Multi-View Geometry from Frozen DINOv2 Features

Gabriel López-Asiaín

1 Introduction

Vision Transformers (ViTs) [1] have revolutionized computer vision, achieving state-of-the-art results across a wide range of tasks thanks to their ability to model long-range dependencies via self-attention. On one hand, models with explicit 3D supervision have demonstrated impressive geometric understanding, enabling tasks such as dense 3D reconstruction directly from image pairs [2]. On the other, self-supervised ViT-based models, trained with no geometric supervision whatsoever, have been shown to exhibit surprising 3D awareness [3], raising a fundamental question: *what geometry do these features implicitly encode?*

In classical computer vision, geometric understanding relied on handcrafted local features such as SIFT, explicitly designed to produce geometrically consistent correspondences across viewpoints. Modern self-supervised representations like DINOv2 [4] were neither trained for geometric robustness nor exposed to camera geometry during training; yet they are increasingly used as backbones for geometric tasks. Our goal is to study how much geometric information frozen self-supervised features retain, and whether it can be extracted without the need of large learned models.

2 Setup

2.1 Data

Multi-view datasets with accurate ground-truth camera poses remain relatively scarce. The largest and most widely used is ScanNet [5], which provides dense RGB-D sequences of indoor scenes with registered camera trajectories; however, access requires institutional approval. For the scope of this project, we rely on the TUM RGB-D SLAM Dataset [6], which provides synchronized RGB streams along with ground-truth camera trajectories obtained from a motion capture system. This gives us access to calibrated intrinsics and ground-truth relative poses between frames, which we use both for supervision and evaluation of fundamental matrix estimation. For tasks that do not require multi-view geometry or ground-truth poses, such as evaluating feature extraction quality, we use Places365 [7].

2.2 Model

We use DINOv2 [4] as our frozen feature extractor. DINOv2 is a Vision Transformer trained with a self-supervised objective combining image-level and patch-level self-distillation. It produces patch-level features that capture both semantic and structural information without any language or task-specific supervision. All input images are resized to 224×224 before being fed to the encoder. The ViT-B/14 variant partitions each image into a grid of $16 \times 16 = 256$ non-overlapping patches of $p = 14$ pixels each, and outputs a feature vector of dimension $D = 768$ for every patch. All DINOv2 parameters remain frozen throughout our experiments.

2.3 Notation

We denote the frozen DINOv2 encoder by Φ. For an input image $\mathcal{I}$, the full output is

$$\Phi(\mathcal{I}) = \left[\, \mathbf{f}_{\texttt{cls}};\;\; \mathbf{f}_{1,1},\, \mathbf{f}_{1,2},\, \ldots,\, \mathbf{f}_{16,16} \right] \tag{1}$$

where $\mathbf{f}_{\texttt{cls}} \in \mathbb{R}^D$ is the `[CLS]` token and $\mathbf{f}_{r,c} \in \mathbb{R}^D$ is the patch token at grid position (r, c). The patch at position (r, c) corresponds to the image region whose top-left corner is at pixel $\big((r-1)\,p,\;\; (c-1)\,p\big)$.

Throughout, we measure similarity between two feature vectors using cosine similarity:

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\, \|\mathbf{v}\|} \tag{2}$$

which equals 1 for identical directions and 0 for orthogonal vectors.

3 Robustness of DINOv2 Features

A defining property of classical local descriptors such as SIFT is their invariance: descriptors are designed to be stable under rotation, scale change, and illumination variation. Before using DINOv2 features for geometric tasks, we want to establish whether they exhibit analogous robustness.

Given an original image $\mathcal{I}$ and a transformed version $\mathcal{I}'$, we extract features from both and compare them. Rather than comparing all patch features, we track a single reference patch and measure the cosine similarity between its feature and that of the corresponding patch in $\mathcal{I}'$, along with the similarity between the two `[CLS]` tokens. We report both metrics averaged over $1{,}000$ images from Places365 [7]. Perfect invariance ($\mathrm{sim} = 1$) is neither expected nor desirable, since a useful representation should encode the transformation itself, not discard it. What we seek to verify is that features degrade gracefully under increasing perturbation rather than collapsing abruptly.

