholded, and the AUC integrates the resulting accuracy up to 30○.

Correspondence. In Table 4, 100 random pixels of the first view of each 8-view clip are matched in the seven other views by the most similar patch feature, and only covisible queries are scored. Features are read at the best of four evenly spaced layers of each encoder. The linear probe of MuM (Nordström et al., 2026) trains one linear map on these features with ScanNet training scans.

Significance. Confidence intervals and paired differences resample scenes, that is scans or sequences, with replacement and keep all clips of each scene, with 1,000 resamples. A paired difference evaluates both models on the same resampled scenes. These intervals do not include the variation between training runs, which Appendix E reports.

C COVISIBILITY LABELS

Label stability. For a source pixel whose reprojection lands on the same target pixel, let $a = |z - d| - t$ be the margin of the depth test of Equation 2, with $d = D_j(\bar{q})$ and $t = \max(\tau_r d n \tau_a)$. Errors in the depth or cameras change this margin by at most

$$|\hat{a} - a| \leq |\hat{z} - z| + |\hat{d} - d| + |\hat{t} - t|, \tag{5}$$

so the label is unchanged whenever this bound is smaller than $|a|$. Larger errors can move the reprojected pixel across a depth edge or out of the image, and then change the label.

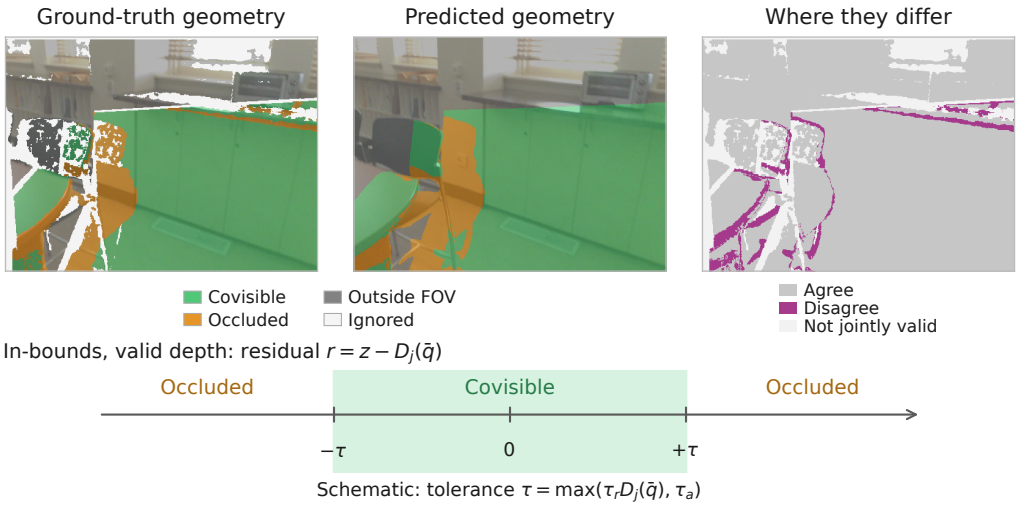


Figure 6: **Labels from ground-truth and predicted geometry on one held-out ScanNet pair.** Left, labels computed from ground-truth depth and cameras; middle, from predicted depth and cameras, with the class colours of Figure 2; right, the pixels where the two disagree. The bottom row sketches the depth test of Equation 2: a residual outside the tolerance, of either sign, gives class 1. The pair was chosen for its mix of classes.

D ADDITIONAL RESULTS

D.1 FINE-TUNING UNDER A MATCHED BUDGET

	ScanNet			7-Scenes	DL3DV	KITTI
pre-training	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓
from scratch						
none, 60k steps	0.0960	8.60	37.9	0.1656	36.34	0.2565
none, 120k steps	0.0863	7.70	37.2	0.1632	35.26	0.2543
released encoder, frozen						
DINOv3 ViT-L/16	0.0323	3.50	15.6	0.0842	25.89	0.1986
MuM ViT-L/16	**0.0294**	3.03	12.5	**0.0792**	**19.94**	**0.1234**
ours, pre-trained on ScanNet						
covisibility	0.0385	3.13	10.7	0.1112	22.15	0.2167
covisibility + masked modeling	0.0316	**2.73**	**9.3**	0.0946	20.38	0.2161

