On Geometric Understanding and Learned Priors in Feed-forward 3D Reconstruction Models

Jelena Bratulić*[1], Sudhanshu Mittal[1], Thomas Brox[1], and Christian Rupprecht[2]

[1]University of Freiburg, Germany [2]University of Oxford, UK

Abstract. Feed-forward 3D reconstruction models such as DUSt3R, VGGT, and Depth Anything 3 (DA3) are transformer–based foundation models that infer camera geometry and dense scene structure in a single forward pass. Trained at scale in a supervised fashion, they raise a central question: do these models build upon geometric principles akin to traditional multi–view pipelines, or do they primarily rely on learned priors arising from the large–scale training setup? We find that epipolar geometry emerges within the intermediate layers of all three models and is causally linked to correspondence patterns in attention heads. To study this, we perform a systematic analysis of their internal representations across three real–world datasets and a controlled synthetic dataset. We quantify geometric understanding by probing intermediate features, analyzing attention patterns to identify correspondence matching patterns, and performing targeted interventions at the attention level. Further, we assess the role of learned priors by applying challenging input–level perturbations, such as occlusions, scene ambiguities, and varying camera configurations, and compare them against classical multi–stage reconstruction pipelines.

1 Introduction

Estimating 3D geometry from a set of 2D images has long been a fundamental challenge in computer vision. Traditional approaches rely on geometry–constrained pipelines that require a multi–step, sequential process: feature extraction, matching, and refinement, followed by camera pose estimation and 3D point cloud reconstruction. These approaches, although mathematically precise, struggle with ambiguities and lack robustness to scene dynamics and partial visibility, and can be computationally expensive due to runtime optimization.

Recent deep learning approaches such as DUSt3R [41], MASt3R [21], VGGT [39], and DA3 [22], have shown that 3D reconstruction can be achieved in a single unified feed–forward step, replacing the traditional multi–stage pipeline. These methods predict camera poses and dense 3D point maps from image sequences using a functional mapping learned from large–scale labeled datasets. As all optimization is shifted to training, they run much faster than classical frameworks,

* The work was done during a PhD visit to the University of Oxford within the ELLIS PhD program. Correspondence to bratulic@cs.uni-freiburg.de

and the priors learned from the data make them significantly more robust to ambiguities and input variations. However, such learned functional mappings are prone to fail in out–of–distribution cases. It is worth noting that the laws of 3D geometry are universal and should not depend on the distribution of the training data. Therefore, the question arises whether these models, while learning from data, internally learn any geometric reasoning mechanisms similar to classical multi–view pipelines, or do their capabilities primarily arise from learned priors? If a geometric mechanism is indeed learned, it would guarantee a similar level of robustness to out–of–distribution cases as with traditional pipelines, while leveraging the complex learned priors in in–distribution scenarios.

In this work, we examine the underlying mechanisms of three representative feed–forward 3D reconstruction models: DUSt3R, VGGT, and Depth Anything 3 (DA3). We aim to mechanistically understand how these models compute their predictions and to determine whether their performance arises from internalized geometric reasoning or predominantly from learned priors. In particular, we investigate whether their representations encode fundamental principles of multi–view geometry, such as correspondences and epipolar geometry.

To this end, we design a controlled study using different object-centric and scenery datasets, and a controlled synthetic dataset with ShapeNet [8] objects under various conditions, with complete scene information and all 3D attributes being available. Our systematic analysis adopts two perspectives: (1) a **geometric** perspective, where we probe the internal representations and perform interventions to assess the causal effect of these representations on the fidelity of the network representation in terms of epipolar geometry, (2) a **data** perspective, where we test the model's robustness to interventions on input images to study the (typically positive) influence of learned priors from data and training.

The studied models are jointly trained to predict dense scene structure and camera geometry directly from a set of images in a single forward pass. Although their architectural designs differ, none is explicitly trained to estimate classical geometric entities such as the fundamental matrix. To evaluate their use of geometric principles, we focus on a core property: epipolar geometry. The fundamental matrix $\mathbf{F} \in \mathbb{R}^{3\times3}$ provides a natural lens for this analysis, as it formalizes the epipolar geometry between different camera views and corresponding points observed in them. We therefore use $\mathbf{F}$ prediction as a proxy to assess how much geometric information is encoded in the model's internal representations.

Contributions. Our extensive study across real-world and synthetic datasets reveals several key insights. (1) We show that the intermediate representations in DUSt3R, VGGT, and DA3 encode epipolar geometry. (2) We identify a geometric phase transition in the middle layers, where fundamental matrix recovery coincides with the emergence of point correspondences in the attention space. (3) Through targeted attention-knockout interventions, we establish a causal link between correspondence formation and the encoding of epipolar geometry. (4) From a data perspective, we show, using VGGT, that such models remain robust to moderate disturbances, often outperforming classical pipelines. Under occlu-

sions, they infer correspondences despite missing evidence, revealing reliance on learned priors and a trade-off between robustness and geometric consistency.

