Teaching DINOv3 About Partial 3D Geometry: A Self-Supervised Geometry-Aware Approach

Supplementary Material

Summary

In this supplementary material, we provide additional details supporting the main paper. We begin with an overview of the notation (Sec. 7) used throughout the method section, followed by extended ablation studies (Sec. 8), like runtime comparisons, and a formal definition of the dataset-wide chirality accuracy (Sec. 9). We further describe implementation details (Sec. 10), compare our feature representation to those used in prior work (Sec. 11 and 12), present failure cases (Sec. 13) and analyse the upright bias (Sec. 14). Additionally, we show more real-world scans (Sec. 15) and provide additional qualitative results (Sec. 16), including feature visualisations. Finally, we present further 2D/3D application examples (Sec. 17) and we give some more insights on rendering with and without texture (Sec. 18).

7. Notation Table

For better understandability, we show an overview of the notation used in the method section in Table 7.

Symbol	Description
$\mathcal{X} = (\mathbf{V}_\mathcal{X}, \mathbf{T}_\mathcal{X})$	3D Full Shape
$\mathbf{V}_\mathcal{X}$	Vertices of shape $\mathcal{X}$
$\mathbf{T}_\mathcal{X}$	Triangles of shape $\mathcal{X}$
$\mathcal{Y} = (\mathbf{V}_\mathcal{Y}, \mathbf{T}_\mathcal{Y})$	3D Partial Shape, $\mathbf{V}_\mathcal{Y} \subset \mathbf{V}_\mathcal{X}$, $\mathbf{T}_\mathcal{Y} \subset \mathbf{T}_\mathcal{X}$
$\mathbf{\Pi}_{\mathcal{Y}\mathcal{X}}$	Correspondence from partial to full shape
$\mathcal{C} = (c^1, ..., c^N)$	Set of all cameras
$\mathbf{I}^i_\mathcal{X}$	Image of full shape from $c^i \in \mathcal{C}$
$\mathbf{Q}^i_\mathcal{X}$	Features computed on $\mathbf{I}^i_\mathcal{X}$
$\mathbf{F}_\mathcal{X}$	Vertex-wise features of shape $\mathcal{X}$
$\mathbf{D}_\mathcal{X}$	Geodesic Distance Matrix of $\mathcal{X}$

Table 7. Overview of Symbols

8. Further Ablation Studies

LoRA Size We evaluate different values for hyperparameters, α and the rank r in LoRA on the BECOS validation set in terms of geodesic error (Table 8). The values $r = 16$ and $\alpha = 32$ yield the best results.

Inference Time Table 9 compares the inference time of different methods. The additional use of stable diffusion in Diff3f results in a longer processing time compared to using DINO features alone. The other feature descriptors, including our method GeoLoRA, have similar runtimes.

r/α	4/8	8/16	16/32
Geo Error (↓)	8.14	6.56	**5.89**

Table 8. **LoRA Hyperparameter Ablation:** We observe the best results with rank $r = 16$ and $\alpha = 32$.

	DINOv2	DINOv3	Diff3f	GeoLoRA
Time [s] (↓)	4.72	4.45	360.45	4.59

Table 9. **Inference Time Comparison:** While DINOv2, DINOv3 and GeoLoRA have similar runtime during inference, Diff3f's runtime is significantly increased due to the need to additionally run inference with Stable Diffusion.

Backbone Scale Replacing the ViT-B backbone with a ViT-L backbone does not lead to significant performance gains in terms of geodesic error (see Table 10). Nevertheless, ViT-L requires 50% more training time and twice the amount of VRAM on the GPU than ViT-B.

PFAUST	DINOv3-B	GeoLoRA-B	DINOv3-L	GeoLoRA-L
-M	11.36	2.88	11.40	**2.43**
-H	13.82	3.33	12.50	**2.65**

Table 10. **Performance of Different Backbone Scales:** We show that with a larger backbone we do not gain significant improvement in terms of geodesic error.

9. Dataset-wide Chirality Accuracy Definition

Given a full dataset U with M partial shapes $U = \{\mathcal{Y}_1, ..., \mathcal{Y}_M\}$, we define the chirality accuracy as:

$$acc_U = \frac{1}{M} \sum_{\mathcal{Y}_i \in U} \text{acc}_{\mathcal{Y}_i} \tag{4}$$

We can define the final chirality accuracy as:

$$acc_{\text{chi}} = \max\{acc_U, 1 - acc_U\} \tag{5}$$

.