Table 5: **Fine-tuning under a matched budget** (ViT-S, ScanNet). Every model trains for 120k steps except the first. Training from scratch spends all of them on depth and cameras; our models spend 60k on pre-training and 60k on fine-tuning. The released encoders are frozen and feed the same ViT-S network, trained for 120k steps; they were pre-trained on far more data, so their rows are a reference rather than a same-data comparison. Rotation and translation-direction errors are in degrees. Best per column in **bold**.

Table 5 gives the fine-tuning results of the controlled setting. Doubling the schedule of training from scratch improves ScanNet AbsRel only from 0.0960 to 0.0863, whereas every pre-trained model reaches below 0.04. The frozen DINOv3 and MuM encoders were pre-trained on far more data and serve as a reference: the joint ViT-S matches DINOv3 on ScanNet depth and is significantly more accurate on ScanNet cameras and on DL3DV, while DINOv3 is more accurate on 7-Scenes and KITTI depth.

D.2 FROZEN ENCODERS ON DEPTH AND CAMERAS

	ScanNet			7-Scenes	DL3DV	KITTI	RE10K
frozen encoder	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓	AUC@30 ↑
released, with a ViT-S multi-view network trained on top							
DINOv3 ViT-L/16	0.0323	3.50	15.6	0.0842	25.89	0.1986	45.0
MuM ViT-L/16	**0.0294**	3.03	12.5	0.0792	19.94	**0.1234**	51.5
released, heads only							
Muskie ViT-L	0.0569	10.31	25.7	0.0951	25.51	0.1629	45.7
ours, ViT-B, heads only							
masked modeling	0.0480	3.50	11.2	0.0893	16.98	0.2236	60.5
covisibility	0.0390	3.07	10.5	0.0854	13.86	0.1730	63.2
covisibility + masked modeling	0.0396	**3.00**	**9.6**	**0.0760**	**10.69**	0.1679	**68.3**

Table 6: **Frozen encoders on depth and cameras.** Each encoder is frozen. DINOv3 and MuM encode each view separately, so a ViT-S multi-view network is trained on top of them together with the depth and camera heads, for 120k steps on ScanNet. Muskie and our ViT-B encoders are multi-view, so only the heads are trained, for 60k steps. Our three ViT-B encoders are pre-trained on the same data for 500k steps. Rotation and translation-direction errors are in degrees. Best per column in **bold**.

Table 6 freezes each encoder and trains depth and camera heads on ScanNet. DINOv3 and MuM encode each view separately, so a ViT-S multi-view network is trained on top of their frozen features together with the heads, as in Table 5, whereas Muskie and our ViT-B encoders are multi-view and

only the heads are trained. Our covisibility ViT-B is significantly more accurate than Muskie on every column except KITTI depth. Muskie was pre-trained with at most eight views, fewer than the 24 used here, but it also trails our encoder on ScanNet under the eight-view protocol of Table 4. DINOv3 and MuM, which benefit from the network trained on top of them, are more accurate on ScanNet depth, and MuM also on KITTI depth, whereas our encoders are significantly more accurate in ScanNet translation direction, on DL3DV and on RealEstate10K. With the same data and the same 500k steps, the covisibility ViT-B is significantly more accurate than the masked modeling ViT-B on every column except 7-Scenes depth, where they are level, and the joint encoder improves on the covisibility ViT-B on five of the seven columns.

D.3 Robustness to noise from VGGT-Ω's labels

	ScanNet			7-Scenes	DL3DV	KITTI
pre-training	AbsRel ↓	rot. ↓	trans. ↓	AbsRel ↓	rot. ↓	AbsRel ↓
none (from scratch, 120k steps)	0.0863	7.70	37.2	0.1632	35.26	0.2543
covisibility, ground-truth labels	**0.0385**	**3.13**	**10.7**	**0.1112**	**22.15**	**0.2167**
labels from VGGT-Ω						
regression of VGGT-Ω's depth and cameras	0.0822	6.63	32.3	0.1580	33.97	0.2459
covisibility, labels from VGGT-Ω	0.0405	3.31	11.2	0.1225	22.74	0.2427

Table 7: **Labels from VGGT-Ω** (ViT-S, ScanNet, fine-tuned). Each pre-trained model spends 60k steps on its pretext target and 60k on depth and cameras. The last two rows use the predictions of VGGT-Ω either as a regression target or to compute covisibility labels. Best per column in **bold**.