2 Related work

Classical 3D reconstruction and multi-view geometry. Classical structure-from-motion (SfM) pipelines [1, 16, 34, 36] comprise several stages: feature detection, correspondence matching, camera pose estimation, and bundle adjustment, and rely on hand-crafted features such as SIFT [24]. A central component of these pipelines is estimating the fundamental matrix $\mathbf{F}$, which encodes the epipolar geometry between two views [16]. Minimal solvers such as the five-point [28] and eight-point [17] algorithms recover $\mathbf{F}$ from sparse correspondences, usually combined with robust estimators like RANSAC [13] to handle noisy correspondences and outliers. Recent learned matching methods such as SuperGlue [32], LightGlue [23], and LoFTR [38] significantly improve correspondence matching. However, geometric relations such as the fundamental matrix are still estimated using classical algorithms. In this work, we use epipolar geometry as a principled lens to study whether modern feed-forward reconstruction models internally encode such geometric structure.

Feed-forward reconstruction models. Recent deep learning approaches challenge the traditional multi-stage paradigm for 3D reconstruction. DUSt3R [41] introduced a feed-forward formulation for dense 3D reconstruction from image pairs without requiring known camera poses. MASt3R [21] improves robustness to large viewpoint changes by augmenting the architecture with a dense local feature head, while preserving DUSt3R's robustness. VGGSfM [40] further demonstrated that fully differentiable end-to-end reconstruction pipelines can outperform classical SfM methods. More recent foundation models, such as VGGT [39] and Depth Anything 3 (DA3) [22], directly infer depth and scene geometry from image collections in a single feed-forward pass, achieving state-of-the-art performance while significantly reducing inference time. Follow-up work further accelerates inference [35,42] or extends these architectures to downstream tasks such as SLAM and real-time reconstruction [26,47]. Despite these advances, the internal mechanisms by which these models represent and reason about geometry remain poorly understood, a gap our work aims to address.

Interpretability and causal analysis. Understanding neural network mechanisms has become increasingly important, with much of the work centered on LLMs [12, 29]. A common approach is the use of probing classifiers [2, 7, 20] to test whether information is decodable from intermediate representations. For computer vision tasks, Network Dissection and related work [5, 6, 45] quantify the CNNs' interpretability at individual units, whereas Chefer et al. [9] showed that naive attention visualization can be misleading for transformers. Different interventions, such as activation patching or causal mediation analysis [18, 44], provide more substantial evidence by measuring the effects of modified activations within a model. However, little work has been conducted on interpretability in geometric understanding for 3D vision models. El Banani et al. [11] do not

analyze feed-forward 3D reconstruction models, but probe 2D vision foundation models with a DPT head for depth, normals, and correspondences. Chen et al. [10] assess 3D awareness and feature quality through a downstream novel-view synthesis task, while An et al. [3] show that cross-attention layers can serve as a zero-shot matcher. While these works show that visual or 3D foundation models can exhibit correspondence-like behavior or encode 3D-aware representations, they primarily provide observational evidence at the output. Recently, Stary et al. [37] analyzed a variant of DUSt3R and studied the role of individual layers and correspondences in multi-view transformers. Maggio et al. [25] analyze the VGGT's layers for image retrieval verification, while Han et al. [15] study the emergent outlier detection ability in VGGT. We study where and how epipolar geometry emerges *inside* models, providing a mechanistic and causal analysis of how correspondence formation supports this geometry, rather than merely showing that correspondences exist. We further analyze when models rely on geometric evidence versus learned priors under perturbations.

3 Background

3.1 Epipolar Geometry

Here we provide a brief overview of epipolar geometry, but an extensive analysis can be found in [16]. Epipolar geometry describes the projective relationship between two cameras viewing the same 3D point. If a 3D point $\mathbf{X} \in \mathbb{R}^3$ is observed by two pinhole cameras, its image projections are

$$\mathbf{x}_1 = \mathbf{K}_1[\,\mathbf{I} \mid \mathbf{0}\,]\mathbf{X}, \quad \mathbf{x}_2 = \mathbf{K}_2[\,\mathbf{R} \mid \mathbf{t}\,]\mathbf{X}, \tag{1}$$

where $\mathbf{K}_i$ are the camera intrinsics and $(\mathbf{R}, \mathbf{t})$ is the relative pose. A ray from the first camera center through $\mathbf{x}_1$ defines an *epipolar plane* with the two camera centers. Its intersection with the second image plane gives the *epipolar line* ℓ_2, on which the corresponding point $\mathbf{x}_2$ must lie. All epipolar lines intersect at the epipole. This geometric relationship is expressed by the *epipolar constraint*:

$$\mathbf{x}_2^\top \mathbf{E}\, \mathbf{x}_1 = 0, \quad \mathbf{E} = [\mathbf{t}]_\times \mathbf{R}, \tag{2}$$

where $\mathbf{E}$ is the *essential matrix* and $[\mathbf{t}]_\times$ is the skew-symmetric matrix representing the cross product with $\mathbf{t}$. For uncalibrated cameras, the relation becomes

$$\mathbf{x}_2^\top \mathbf{F}\, \mathbf{x}_1 = 0, \quad \mathbf{F} = \mathbf{K}_2^{-\top} \mathbf{E} \mathbf{K}_1^{-1}, \tag{3}$$

where $\mathbf{F}$ is the *fundamental matrix*.