10. Additional Implementation Details

For the temperature parameter, we follow [3] and use $\tau = 0.07$. We run the training on one A40 GPU and five cores

of a Xeon Gold 6248R CPU. The training takes approximately 5 days for $50,000$ iterations for the longest-lasting BECOS dataset. The training time can vary between 2-5 days between the different datasets according to different number of vertices. As a prompt to Stable Diffusion inside Diff3f, we use "animal" for animal shapes, "human" for human shapes and "centaur" for centaur shapes.

11. ULRSSM: Comparison to Original Features

Instead of using foundation features or handcrafted descriptors, an alternative approach is to start with the raw XYZ coordinates as input. Since previous methods have shown promising results, for fairness, we also compare our approach with ULRSSM [12] using XYZ coordinates as input. In Table 11, we show that the performance falls far behind the use of foundation features. The use of XYZ is sensitive to shape orientations, which are particularly ambiguous in partial shapes.

Geo Error (↓)	XYZ	Ours
BECOS	40.46 (40.18)	**11.62 (11.54)**
SHREC16CUTS	7.04 (3.2)	**3.01 (1.97)**
SHREC16 HOLES	13.75 (8.2)	**8.07 (6.88)**
PFAUST-M	7.65 (7.49)	**1.61 (1.61)**
PFAUST-H	9.39 (9.24)	**2.29 (2.19)**

Table 11. **Comparison of XYZ features vs GeoLoRA features:** We compare the geodesic error ($\times 100$) on different partial-to-full datasets with ULRSSM. On all datasets, our features outperform the originally used XYZ features. The numbers in brackets "()" indicate geodesic error results after test-time adaptation [12].

12. Left-Right Prediction: Comparison to Original Features

In the main paper, we compare the GeoLoRA features with those of DINOv3. For completeness, here we report a comparison with the features originally used in [58], which combine DINOv2 and Stable Diffusion features (Table 12).

	Original	Ours
Chirality Accuracy (↑)	57.40	**91.42**

Table 12. **Comparison to DINOv2 + Stable Diffusion Features:** We compare the chirality (left-right) accuracy of the features used in [58] to our GeoLoRA features. Our method's features show higher accuracy.

13. Failure Cases

The datasets used to train our models primarily contain upright shapes, resulting in most failure cases involving shapes in different poses, such as a rotated person in a pike position or a person lying on their side with their hands on the ground. We show two examples in Figure 7.

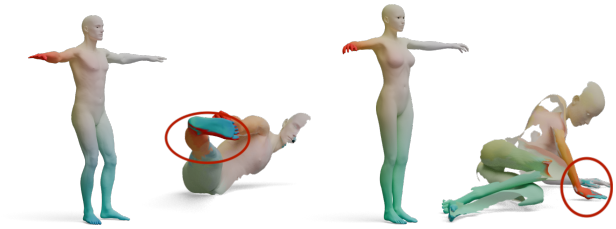

Figure 7. **Failure Cases:** We show two failure cases of our method. The method primarily struggles with shapes in challenging poses, where the shapes are not positioned upright.

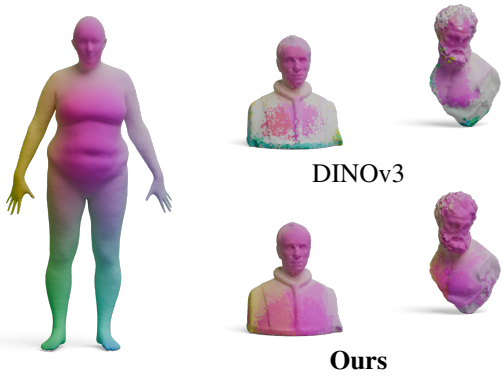

Figure 8. **Additional Results on Real-World Statues from [1]:** DINOv3 fails to find good correspondences between the partial statues and the full template shape, where GeoLoRA finds reasonably dense correspondences.

14. Upright Bias

For consistency with the literature, we assume that shapes are oriented upright. However, we train GeoLoRA on FAUST (for 70K iterations), sample random rotation along all axes, and test upside-down partial shapes. GeoLoRA achieves 6.69 geodesic error, while DINOv3 performs $10\times$ worse (63.84).

15. Additional Real-World Experiments

Additional Real World Scans of Statues We demonstrate further real-world scans of statues from [1] and match them to a template shape from FAUST. Here, GeoLoRA shows qualitatively better matchings than DINOv3.

Real-World Validation Quantitative evaluation on real-world partial scans is limited by the lack of ground-truth data availability. Therefore we evaluate the annotated real scan of SHREC20. Our method significantly reduces geodesic error compared to DINOv3 (5.92 vs. 16.26), indicating improved robustness on real-world data.

