SGSoft: Learning Fused Semantic-Geometric Features for 3D Shape Correspondence via Template-Guided Soft Signals

Supplementary Material

A. Extended Method Details

A.1. Canonical Geodesic Correspondence Field Construction

The geodesic correspondence field $\tilde{S}$ serves as the supervisory signal of our framework. It provides stable, continuous, and topology-invariant correspondence likelihoods defined entirely on the canonical template mesh. This field is computed *once* on the template and reused across all augmented meshes, enabling consistent supervision under arbitrary remeshing, while avoiding repeated geodesic computation on each variant.

We denote the canonical template mesh as $T = (V_T, F_T)$ with vertex set $V_T = \{t_1, \ldots, t_{N_T}\}$. For each augmented mesh vertex x_i, a one-to-one mapping $h(i)$ is maintained through the augmentation chain, linking x_i back to its canonical vertex $t_{h(i)}$. This mapping allows the precomputed template field $\tilde{S}_{h(i),v}$ to be directly used as the correspondence distribution for x_i, eliminating the need for geodesic recalculation on augmented meshes. (See Fig. 1)

Step 1. Intrinsic Precomputation on the Template. We first compute a set of intrinsic geometric quantities on the canonical template mesh:

- local vertex area $A[v]$ is computed as one-third of the total area of its incident one-ring faces.
- global median vertex area a_{med}.
- curvature estimate $\kappa[v]$ (e.g., mean curvature magnitude [15]).
- geodesic distances $d_{\text{geo}}(p, v)$ for all pairs of vertices on T, computed using an intrinsic solver (e.g., heat method [3]).

Step 2. Adaptive Neighborhood Size. For each template vertex t_i, we compute a density coefficient inversely proportional to its local surface area:

$$\rho_i = \frac{a_{\text{med}}}{A[i] + \epsilon}, \tag{1}$$

where $A[i]$ is the vertex area and a_{med} is the global median area. The density ρ_i is clipped to a predefined range and translated into an adaptive neighborhood size:

$$K_i = \text{clip}(\text{round}(K_{\text{base}} \cdot \rho_i^{\alpha}), K_{\min}, K_{\max}). \tag{2}$$

We set $K_{\text{base}} = 32$, $K_{\min} = 8$, and $K_{\max} = 50$.

This strategy allocates denser neighborhoods in geometrically detailed regions (e.g., fingers, facial features) and avoids redundant sampling over flat areas, yielding a balanced intrinsic support. By additionally capping the neighborhood size, we maintain computational efficiency and sparsity, while preventing over-dense connections in locally uniform regions.

Step 3. Base Geodesic Kernel. Given the precomputed geodesic distance matrix, we define an intrinsic neighborhood

$$\mathcal{N}_i^{(K_i)} = \text{Top-}K_i \text{ vertices } t_v \text{ with smallest } d_{\text{geo}}(t_i, t_v). \tag{3}$$

For each $v \in \mathcal{N}_i^{(K_i)}$, we construct a geodesic Gaussian kernel:

$$\hat{S}_{i,v} = \exp\left(-\frac{d_{\text{geo}}(t_i, t_v)^2}{\sigma^2}\right), \tag{4}$$

which represents the base intrinsic correspondence kernel defined purely by geodesic distances on the canonical template surface. This kernel is later modulated by geometric and semantic constraints to obtain the final correspondence field $\tilde{S}$.

Step 4. Geometric and Semantic Modulation. We further refine the intrinsic field by incorporating geometric saliency and semantic constraints.

Geometric saliency weighting. To emphasize high-curvature structures, we define a curvature-aware weight:

$$A_{i,v} = (1 + |\kappa_v|)\ \mathbb{1}\left[t_v \in \mathcal{N}_i^{(K_i)}\right], \tag{5}$$

where κ_v denotes the mean curvature magnitude at t_v.