Both the essential matrix $\mathbf{E}$ and fundamental matrix $\mathbf{F}$ have rank 2, as $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$ inherits the rank deficiency from the skew-symmetric matrix $[\mathbf{t}]_\times$. This constraint reflects the fact that all epipolar lines intersect at the epipole. When estimating $\mathbf{F}$ from point correspondences, the rank-2 constraint is typically enforced by projecting the unconstrained solution onto the manifold of

rank-2 matrices. Given an initial estimate $\hat{\mathbf{F}}$, we compute its singular value decomposition $\hat{\mathbf{F}} = \mathbf{U}\text{diag}(\sigma_1, \sigma_2, \sigma_3)\mathbf{V}^\top$ where $\sigma_1 \geq \sigma_2 \geq \sigma_3$, and set the smallest singular value to zero:

$$\mathbf{F} = \mathbf{U}\,\text{diag}(\sigma_1, \sigma_2, 0)\,\mathbf{V}^\top. \tag{4}$$

This produces the closest rank-2 approximation in the Frobenius norm.

Evaluating Epipolar Geometry. We can assess the quality of an estimated fundamental matrix using two metrics that quantify deviations from epipolar consistency.

The algebraic error

$$e_{\text{alg}} = \mathbf{x}_2^\top \mathbf{F}\,\mathbf{x}_1 \tag{5}$$

measures the constraint satisfaction (which should be 0), but has no geometric meaning directly and is scale-dependent. More geometrically meaningful measures are the Sampson error and the smallest singular value of $\mathbf{F}$. The *Sampson error* [31] represents a first-order approximation of the true geometric error:

$$d_{\text{Sampson}} = \frac{(\mathbf{x}_2^\top \mathbf{F}\,\mathbf{x}_1)^2}{\|(\mathbf{F}\mathbf{x}_1)_{1:2}\|^2 + \|(\mathbf{F}^\top \mathbf{x}_2)_{1:2}\|^2} \tag{6}$$

Sampson error approximates the reprojection error, avoiding iterative optimization. In practice, well-calibrated systems typically yield Sampson errors below 1–2 pixels, though acceptable values depend on image resolution, focal length, and the application's precision requirements. The *smallest singular value* of $\mathbf{F}$ must be zero to satisfy the rank-2 constraint of the fundamental matrix. In practice, the smallest singular value that is less than 10^{-3} of the largest singular value is generally considered acceptable.

3.2 Feed-forward 3D Reconstruction Models

Here, we analyze three representative feed-forward 3D reconstruction models covering different architectural paradigms.

DUSt3R [41] introduced the paradigm of feed-forward dense 3D reconstruction from arbitrary image pairs without requiring prior camera calibration or pose information. The model predicts pointmaps expressed in the reference frame of one view using an asymmetric transformer encoder–decoder architecture. For multi-view inputs, pairwise predictions are aligned into a common coordinate frame through a global alignment step performed at test time. This formulation unifies depth, pose, and dense reconstruction within a single model and laid the foundation for later feed-forward geometry models such as VGGT and DA3.

Visual Geometry Grounded Transformer (VGGT) [39] is a transformer-based model that jointly predicts camera intrinsics, extrinsics, depth maps, pointmaps, and 3D point trajectories from multiple views. It builds on a DINOv2 [30] backbone with alternating frame-wise and global self-attention layers to aggregate information across views. VGGT is trained on a large mixture of

publicly available 3D datasets using direct supervision for camera pose estimation, depth prediction, 3D point cloud reconstruction, and 3D point tracking.

Depth Anything 3 [22] predicts scene geometry from arbitrary visual inputs using a transformer backbone trained in a teacher–student framework. Unlike the multi-task supervision used in VGGT, DA3 formulates the problem as the joint prediction of depth maps and ray maps using a unified prediction head. The model can optionally condition on input camera poses at inference and achieves state-of-the-art performance on camera pose estimation and geometric accuracy benchmarks, surpassing prior models, including VGGT.

3.3 Dataset Design

We conduct our study on synthetic and real-world datasets. Since all three models are trained on large, diverse datasets, ensuring meaningful out-of-distribution evaluation and analysis of certain scenarios requires careful dataset selection. Thus, we construct a custom ShapeNet-based synthetic dataset that provides complete geometric ground truth and allows systematic scene manipulation. Here we briefly describe the three real-world and custom synthetic datasets, while more details and visual examples are provided in the supplementary.

Real-world datasets. *DTU MVS* [19] is a controlled dataset of 124 object-centric scenes photographed from 49 fixed positions under structured-light ground truth. *ETH3D* [33] is a precise, high-resolution, multi-view dataset covering diverse indoor and outdoor scenes. *MipNeRF360* [4] comprises nine unbounded outdoor and indoor scenes captured with a consumer camera and reconstructed via COLMAP. Camera poses and correspondences are derived from COLMAP sparse reconstructions, while for DTU, we use depth maps from MVSNet [43]. To ensure diversity in viewpoint separation while improving control over the data distribution, we bin image pairs by the true 3D angular distance between the cameras and sample from the bins for MipNeRF360 and DTU. For ETH3D, we include all pairs since its scenes, although diverse, contain only a relatively small number of images. We provide additional analysis using the semantic correspondences from the SPair71k dataset [27] in the supplementary.

Custom ShapeNet dataset. To model diverse scenarios, we generated a controlled synthetic dataset from ShapeNet [8] assets rendered in Blender, which enables precise control over scene properties and complete geometric ground truth (camera parameters, depth, correspondences, fundamental and essential matrices) for precise analysis. We selected two types of objects: 10 different categories of *unique* assets with distinctive features and 5 different categories of *symmetric* assets (e.g., bottles, vases) that introduce correspondence ambiguity. We use focal lengths ranging from 24 mm to 100 mm with a standard 36 mm camera sensor, and distinguish between 4 camera configurations: a stereo setup with parallel image planes and a horizontal baseline, and three different non-parallel image planes with different angular baselines, categorized as small (10–25 degrees), medium (45–75), and large (90–120). For robustness experiments, we created ambiguous scenes with repetitive objects.