3.1 Experiments

We evaluate robustness along five axes, each isolating a different source of variation. Empty regions introduced by a transformation are padded with the mean pixel value. The reference patch is the center of the grid $(8, 8)$ in all cases except translation, described below.

- **Rotation.** We rotate $\mathcal{I}$ by $\theta \in \{15°, 30°, \ldots, 180°\}$ about its center. The reference patch stays at the same grid position, but the pixel content within it is rotated, the global context is altered by shifted neighbors and padding, and the fixed square patch no longer covers exactly the same physical region.
- **Scale.** We resize $\mathcal{I}$ by a factor $\sigma \in \{0.25, 0.5, 0.75, 1.25, 1.5, 2\}$ about the image center, cropping or padding back to the original resolution. The reference patch remains centered on the same point, but covers a smaller region at higher detail for $\sigma > 1$ and a larger region at lower detail for $\sigma < 1$.
- **Illumination.** We multiply all pixel intensities by a gain $\alpha \in \{0.25, 0.5, 0.75, 1.25, 1.5, 2.0\}$ and clip to $[0, 255]$. Since the transformation is uniform across all pixels, the spatial structure is unchanged and any change in the features comes solely from the brightness shift.
- **Translation.** We shift $\mathcal{I}$ by $(k_x \cdot 14,\ k_y \cdot 14)$ pixels with $k_x, k_y \in \{1, 2, 3, 4, 5\}$ and pad the empty region. Because the shift is an exact multiple of the patch stride, the displaced patch observes identical pixels, so any change is entirely due to positional embeddings and global context.
- **Noise.** We add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma \in \{5, 10, 20, 40, 60, 80\}$ and clip to $[0, 255]$. The underlying content is preserved in expectation, so the drop reflects sensitivity to high-frequency perturbations.

Figures 1 and 2 summarize the results. Rotation produces a steady decline, with a mild recovery at $90°$ likely because the square patch still covers exactly the same physical region, and another at $180°$ due to the approximate symmetry of many natural images. The lowest similarities occur near $150°$, where both effects are absent. Scale perturbations affect both tokens, but the `[CLS]` token is consistently more robust. Both tokens are more stable when zooming in ($\sigma > 1$) than when zooming out, with the center patch dropping to ~ 0.37 at $0.25\times$ but remaining above 0.70 for all $\sigma > 1$.

Illumination is absorbed almost perfectly, with patch similarity above 0.88 even at $2\times$ gain. Translation shows a gradual drop to ~ 0.86 at a $(5, 5)$-patch displacement despite the shifted patch observing identical pixels, indicating that positional embeddings and global context account for a small but non-negligible share of the overall feature variation. Additive noise causes the sharpest degradation, with both tokens falling below 0.40 at $\sigma_n = 80$.

To put these numbers in context, we measure the self-similarity of unperturbed images: the center patch has mean cosine similarity 0.57 with its eight immediate neighbors and only 0.23 with non-adjacent patches. A transformed feature in the 0.5–0.6 range has therefore drifted into the similarity regime of neighboring but distinct patches.