16. Additional Qualitative Results

Feature Quality We show more qualitative comparisons to other features in Figure 11. Additionally, we show further examples of our feature predictions in Figure 12.

Partial Shape Matching We show qualitative results for partial-to-partial shape matching in Figure 13 and additional results for partial-to-full shape matching in Figure 14.

17. Further 2D/3D Applications

While we focus on learning features for 3D partial shape matching, we find their use in other contexts exciting. Hence, we use our GeoLoRA model (trained on BeCoS) to perform image matching on SPair-71k [42] dataset images. In several cases, our approach provides better matching than DINOv3, especially in images where foreground elements differ in their view (see Table 9). Additionally, we show that our method can also be used for 3D keypoint matching for shapes of the BeCoS test set (see Table 10).

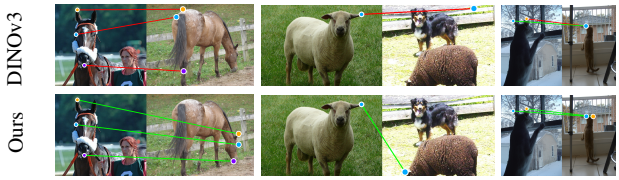

Figure 9. **2D Keypoint Matching:** We show examples for keypoint matching on the SPair-71k [42] dataset. Our method shows superior matching results in these examples.

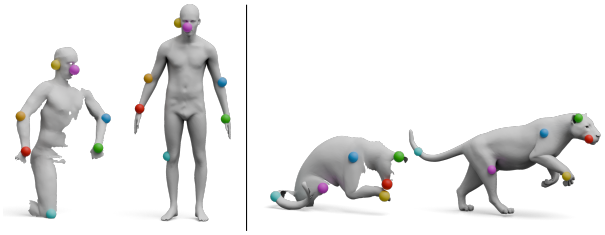

Figure 10. **3D Keypoint Matching:** Our method can be used for 3D keypoint matching. We show examples of the BeCoS test set.

18. Texture & Rendering

All partial shape matching benchmarks are textureless, so meshes are rendered without texture (single colour with shadow on the surface, see Fig. 2 Image Column). Using SMPL [41], we empirically test texture sensitivity and find that adding texture degrades performance (geodesic error 3.65 vs. 1.32). We hypothesise that texture helps identify correspondences within the same (e.g., person), but not across different identities (e.g., due to different clothing).