4 Geometric interpretability

DUSt3R, **VGGT**, and Depth Anything 3 (**DA3**) are trained to predict camera geometry and dense 3D structure through supervised objectives specific to their respective formulations. To evaluate whether these models develop geometric reasoning beyond their explicit training targets, we examine their ability to represent the fundamental matrix via probing. Since the fundamental matrix is never directly supervised, this independent probing avoids confounding effects from task-specific heuristics and strong supervision on camera pose and intrinsics, although F can be derived algebraically from pose and intrinsics. Further, layer-wise probing also allows us to localize where this geometric information is encoded in the network layers, providing deeper insight into how geometric structure develops across the network.

4.1 Do these models encode geometry? If so, where?

We probe intermediate layers to determine where, and to what extent, information about the epipolar geometry can be recovered from these representations.

Experiment setup. We train simple two-layer MLP probes on intermediate representations at each layer to predict the fundamental matrix. Given the differences between the architectures, we use camera tokens from each layer for **VGGT** and **DA3**, and averaged features after each decoder block for **DUSt3R**. Each probe utilizes a hidden dimension of 512 and is optimized for 30 epochs using the squared Sampson distance as the training loss on ground-truth correspondences, without directly enforcing the rank-2 constraint. We train probes on each dataset separately, and report the root Sampson error in pixel space

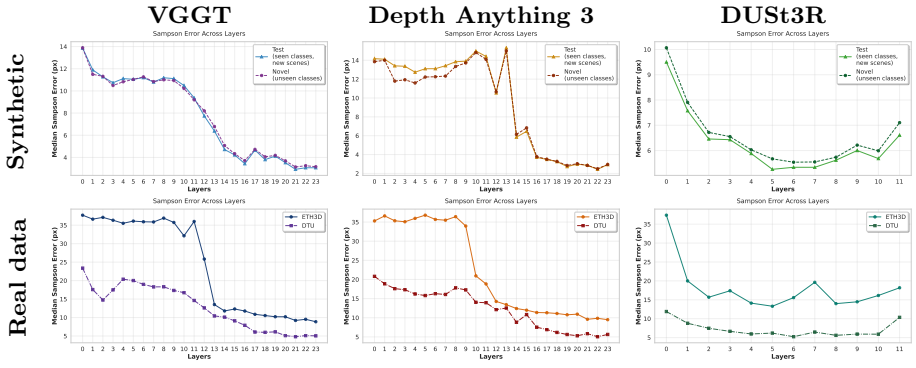

Fig. 1: Probing internal representations for fundamental matrix approximation. We successfully recover the fundamental matrix from all three models on both synthetic and real-world datasets. However, the emergence trends differ across models: models with camera-tokens (**VGGT** and **DA3**) exhibit a similar pattern, with clear change in the middle layers and strong performance in the final layers. **DUSt3R** shows strong performance already in the early layers, with slight degradation towards the end.

across unseen data splits. We further perform a similar analysis using image triplets and probing for all available fundamental matrices. Additional details and results are provided in the supplementary.

Results. Figure 1 shows that the fundamental matrix can be reliably recovered from all three models (**VGGT**, **DA3**, **DUSt3R**) across different datasets (ShapeNet, ETH3D, DTU). The location and the trends of the emergence, however, differ. In **VGGT** and **DA3**, the ability to recover the fundamental matrix emerges roughly in the middle layers (layers 12–14). Performance continues to be refined through the final layers. The sudden improvement in error within a few layers suggests that the necessary information for establishing epipolar geometry is formed in these layers. The layers in **DA3** correspond to those immediately after global attention is introduced. As a second criterion, we also check the smallest singular value of the predicted $\mathbf{F}$. We observe that it attains its minimum within the same set of layers and is approximately 2.5e−4 smaller than the maximum value, further supporting the emergence of epipolar geometry within these layers (see supplementary).

DUSt3R, on the other hand, exhibits a different trend, in which the fundamental matrix emerges in the early decoder blocks and refines throughout the layers with slight divergence in the last layers. We speculate that this divergence arises due to independent target heads in the model.

Since all models encode the fundamental matrix in their representations, with the emergence followed by a sudden improvement over a few layers, this suggests that these layers serve a specific purpose. An obvious candidate is that the network's attention maps capture a correspondence structure across views.

Epipolar geometry emerges in the internal representations of feed-forward multi-view transformers. In camera-token based architectures, it appears in the middle layers and is progressively refined towards the final layers.

4.2 How do attention maps encode correspondences?

Probing reveals the emergence of learned epipolar geometry in the intermediate layers. However, we do not yet know how the model recovers this information. To this end, we aim to discover the underlying correspondence-matching patterns in the attention maps of those layers across two views. Following [3], we hypothesize that these correspondences are computed by global (cross) attention layers.