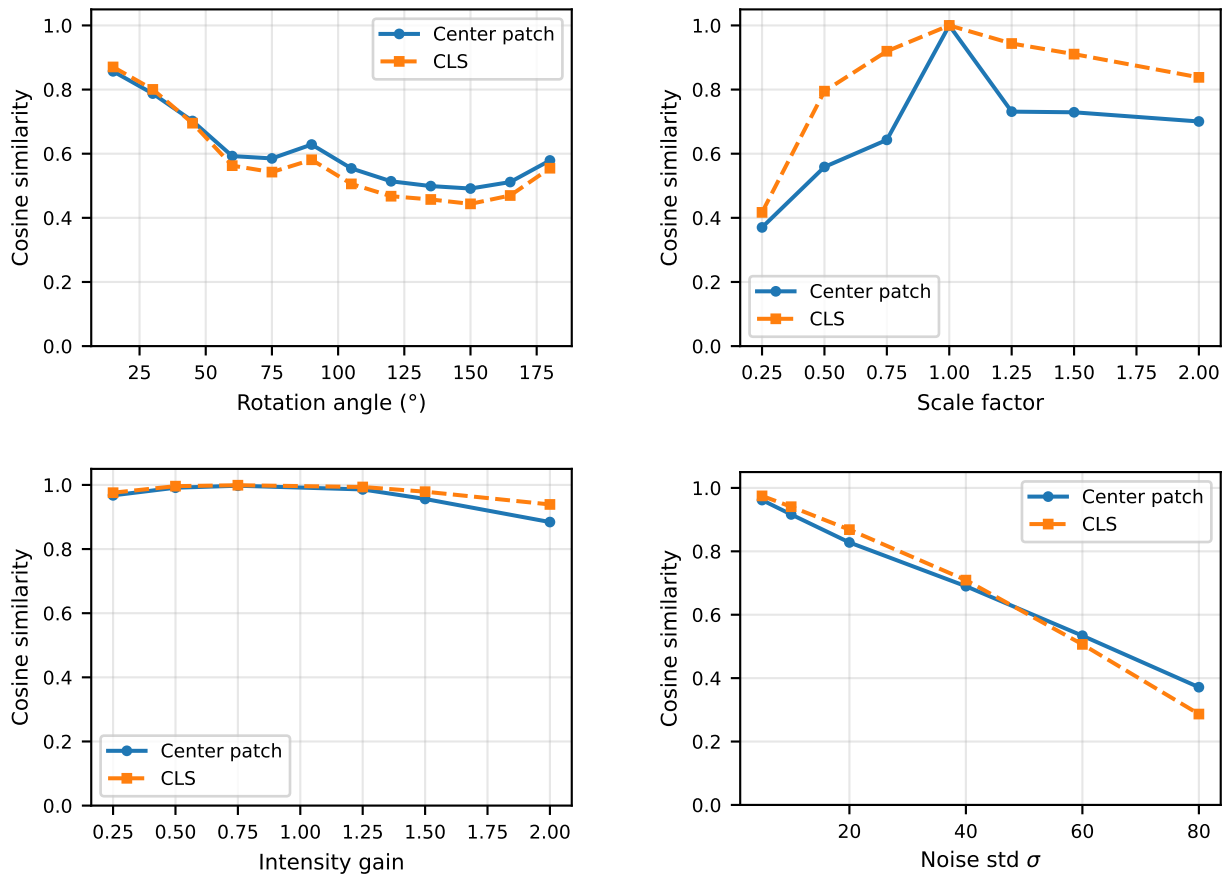

Figure 1: Cosine similarity between original and transformed features as a function of transformation intensity for rotation, scale, illumination, and additive Gaussian noise.