Table 2 computes the covisibility from a weaker model. In Table 7, they come from the predictions of VGGT-Ω. These labels come within 0.002 ScanNet AbsRel of ground-truth labels. Used instead as a dense regression target during pre-training, the same predictions give 0.082: the discrete target transfers much better than the geometry it is computed from.

D.4 PRINCIPAL COMPONENTS OF FROZEN FEATURES

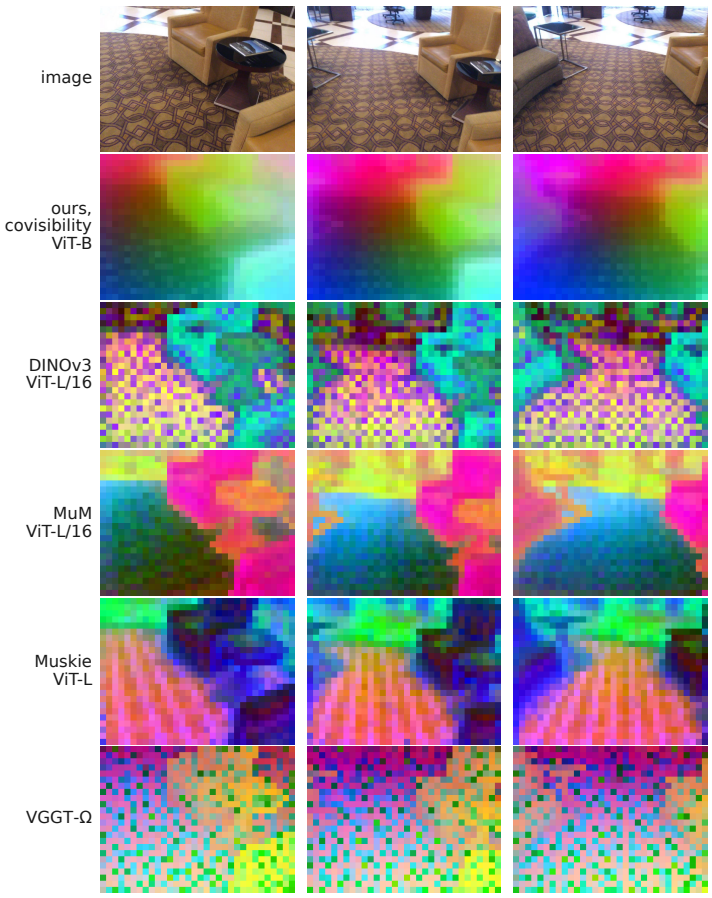


Figure 7: **Principal components of frozen patch features** on the three views of Figure 3 (three of the eight views of a held-out ScanNet clip), given to every encoder at 384×512 pixels (24×32 patches per view; ours was pre-trained at 192×256). For each encoder, one PCA basis is fitted over all eight views and its first three components are shown as colours, so colours can be compared across the views of a row but not between rows. Each encoder is read at the layer where its features match best in Table 4; ours is the covisibility ViT-B of Table 6.

The first two components of our encoder vary smoothly across the scene: a linear function of the 3D position of each patch explains 90% and 89% of their variance, against at most 71% for DINOv3, MuM and Muskie, which instead separate regions such as the carpet and the armchairs. Image position explains much of it too (88% and 76%), since the clip covers a small part of the room. VGGT-Ω's second component also follows 3D position (83%), but its features at this layer are dominated by high frequency patterns.