Experiment setup. We analyze the query-key (QK) attention space to identify correspondence matching. For each patch token in the source image, we locate the patch in the target image that receives the highest attention and calculate accuracy based on whether this target patch corresponds to the true matching patch. When a source patch maps to multiple target patches due to patch discretizations, we count a hit if any of the correct positions receive the maximum attention. We evaluate all unique patch correspondence pairs and report the average correspondence accuracy over our controlled ShapeNet's test set of 5 scenes from different object categories, taken at a 50 mm focal length, averaged

across sampling modes that cover different camera-pair configurations. We analyze all available cross- and global-attention layers in the model: 24 for **VGGT**, 8 for **DA3** (starting at layer 9), and 12 for **DUSt3R**. We compare the forward direction (view 1→view 2) and reverse (view 2→view 1). Further, we analyze feature-space similarity to assess whether correspondence-matching information is retained in the representation outside of attention space. Details, real-world results, and feature-similarity results are provided in the supplementary.

Results. Figure 2 shows strong correspondence matching activity in multiple layers and heads across all three models. In **VGGT**, strong correspondence matching across multiple attention heads is observed in the middle layers, roughly between layers 10 and 16. This behavior begins around layer 10 and precedes the sharp improvement in fundamental matrix recovery at layers 13 and 14 (see Fig. 1). This suggests a connection between correspondence estimation and the emergence of epipolar geometry within the network. We observe a similar trend for **DA3**, which, despite having layer 9 as the first global-attention layer, has the strongest correspondence matching across the middle layers of the complete architecture. **DUSt3R**, however, exhibits strong correspondence-matching performance from the early decoder layers, while performance continues to improve in the middle layers. This is again well aligned with the probing results in Fig. 1, where **DUSt3R** exhibited decent results already in the early layers.

Further, both **VGGT** and **DA3** have symmetrical patterns, meaning that the same heads and layers perform correspondence matching equally well in both forward (F 1→2) and reverse (R 2→1) directions. On the other hand, **DUSt3R**'s activation patterns are asymmetric, with the reverse direction exhibiting stronger correspondence-matching performance and a greater number of active heads. We

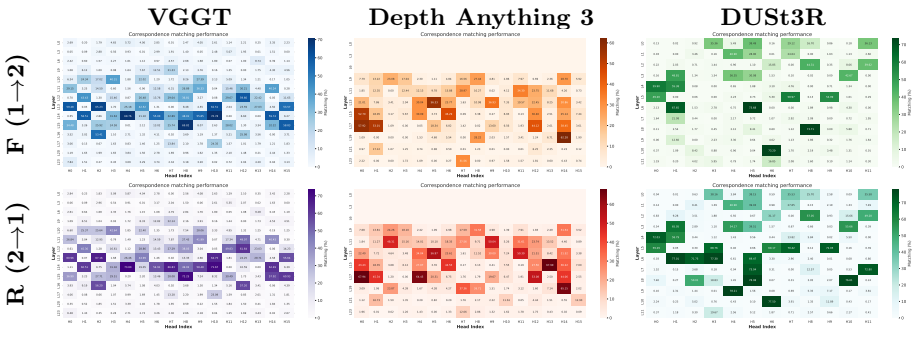

Fig. 2: Correspondence matching per attention head. We compare correspondence matching performance using the QK attention space for forward (view 1→view 2) and reverse (view 2→view 1) directions. Strong matching performance is established in the layers, which immediately precede the transition observed in Fig. 1. Middle layers are active for both **VGGT** and **DA3**, whereas **DUSt3R** exhibits correspondence matching already in the first block and asymmetrically active heads in different directions, stemming from its asymmetric decoders.

believe the difference stems from **VGGT** and **DA3** using global attention layers, whereas **DUSt3R** has asymmetric cross-attention decoders.

We also analyzed QK-space patterns across other layers to identify interpretable behavior. We observed that early layers in **VGGT** exhibit semantic alignment, where visually similar parts are matched (e.g. a chair leg in one view matches another chair leg in the second view), but not necessarily to the correct instance. More details are included in the supplementary.

Point correspondence matching appears in the mid-layer attention heads as a key mechanism, enabling geometric alignment across views.

4.3 Do these attention patterns play a causal role?

The previous observations establish a correlation between the emergence of epipolar geometry via fundamental matrix recovery and the correspondence-matching patterns in the same set of intermediate layers. To determine whether the attention patterns are truly a cause of the geometric encoding observed in Fig. 1, we perform targeted interventions on the attention heads that we hypothesize are responsible for the epipolar representation.

Experiment setup. We intervene on the QK attention space using attention knockout, a method similar to Geva et al. [14], where we zero out activations for specific layer-head combinations of the attention map for different levels of locality, ranging from zeroing out the whole attention head (*Whole head*), to the most local with just the corresponding patches (*Patch* in Fig. 3). We select interventions based on the high correspondence-matching performance of the layer-head combinations, along with two baselines: early/late layer-head combinations with lower matching performance. We do not intervene in the task-specific decoder heads, which are shared across clean and intervened runs; therefore, they cannot, by themselves, explain the intervention-induced gaps.

To perform the analysis, we require dense correspondence ground truth. There, we conduct the experiment on our controlled ShapeNet data and report median results across all sampling modes at a 50 mm focal length over the test set of 5 categories. Additional combinations, details on the intervention locality, and results are provided in the supplementary material. We report the difference in the Sampson distance between the baseline and the intervened model, where we approximate the fundamental matrix using the model's camera pose predictions. For **DUSt3R**, this requires post-processing steps that can result in invalid camera poses, which we consider a failure case.