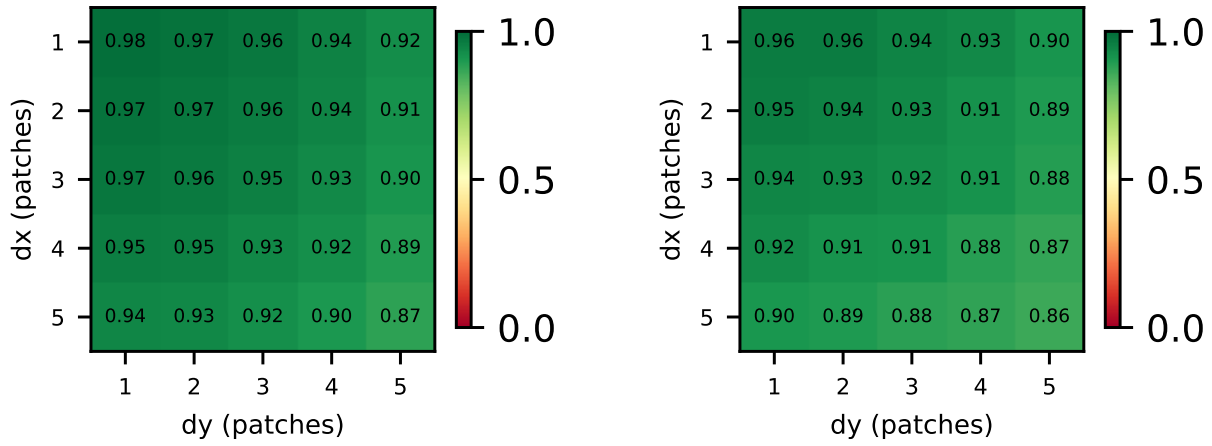

Figure 2: Cosine similarity under grid-aligned translation, displayed as heatmaps over (k_x, k_y). Left: `[CLS]` token. Right: center patch.

4 Pose estimation via classical correspondence using DINOv2 features

Having established that DINOv2 features degrade gracefully under common transformations, we now ask whether they can support geometric reasoning across views. Given two images $\mathcal{I}_1$ and $\mathcal{I}_2$ of the same scene taken from different viewpoints, can we identify pairs of descriptors $(\mathbf{f}_i,\, \mathbf{g}_j)$ that correspond to the same physical point, and can we use those correspondences to recover the relative camera pose?

4.1 Method

The simplest approach to finding correspondences is to match each patch in $\mathcal{I}_1$ to the most similar patch in $\mathcal{I}_2$ by cosine similarity. To make this more robust, we replace point-to-point comparison with a weighted average over a 3×3 neighborhood. For each candidate pair (i, j), we compute the cosine similarity at all nine offset positions and aggregate them using the kernel

$$\mathbf{W} = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{3}$$

so that the center patch contributes the most and diagonal neighbors the least. The best match for patch i is then $j^* = \arg\max_j \; s_{i,j}$, and the resulting scores $\mathbf{s}_i \in \mathbb{R}^N$ form the input to the confidence scoring stage described below.

Confidence scoring. Not all matches are reliable: the queried patch may have no true correspondence in $\mathcal{I}_2$, or repeated structures may produce ambiguous candidates. A good confidence measure should reflect two complementary quantities: the *absolute quality* of the best match and its *distinctiveness* relative to the other candidates. Consider the softmax distribution over the similarity vector with temperature τ:

$$P_j = \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{N} \exp(s_{i,k}/\tau)} \tag{4}$$

The probability P_{j^*} of the best match is high when two conditions are met simultaneously: the numerator is large, which requires s_{i,j^*} to be close to 1, and the denominator is small, which requires all other similarities to be low. In other words, the numerator captures absolute quality and the denominator captures distinctiveness. Taking the log separates these two terms additively:

$$\log P_{j^*} = \underbrace{\frac{s_i^{(1)}}{\tau}}_{\text{quality}} \; - \; \underbrace{\log \sum_{k=1}^{N} \exp(s_{i,k}/\tau)}_{\text{distinctiveness penalty}} \tag{5}$$

where $s_i^{(1)} = \max_j s_{i,j}$. We generalize this by introducing a parameter $\lambda \in [0, 1]$ that controls the relative importance of each term:

$$c(\mathbf{f}_i) = \lambda \, \frac{s_i^{(1)}}{\tau} \; - \; (1 - \lambda) \cdot \log \sum_{k=1}^{N} \exp(s_{i,k}/\tau) \tag{6}$$

When $\lambda = 1$ the score reduces to the scaled best similarity, and when $\lambda = 0$ it equals the negative log-sum-exp, which penalizes cases where many candidates have high similarity. Intermediate values balance both terms. The temperature τ controls the sensitivity to differences between scores: lower values amplify small gaps in similarity, making the score more discriminative but also more sensitive to noise. We can use confidence as an initial filter, discarding matches below a threshold or below a certain percentile.