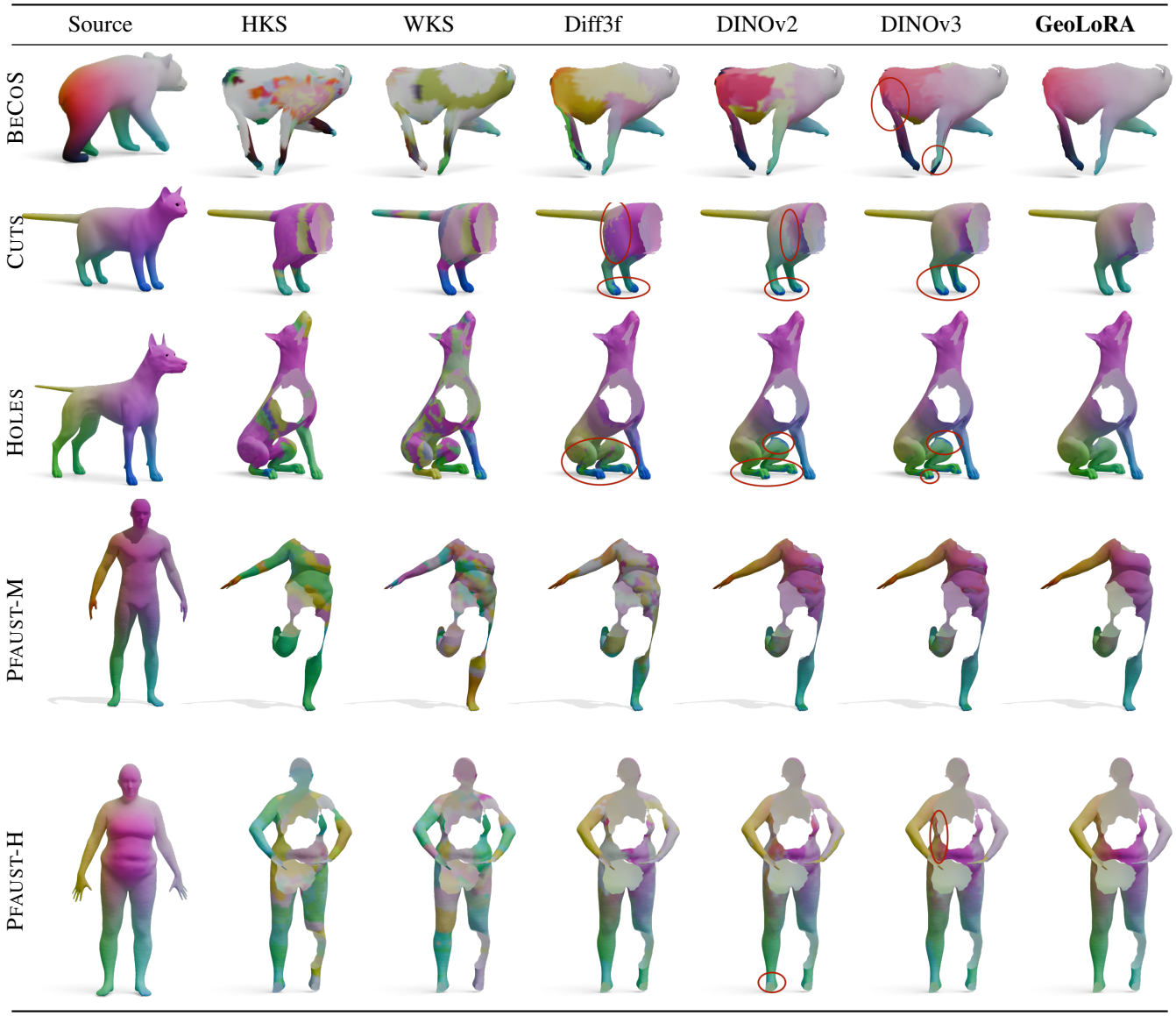

Figure 11. **Qualitative Results of Feature Quality:** We show colour-coded correspondences based on the features computed with different feature extractors. GeoLoRA exhibits improved performance, especially in left-right and front-back predictions.

Figure 12. **More Qualitative Results:** We show more qualitative results of our partial-to-full features.