Results. Figure 3 demonstrates strong causal effects for all three models: across all four locality levels (whole heads, image map, block, and patch), the same layer-head combinations reduce Sampson error, whereas other early/late layers do not. This rules out OOD disruption as the explanation and shows that correspondence-matching heads have functional information for geometric reasoning. In **VGGT** and **DA3**, disrupting the high-performing attention heads in the middle layers degrades the Sampson distance error to such a large degree

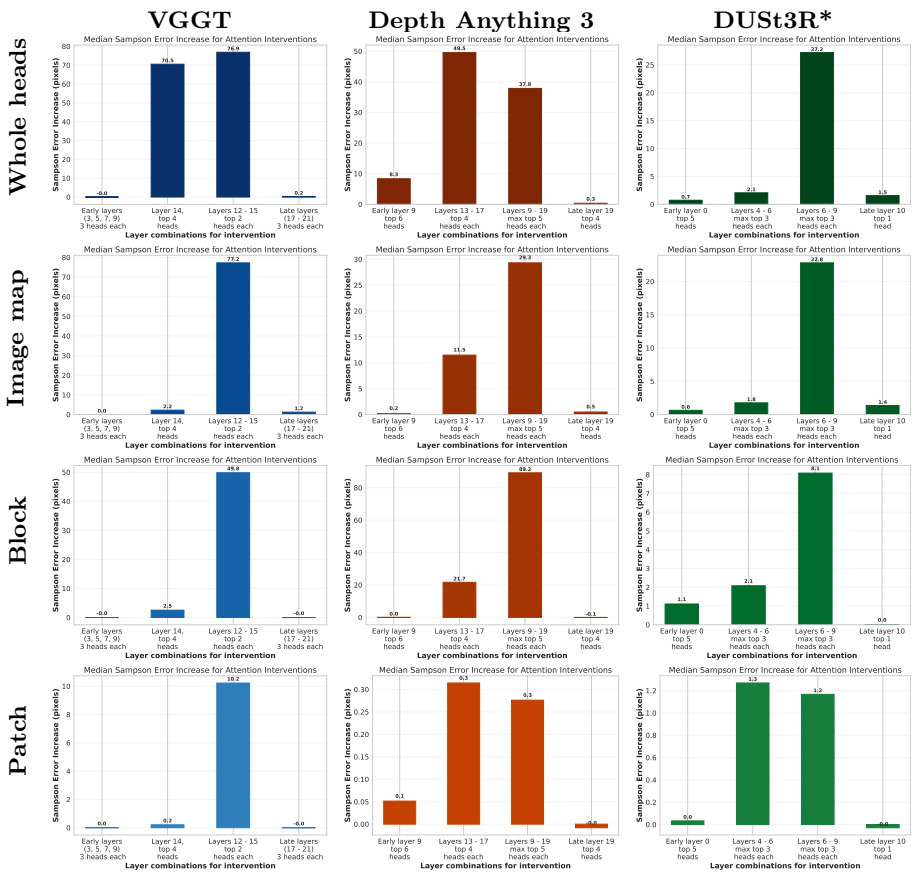

Fig. 3: Performance degradation under attention interventions on 4 localities. Disrupting random early or late layers has minimal impact on performance, whereas intervention on high correspondence-matching layers causes significant degradation. The degradation is consistent across all models and at all levels of locality. *For **DUSt3R**, we report results only for valid cases for Image map and Whole heads, since 72% of cases result in invalid predictions.

that it can be regarded as a complete failure. Intervening on random early or late layers, on the other hand, has no significant impact. There are two types of effective interventions: (1) multiple heads within a single layer, and (2) fewer heads distributed across multiple layers. These correspond to the middle bars in Fig. 3. For instance, in **VGGT**, disabling only four heads in a single layer already leads to substantial degradation.

However, **DUSt3R** exhibits a slightly different behavior from **VGGT** and **DA3**. Intervening in the early layers significantly affects performance, consistent with the strong correspondence-matching observed in those layers. Thus, while the layer location differs, the overall trend remains aligned with the other models.

Moreover, for the first two intervention types (*Whole heads* and *Image map*), where we zero out the entire head or the full attention map (keys set to 0), **DUSt3R** frequently produces invalid camera pose predictions. This suggests that its cross-attention mechanism is more sensitive to disruption than the global attention used in **VGGT** and **DA3** in this setting.

These results support our hypothesis: disabling attention heads that exhibit strong correspondence matching leads to significant degradation in the encoded epipolar geometry. However, this does not conclusively prove that correspondence matching alone is responsible for the emergence of epipolar geometry. These heads may also encode additional mechanisms that contribute to the effect. This becomes evident in more localized interventions: while the same degradation trend persists, but the error scale change. Even when exact patch-to-patch correspondences in the QK space are removed, nearby patches still provide sufficient signals for the model to partially recover geometric information.

Intervening on the attention patterns for point correspondences shows a causal relationship between those attention heads and the epipolar geometry representation by the model.

5 The role of learned priors in VGGT

While our geometric analysis includes three models, we focus on VGGT here to study how much robustness arises from learned priors rather than geometry. We consider VGGT a strong representative model for this study because it provides a cleaner experimental setup and operates without pose-conditioning at inference, unlike DA3. Additionally, there is no clear indication of a fundamental difference between VGGT and DA3; therefore, we focus on VGGT to get deeper insights.