The accepted matches are passed to RANSAC with the eight-point algorithm to obtain an estimate of the fundamental matrix $\hat{\mathbf{F}}$. Correspondences whose epipolar distance under $\hat{\mathbf{F}}$ exceeds a given threshold are discarded as outliers. Since we operate on patch centers rather than pixel-level keypoints, this threshold must be generous to account for spatial quantization, and $\hat{\mathbf{F}}$ should be understood as a coarse estimate.

The epipolar distance of a correspondence $(\mathbf{p}_i, \mathbf{p}_{j^*})$ under a fundamental matrix $\mathbf{F}$ is defined as

$$d(\mathbf{p}_i, \mathbf{p}_{j^*}) = \frac{|\mathbf{p}_{j^*}^\top \mathbf{F} \mathbf{p}_i|}{\sqrt{(\mathbf{F}\mathbf{p}_i)_1^2 + (\mathbf{F}\mathbf{p}_i)_2^2}} \tag{7}$$

which gives the perpendicular distance, in pixels, from $\mathbf{p}_{j^*}$ to the epipolar line induced by $\mathbf{p}_i$. Since correspondences are quantized to a grid of spacing p, a correct match can still incur a distance on the order of the patch size p due to discretization alone. Matches with d significantly larger than p are considered geometrically incorrect. We use this metric with the ground-truth $\mathbf{F}$ to evaluate the quality of surviving matches independently of $\hat{\mathbf{F}}$.

Refinement of matches The patch-level matches from the previous stage localize correspondences to a 16×16 grid, which limits both the accuracy of individual matches and the quality of the estimated $\hat{\mathbf{F}}$. However, each accepted match defines a small neighborhood in which the true correspondence is likely to lie. We exploit this by upsampling the features and refining each match locally.

We use FeatUp [8], which provides a pretrained upsampler for DINOv2 that maps the image to a full 224×224 resolution grid while preserving the original embedding space, so cosine similarity remains a valid matching criterion.

For each inlier $(\mathbf{p}_i, \mathbf{p}_{j^*})$ from the coarse stage, we search for the best match within a local window in the 224×224 upsampled feature map, centered at $\mathbf{p}_{j^*}$. This gives us sub-patch localization while keeping the search space small. With the refined correspondences, we re-estimate $\hat{\mathbf{F}}$ via RANSAC and discard outliers again, yielding both a tighter fundamental matrix and a more precise set of matches.

4.2 Example

We begin by testing the pipeline on a single scene from the TUM RGB-D dataset: the `fr1/desk` sequence, which provides two views of a cluttered office desk with moderate viewpoint change. Figure 3 shows the image pair. The original images have resolution 480×640; they are resized to 224×224 before being fed to the encoder, so each of the $16 \times 16 = 256$ patches corresponds to a 30×40 pixel region in the original image.