Semantic plausibility filtering. To suppress implausible correspondences (e.g., hand-to-foot), we apply a same-part mask:

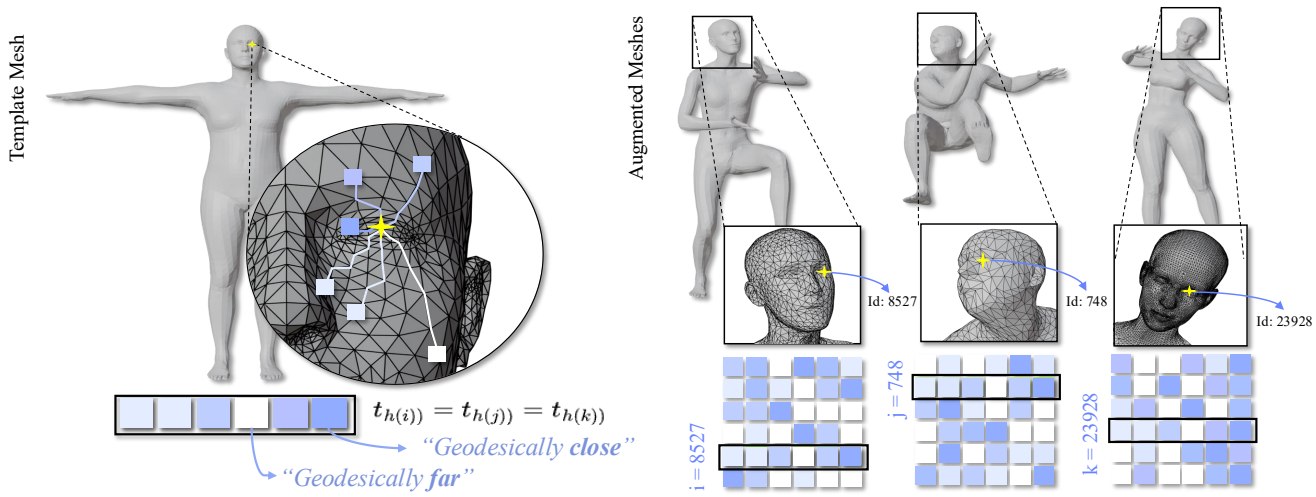
$$\Pi_{i,v} = \mathbb{1}[\ell(t_v) = \ell(t_i)], \tag{6}$$

where $\ell(\cdot)$ denotes the semantic part label of a vertex.

Step 5. Final Geodesic Correspondence Field. The geodesic-aware soft correspondence field is defined as:

$$\tilde{S}_{i,v} \propto A_{i,v} \cdot \hat{S}_{i,v} \cdot \Pi_{i,v}, \qquad \sum_v \tilde{S}_{i,v} = 1, \tag{7}$$

This yields a compact, sparse, and intrinsically geometry-aware correspondence distribution centered at vertex t_i.



(a) Precomputed Geodesic Correspondence Field on a Canonical Template Mesh

(b) Lifting Hard Anchors into Geodesic Correspondence Field $\tilde{S}$

Figure 1. **Topology- and deformation-consistent geodesic correspondence field.** (a) We precompute a continuous geodesic correspondence field on a canonical template mesh, where each hard anchor is lifted to an intrinsic geodesic neighborhood defined by surface distances. (b) The field is consistently reused across remeshed and deformed shapes via hard anchor mapping, yielding a topology- and deformation-invariant correspondence field $\tilde{S}$.

Canonical Reuse Across Augmented Meshes. Since the correspondence field $\tilde{S}$ is defined entirely on the canonical template using intrinsic geodesic distances, it is computed *once* and reused for all augmented meshes. For an augmented vertex x_i with canonical anchor $h(i)$, we directly assign

$$\tilde{S}_{x_i}(v) = \tilde{S}_{h(i),v}. \tag{8}$$

This provides consistent, topology-invariant supervision under remeshing and deformation, while completely eliminating the need for recomputing geodesic distances on each augmented shape.

A.2. Hierarchical Semantic Aggregation

To integrate semantic information across multiple transformer depths, we adopt a hierarchical aggregation strategy that fuses intermediate features from selected layers of the Uni3D [20] backbone. Shallow layers capture fine geometric details, while deeper layers encode global semantic context.

Specifically, for the 40 layers of Uni3D transformer backbone, we extract features from layers $\ell = \{8, 16, 24, 39\}$ and perform hierarchical fusion over these representations.