Recent models such as VGGT and DA3 are trained on large datasets and leverage pretrained self-supervised models such as DINOv2. This raises an important question: to what extent does the model's advantage come from data-based appearance priors and pre-training priors, and how well do these capabilities generalize beyond the training distribution? To explore this, we evaluate the robustness of VGGT's geometric predictions under controlled input perturbations. We consider (i) appearance-based 2D perturbations, such as occlusions, and (ii) rendered 3D perturbations with structural ambiguities, light-source, and camera configuration changes. To assess the model's robustness, we compare VGGT against classical geometry-based methods and learned correspondence methods.

5.1 How well does VGGT handle occlusions?

To evaluate how well VGGT performs under missing visual information, we examine the impact of occlusions on correspondence matching and resulting geometric predictions. We simulate occlusions by masking key regions of the image. In principle, the model could address this challenge in multiple ways: by

utilizing global geometric reasoning, inferring missing information from learned priors, or simply ignoring the occluded areas.

Experiment setup. We intervene on input images by randomly selecting patches with known correspondences and masking them along with their surrounding regions (3×3 patch neighborhoods); see Fig. 4. We perform inference twice: first on the clean image pair and then on a pair in which one image is partially occluded. We report the change in Sampson distance error (in pixels) and the number of attention heads in which the correspondence is matched, averaged over all intervened correspondences in a scene. Results are averaged over 40 scenes at a focal length of 50 mm, with small viewpoint differences (10-25 degrees) in our controlled ShapeNet dataset. We also compare VGGT's performance with the 8-point algorithm on all matched correspondences from SIFT [24] and SuperGlue [32]. See supplementary for additional analysis.

Results. Figure 4 (left) shows the results for one case in which we masked four object patches for occlusion analysis, together with their corresponding ground-truth patches. For each occluded patch, we show the number of heads in which the occluded correspondence was matched (green for the clean version and red for the occluded version). Correspondence matching, measured by the number of matched heads (a maximum of 24 layers × 16 heads = 384 heads), shows that matched heads decrease for corrupted patches but do not collapse entirely. Figure 4 (right) reports the average number of matching heads before and after occlusion for the occluded patches. While there is a clear drop in matching heads, approximately 50% of heads are retained, indicating that the model does not ignore occluded regions but recovers correspondences through some inpainting behavior. This phenomenon may arise from the mask-based IBOT [46] loss in DINOv2 [30], which is trained to recover the masked content.

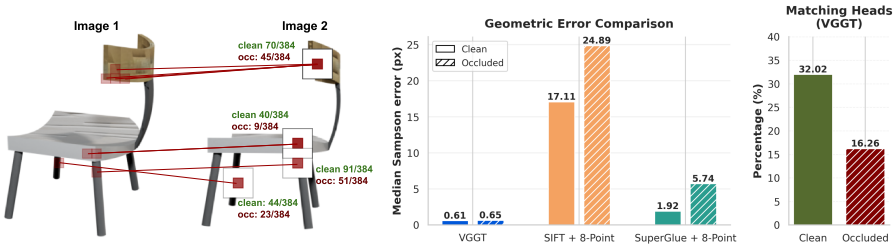

Fig. 4: Occlusion experiment. (Left) We occlude image two from a pair with white patches and report the correspondence matching ability in all layers only for the masked patches. The occluded model shows lower matching performance yet remains above zero. **(Center)** We measure the changes in the Sampson error for clean and occluded image pairs where VGGT has a minimal degradation in performance, but others are significantly affected. **(Right)** We show the average number of heads with correspondence matching on occluded patches for the clean and occluded models.

Additionally, VGGT shows only a slight degradation in the Sampson error, indicating that the epipolar geometry is still restored, whereas the other multi-stage baselines show clear degradation, shown in Figure 4 (center).

VGGT shows robustness to occlusions, preserving epipolar geometry by partially recovering matching heads for the occluded correspondences.

5.2 How robust is VGGT to scene perturbations?

We investigate the role of learned priors under controlled 3D perturbations that introduce structural ambiguity (e.g., symmetric and repetitive objects) and variations in lighting, focal length, and camera configuration. As expected, we observe that VGGT remains stable under moderate appearance and viewpoint changes, including focal length variations, with only limited degradation in geometric accuracy. Notably, when we generate challenging scenes with repetition and symmetry, for example, identical objects arranged in a circular pattern, classical pipelines struggle to establish reliable correspondences, whereas VGGT remains comparatively robust when asymmetric lighting introduces subtle cues that help disambiguate the scene. In contrast, under fully symmetric lighting and object configurations, where no disambiguating signals are present, all methods fail as expected, confirming that the task is geometrically underconstrained.

Overall, these results indicate that VGGT benefits from learned data-based appearance priors to resolve partial ambiguities, outperforming both classical and learned correspondence-based methods in such regimes, while still respecting the fundamental limits imposed by multi-view geometry when no disambiguating evidence exists. Detailed experimental setups, quantitative results, and additional analyses are provided in the supplementary material.

VGGT leverages learned visual cues to resolve partially ambiguous scenes, but like classical methods, it fails when the scene is fully geometrically symmetric.

6 Discussion

Our geometric analysis provides evidence that modern feed-forward 3D reconstruction models such as DUSt3R, VGGT, and Depth Anything 3 encode epipolar structure in their intermediate layers. A closer examination of these layers shows that several attention heads also capture point correspondences across views. The strong alignment between these emergent correspondences and geometric constraints suggests that these models organize information (at least partially) geometrically. Intervening on the most involved attention heads further confirms their functional role in encoding geometry.