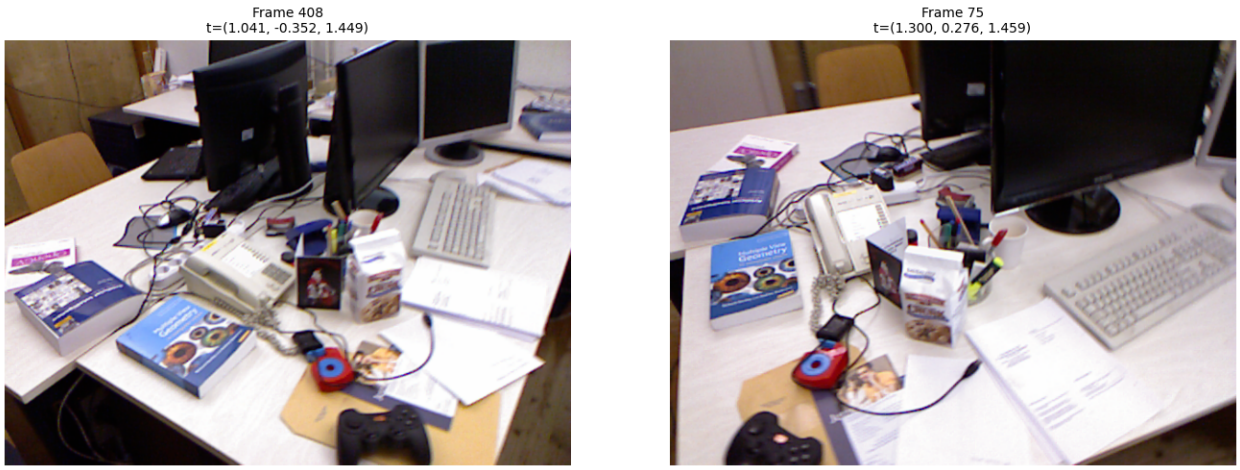

Figure 3: Two views of the `fr1/desk` sequence from the TUM RGB-D dataset, used as a running example throughout this section.

We compute the nearest neighbor of every patch using the weighted similarity described above, producing 196 candidate correspondences (patches in the border of the image are excluded). Evaluating these directly against the ground-truth $\mathbf{F}$, we find that 113 out of 196 matches (57.7%) have an epipolar distance below 30 pixels.

Since τ and λ only affect the confidence score and not the underlying matches themselves, we can perform a grid search over both parameters without recomputing correspondences. Across several image pairs, we find that $\tau = 0.03$ and $\lambda = 0.5$ (e.g. just the scaled log softmax) provide the best separation between inliers and outliers. Figure 4 shows the resulting confidence distributions for geometrically correct and incorrect matches. While the distributions differ in location, the overlap is substantial: the confidence score carries some discriminative signal but is far from a reliable classifier on its own.

We take the top 50% of matches by confidence and run RANSAC with the eight-point algorithm and threshold 25px to obtain $\hat{\mathbf{F}}$. Using this estimate, we discard all matches with epipolar distance above 30px, leaving 98 correspondences of

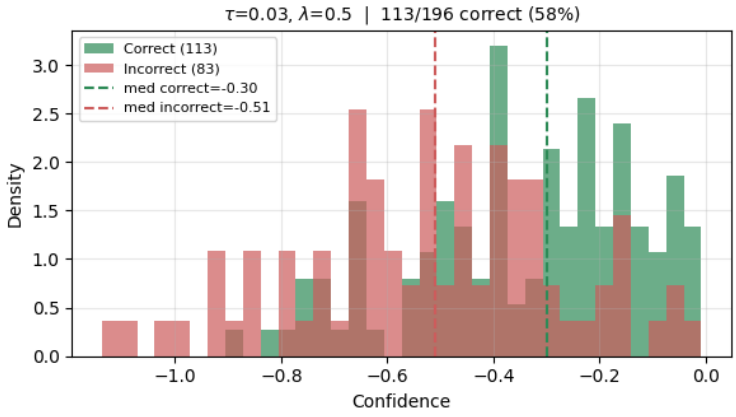

Figure 4: Distribution of the confidence score $c(\mathbf{f}_i)$ for geometrically correct (epipolar distance < 30px) and incorrect matches on the `fr1/desk` pair, with $\tau = 0.03$ and $\lambda = 0.5$. The two distributions overlap substantially, indicating that the confidence score alone is not sufficient to reliably separate inliers from outliers.

which 91% have a true epipolar distance (under the ground-truth $\mathbf{F}$) below 30px. This suggests that the patch-level matches are too coarse to yield a precise $\hat{\mathbf{F}}$ but still useful for outlier rejection. Figure 5 shows the top 15 matches by confidence overlaid on the image pair. Notably, some of the most confident correspondences land on featureless regions of the desk surface, suggesting that the model finds these textureless areas easy to match due to their low visual ambiguity across views.