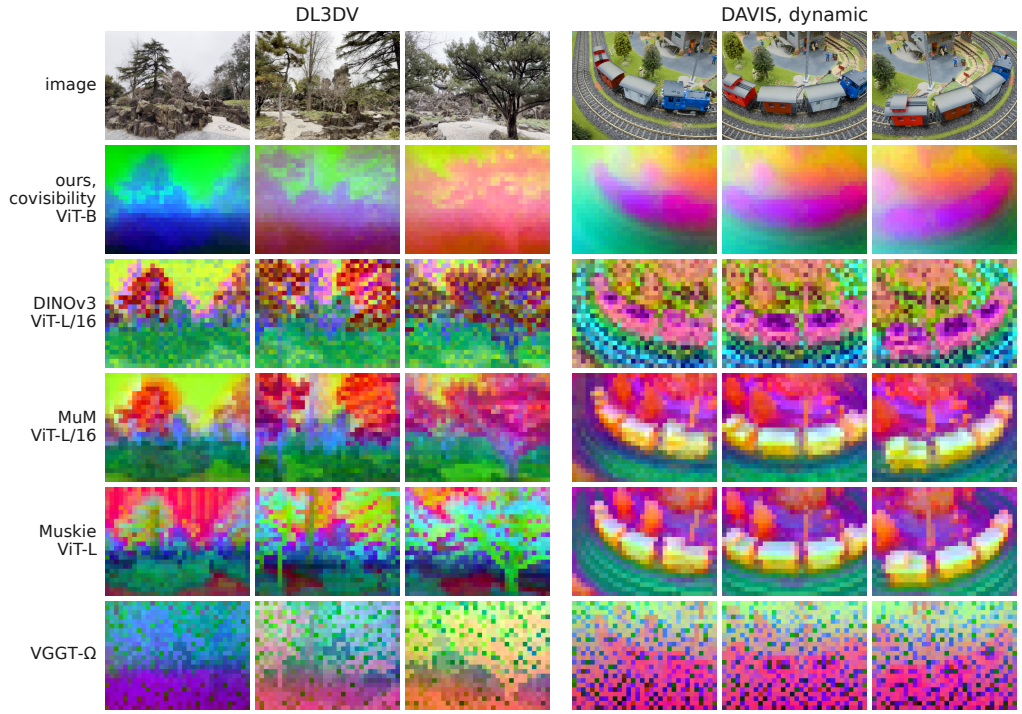


Figure 8: **Principal components of frozen patch features on two outdoor clips**, in the format of Figure 7: a held-out DL3DV scene and a DAVIS video of a moving model train, both outside the training data of our models. Each clip has eight frames at 384×512 pixels, of which three are shown.

Figure 8 repeats this analysis on two outdoor clips. The leading components of our encoder again vary smoothly across each view rather than separating objects, whereas DINOv3, MuM and Muskie separate regions such as trees, ground or the train, and the features of VGGT-Ω at this layer remain dominated by high-frequency patterns. On the DL3DV clip, whose views are far apart, the colors of our encoder change more between views than on ScanNet. These clips have no ground-truth depth, so we cannot measure how much of this variation follows 3D position rather than image position.

E Training-seed variability

	ScanNet		7-Scenes	DL3DV	KITTI
recipe and run	AbsRel ↓	rot. ↓	AbsRel ↓	rot. ↓	AbsRel ↓
from scratch, 120k steps					
first run	0.0863	7.70	0.1632	35.26	0.2543
second run	0.0837	6.79	0.1612	34.87	0.2524
covisibility, encoder frozen					
first run	0.0586	4.34	0.1231	25.22	0.2560
second run	0.0601	4.93	0.1344	26.34	0.2103
covisibility, encoder fine-tuned					
first run	0.0385	3.13	0.1112	22.15	0.2167
second run	0.0389	3.24	0.1165	23.73	0.2022

Table 8: **Two training runs per recipe** (ViT-S, ScanNet), with different random seeds. Rotation errors are in degrees.

Table 8 repeats three recipes of the ablation setting with a second seed. The two runs differ by 1 to 3% in ScanNet AbsRel, less than the differences between the training strategies compared in Table 1. The scene-bootstrap intervals do not include this variation.