While the overall behavior is consistent across the models, we observe small differences in where exactly this information emerges. In VGGT and DA3, which

use global attention in the decoder and predict in a unified 3D space, correspondences and geometric information appear in the middle layers. DUSt3R, in contrast, encodes this information early in the decoder layers. These results show that models share similar geometric mechanisms, but their architecture changes where this encoding occurs.

Our analysis of learned data priors reveals that VGGT partially fills in missing correspondences in scenes with occlusions. While input perturbations degrade both correspondence matching and geometric accuracy, the degradation is not significant. Particularly, the correspondence matching performance remains above zero in the occluded patches. This non-zero matching indicates that the model hallucinates missing correspondences rather than ignoring the occluded areas. While hallucinations could be beneficial if correct, it is fundamentally impossible to resolve single-view ambiguity when one of the two images is occluded, making this behavior potentially undesirable for reconstruction tasks. At the same time, VGGT remains robust in scenes with local feature ambiguity, leveraging global context and subtle cues such as shadows or lighting that classical pipelines fail to exploit.

7 Conclusion

We presented a systematic study of modern feed-forward 3D reconstruction models, focusing on DUSt3R, VGGT and Depth Anything 3, to determine whether they employ interpretable geometric concepts or rely primarily on learned priors from data and model. Our results show that these models encode geometric structure and strongly indicate that they use point correspondences to do so. Further, our input perturbation experiments reveal that the models rely strongly on learned priors, which provide robustness to input variations but can also lead to undesirable hallucinations.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant numbers 499552394–SFB 1597 Small Data and 539134284, through EFRE (FEIH_2698644) and the state of Baden-Württemberg, as well as by the German Ministry for Economy and Climate Protection via a decision by the German parliament (19A23014R). The compute used in this project was funded by the DFG under grants 417962828 and 539134284. Christian Rupprecht acknowledges funding by the European Union (ERC, Volute, 101222037). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Jelena Bratulić is part of the European Lab for Learning and Intelligent Systems (ELLIS) PhD program; her research visit within the ELLIS PhD program was supported through the ELSA mobility program. We want to thank Jianyuan Wang, Paul Englester, Orest Kupyn, Simon Schrodi, Karim Farid, and Rajat Sahay for valuable discussions and feedback.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)
2. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes (2018), https://arxiv.org/abs/1610.01644
3. An, H., Kim, J., Park, S., Jung, J., Han, J., Hong, S., Kim, S.: Cross-view completion models are zero-shot correspondence estimators. In: CVPR (2025)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. CVPR (2022)
5. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR (2017)
6. Bau, D., Zhu, J.Y., Strobelt, H., Lapedriza, A., Zhou, B., Torralba, A.: Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences (2020). https://doi.org/10.1073/pnas.1907375117, https://www.pnas.org/content/early/2020/08/31/1907375117
7. Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., Glass, J.: What do neural machine translation models learn about morphology? In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics, Vancouver (July 2017)
8. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
9. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: CVPR (2021)
10. Chen, Y., Chen, X., Chen, A., Pons-Moll, G., Xiu, Y.: Feat2gs: Probing visual foundation models with gaussian splatting. In: CVPR (2025)
11. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3D Awareness of Visual Foundation Models. In: CVPR (2024)
12. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
14. Geva, M., Bastings, J., Filippova, K., Globerson, A.: Dissecting recall of factual associations in auto-regressive language models. In: EMNLP (2023)
15. Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., Feng, C.: Emergent outlier view rejection in visual geometry grounded transformers. arXiv preprint arXiv:2512.04012 (2025)
16. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
17. Hartley, R.: In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(6), 580–593 (1997). https://doi.org/10.1109/34.601246

18. Heimersheim, S., Nanda, N.: How to use and interpret activation patching (2024), https://arxiv.org/abs/2404.15255
19. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. IEEE (2014)
20. Laina, I., Asano, Y.M., Vedaldi, A.: Measuring the interpretability of unsupervised representations via quantized reversed probing. In: ICLR (2022)
21. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV (2024)
22. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
23. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
25. Maggio, D., Carlone, L.: VGGT-SLAM 2.0: Real-time dense feed-forward scene reconstruction. Robotics: Science and Systems (2026)
26. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl(4) manifold (2025), https://arxiv.org/abs/2505.12549
27. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
28. Nister, D.: An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(6), 756–770 (2004). https://doi.org/10.1109/TPAMI.2004.17
29. Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: In-context learning and induction heads. Transformer Circuits Thread (2022), https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
31. Sampson, P.D.: Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm. Computer graphics and image processing **18**(1), 97–108 (1982)
32. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020), https://arxiv.org/abs/1911.11763
33. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. CVPR (2017)
34. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. CVPR (2016)
35. Shen, Y., Zhang, Z., Qu, Y., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
36. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings. pp. 835–846. ACM Press, New York, NY, USA (2006)

37. Stary, M., Gaubil, J., Tewari, A., Sitzmann, V.: Understanding multi-view transformers (2025), https://arxiv.org/abs/2510.24907
38. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
39. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
40. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: CVPR (2024)
41. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
42. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR (June 2025)
43. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. European Conference on Computer Vision (ECCV) (2018)
44. Zhang, F., Nanda, N.: Towards best practices of activation patching in language models: Metrics and methods. In: ICLR (2024)
45. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Interpretable basis decomposition for visual explanation. In: ECCV (2018)
46. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)
47. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)