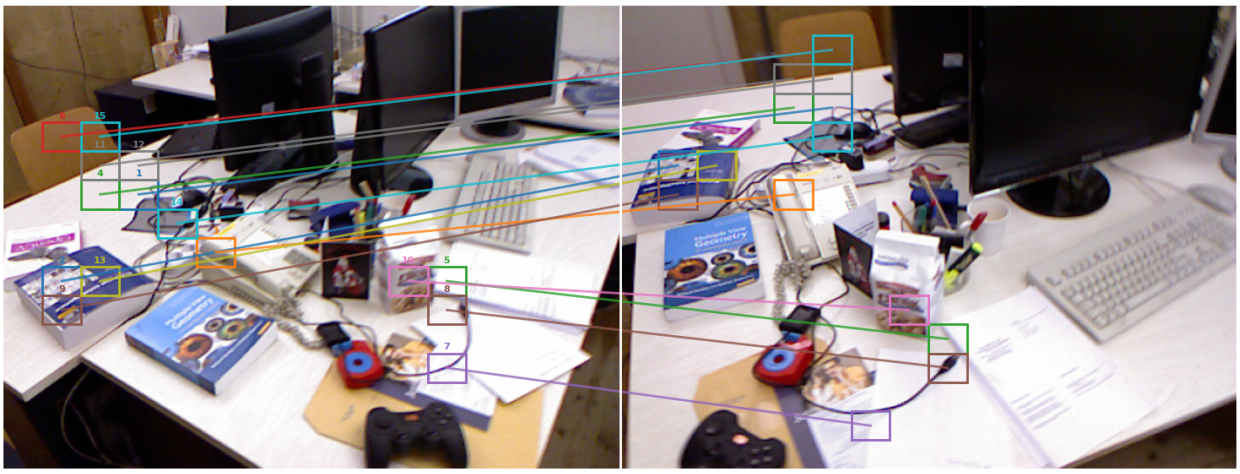

Figure 5: Top 15 matches by confidence on the `fr1/desk` pair. Lines connect matched patch centers across the two views. Several high-confidence matches correspond to textureless regions of the desk.

We now apply the refinement stage: for each inlier from the coarse step, we search for a finer match within a small window in the dense 224×224 feature map provided by the FeatUp upsampler, then re-estimate $\hat{\mathbf{F}}$ via RANSAC with a tighter inlier threshold of 10px. This produces a final set of 52 correspondences. All 52 fall below the 30px threshold under the ground-truth $\mathbf{F}$, and 83% of them are below 10px. The median epipolar distance drops from 9.5px in the coarse stage to 5.3px after refinement, confirming that the sub-patch localization yields more precise matches at the cost of a smaller set.

Since the camera intrinsics K are known, we can evaluate the quality of $\hat{\mathbf{F}}$ by recovering the relative pose of the second camera. Specifically, we compute the essential matrix as $E = K^\top \hat{\mathbf{F}}\, K$ and recover the rotation R and translation t (up to scale) using `cv2.recoverPose`, which selects the decomposition that places reconstructed points in front of both cameras.

Table 1 compares the recovered poses against the ground-truth values for both stages. While neither estimate is completely off, significant errors in orientation and translation persist, making them far from precise. Although the refined correspondences are demonstrably more precise in terms of epipolar distance, the recovered pose does not improve significantly, suggesting that the refined stage is not finding a better $\hat{\mathbf{F}}$.

Table 1: Recovered relative camera pose: coarse and refined estimates vs. ground truth.

	Ground Truth	Coarse	Refined
x (m)	1.2997	1.0335	1.0278
y (m)	0.2758	0.3144	0.3253
z (m)	1.4588	1.3185	1.4972
Roll (°)	−128.63	−133.28	−137.40
Pitch (°)	0.75	−9.33	−1.22
Yaw (°)	68.82	57.33	67.17

4.3 Results

To evaluate the method beyond a single example, we run the full pipeline on multiple image pairs from the TUM RGB-D dataset. We use the `fr1/room` sequence, sampling every 10 frames to ensure sufficient viewpoint change between pairs. Figure 6 shows the sampled frames.