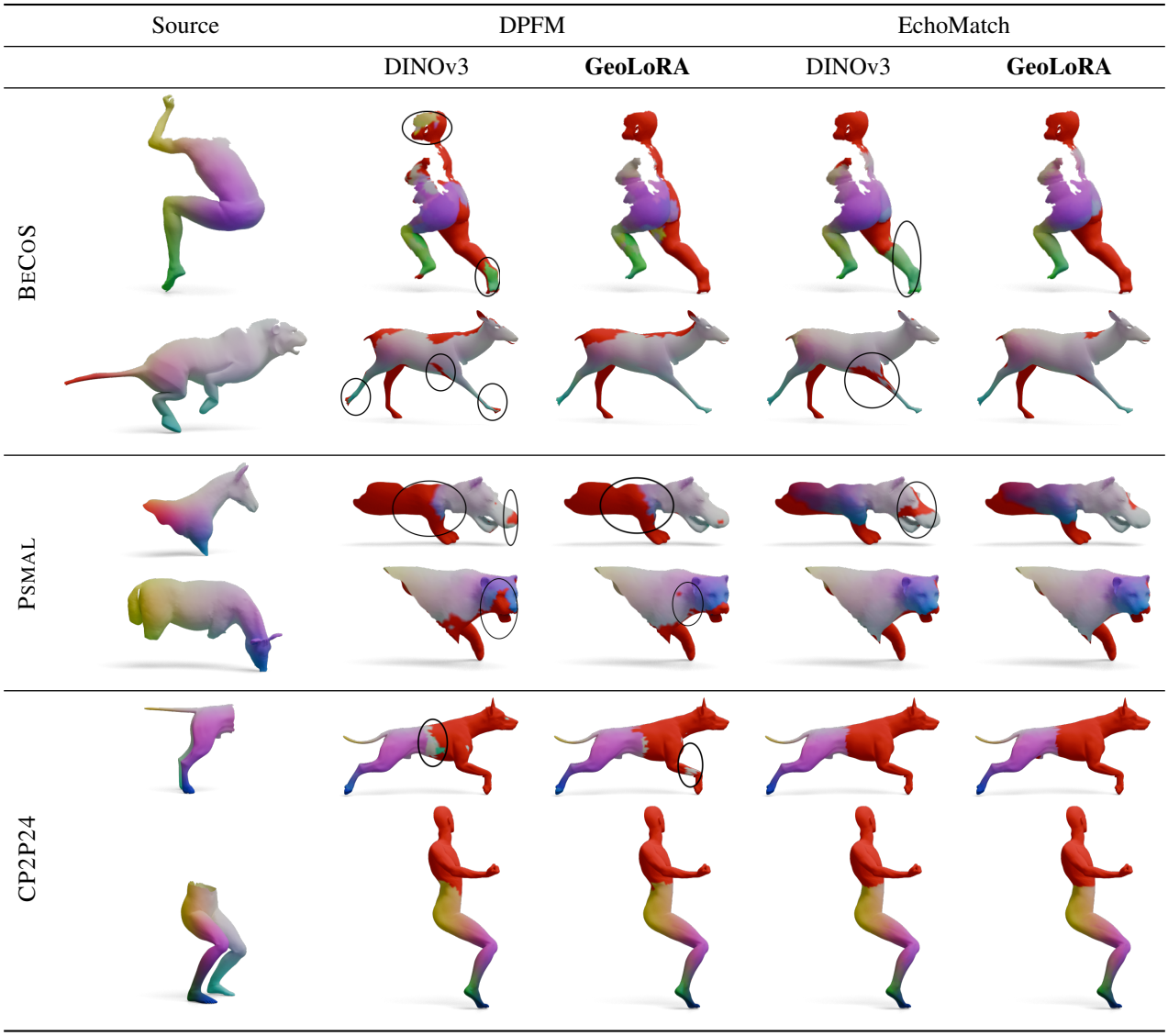

Figure 13. **Qualitative Results of Partial-to-Partial Matching:** We show colour-coded correspondences for partial-to-partial shape matching methods. We compare DINOv3 features with GeoLoRA features. The red area indicates points that are predicted not to be in the overlapping region. On the BECOS dataset and the PSMAL dataset, we demonstrate improved performance with GeoLoRA features compared to DINOv3 features. On CP2P24 we observe similar matching quality.

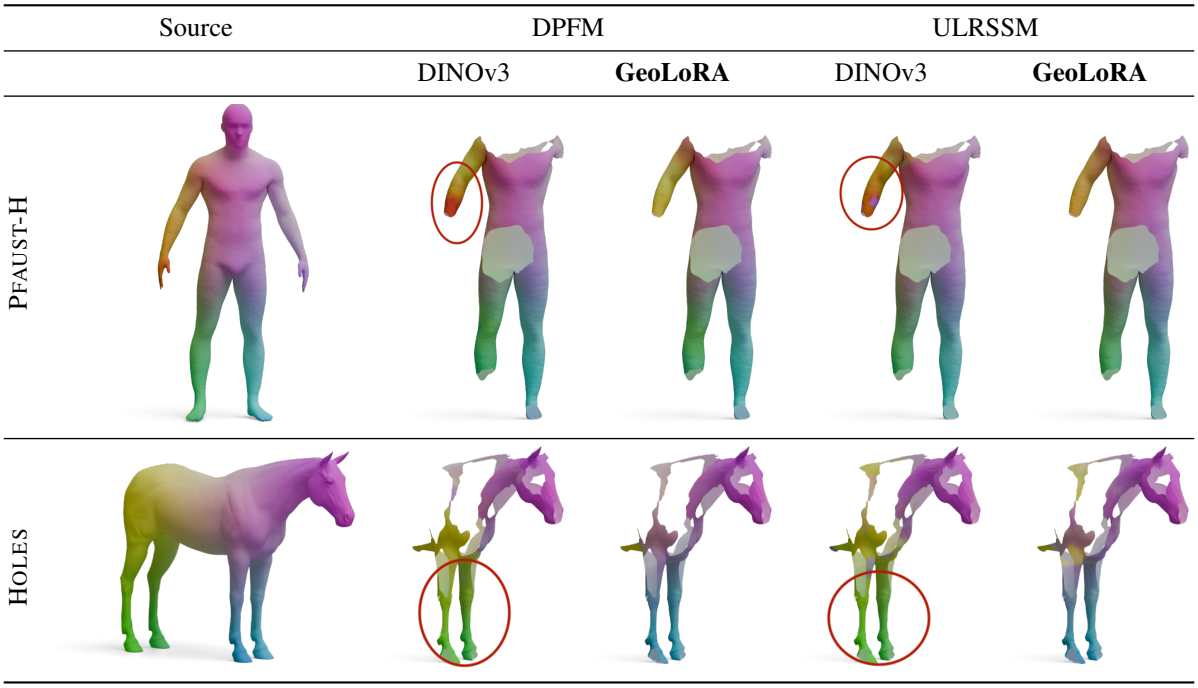

Figure 14. **Qualitative Results of Partial-to-Full Shape Matching Methods:** We show two more examples of partial-to-full shape matching on the PFAUST-H and SHREC16 HOLES dataset. We can see improved matching quality with our GeoLoRA features, especially in the extremities of the shapes, such as the left arm of the human and the legs of the horse.