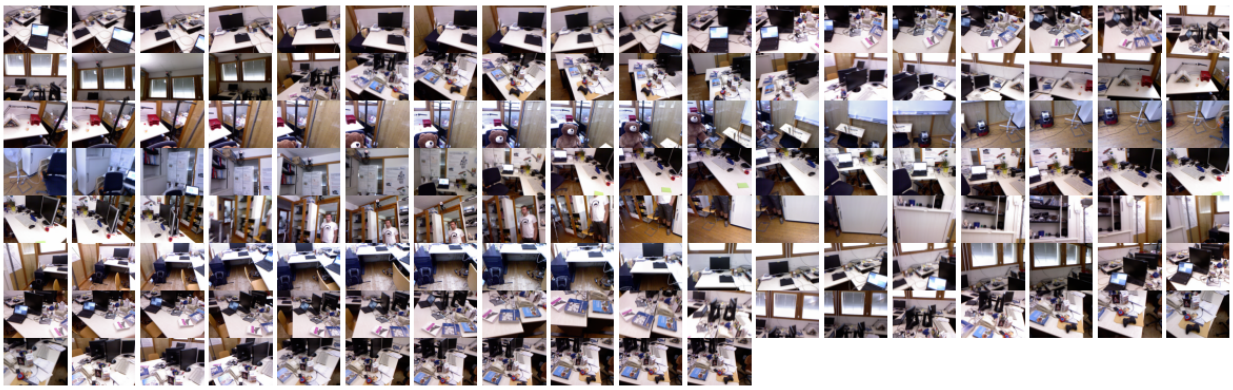

Figure 6: Sampled frames from the `fr1/room` sequence, taken every 10 frames. Consecutive pairs are used as input to the matching pipeline.

Out of 136 consecutive pairs, 107 (78.7%) produce a valid fundamental matrix estimate; the remaining 29 fail due to insufficient inliers at the RANSAC stage, typically when the viewpoint change between frames is too small or too large. For the valid pairs, the coarse stage yields on average 139 inliers with 87.7% below the 30px threshold and a median epipolar distance of 10.7px. After refinement, the inlier count drops to 90 but the median epipolar distance improves to 8.1px, with 62.2% of matches below 10px. The recovered poses show a similar pattern to the desk example: the median position error decreases slightly from 0.131m to 0.123m and the median orientation error from 7.8° to 6.6°, a modest but consistent improvement that does not fundamentally change the accuracy of the estimate.

5 Conclusions and Remarks

The experiments in this project show that frozen DINOv2 features are reasonably robust descriptors at the patch level. Matching patches by cosine similarity alone produces correspondences that are correct more often than not, and the confidence score helps filter out the worst outliers. The main limitation is inherent to the representation: each feature describes a $p \times p$ region, not a point, and this spatial quantization caps the precision of any downstream geometric estimate.

We hoped that upsampling the features with FeatUp would allow for sub-patch refinement, but the improvements

were poor. A likely explanation is that neighboring features in the upsampled map are so similar to each other that fine-grained discrimination within a small window becomes unreliable. Whether this is a fundamental property of DINOv2 features or an artifact of the upsampling method is unclear.

A more direct way to answer whether these features encode recoverable geometry would be to train a probe on them. We attempted the simplest version of this: regressing the fundamental matrix from the pair of `[CLS]` tokens. This was computationally feasible but overfit immediately on the TUM RGB-D data, which provides too few scenes for supervised learning. A more expressive model taking both $16 \times 16 \times 768$ feature grids as input would be substantially heavier to train and would need even more data, so we considered it was beyond the scope of this project.

Nonetheless, we hope the experiments provided insight into the behavior of self-supervised ViT features under geometric transformations and their limitations as pure geometric descriptors.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[2] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2024. URL https://arxiv.org/abs/2312.14132.

[3] 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, 2024. URL https://arxiv.org/abs/2404.08636.

[4] 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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. URL https://arxiv.org/abs/1702.04405.

[6] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding, 2016. URL https://arxiv.org/abs/1610.02055.

[8] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. URL https://arxiv.org/abs/2403.10516.