Given features $\{H^{(\ell)}\}_{\ell \in \mathcal{I}}$ from selected layers $\mathcal{I}$, we project them into a shared space and aggregate them with learned layer weights:

$$f_{\text{sem}} = \sum_{\ell \in \mathcal{I}} \alpha_\ell \, H^{(\ell)}, \quad \alpha_\ell = \text{softmax}(a_\ell). \tag{9}$$

This weighted fusion balances contributions from low- and high-level features, producing a unified semantic embedding that captures discriminative semantics and global contextual structure.

A.3. Ablation Studies

We analyze the contribution of three core components, namely (A) adaptive neighborhood size (density-aware sampling), (B) geometric saliency weighting, and (C) hierarchical semantic aggregation. As shown in Table 1, removing any of these components degrades performance across benchmarks.

w/o A significantly increases the error on SCAPE and SHREC19, highlighting the importance of density-aware intrinsic sampling. In contrast, DT4D-Inter contains large global shape variations across subjects. In such settings, a uniform neighborhood leads to more stable global matching, whereas density-aware sampling, which is optimized for local isometric detail, can introduce an unnecessary locality bias.

w/o B leads to the largest degradation on SCAPE and SHREC19, demonstrating that curvature-based saliency is essential for dynamically posed and highly articulated structures.

w/o C also causes substantial degradation on all benchmarks, indicating that multi-level semantic fusion is crucial for robust part discrimination and consistent semantic alignment across large pose and shape variations.

Overall, the full model (A+B+C) achieves the best per-

formance, and each component provides a complementary benefit to dense correspondence accuracy.

Table 1. **Ablation study on three extended core components of SGSoft.** Reported values are mean geodesic error (↓) across SCAPE, SHREC19, and DT4D-Inter.

Variant	SCAPE	SHREC19	DT4D-Inter	Avg.
SGSoft (full; A+B+C)	**2.9**	**4.0**	8.3	**5.03**
w/o A (Uniform neighborhood)	7.1	8.7	**8.0**	7.93
w/o B (No geometric saliency)	8.0	11.9	8.3	9.40
w/o C (No hierarchical aggregation)	4.8	12.5	11.5	9.60

B. Training Dataset and Implementation Details

B.1. Training Dataset

We construct our training set by randomly sampling body shapes and poses using the SMPL model [10]. Pose parameters are generated in the VPoser latent space to ensure physically plausible and realistic human motions. From this process, we obtain 183 base human meshes, as illustrated in Fig. 2. Each base mesh is further diversified with 10 topological variations, resulting in a total of 1,830 training meshes, each associated with a precomputed geodesic correspondence field.

B.2. Implementation Details

Uni3D Backbone. We use Uni3D with the EVA-Giant-Patch14-560 backbone (1B parameters) as our 3D semantic encoder. Each input point cloud is partitioned into $G = 512$ local patches, each containing $M = 64$ points, using Farthest Point Sampling with geodesic-aware assignment.

Curriculum Learning. We adopt a simplified two-phase curriculum learning strategy. During a warm-up stage (1 epoch), training focuses solely on the part classification loss to stabilize semantic alignment. This is followed by a transition stage of 8 epochs, during which classification and contrastive losses are smoothly blended using a cosine schedule. The final loss weights are set to $(w_{\text{part}}, w_{\text{cont}}) = (0.6,\ 0.3)$. We additionally apply a symmetry consistency regularization with weight 0.3.

Training. We train the model for a total of 13 epochs with batch size 1. Training takes approximately 17 hours on a single NVIDIA A6000 GPU (48GB). To ensure reproducibility, the full codebase and trained models will be released.

C. Baselines

The selected baselines in main paper are drawn from related studies **where official codes are publicly available.**

(i) **Deformation-based feed-forward.** NJF [1] learns a neural deformation field, predicting dense correspondences via forward regression within a canonical domain.

(ii) **2D-pretrained feature distillation.** Diff3F [4] transfers 2D diffusion features to 3D surfaces via multi-view rendering conditioned on depth and normal maps.

(iii) **Hybrid with Functional Maps.** DiffuMatch [12] combines 2D diffusion features with iterative functional map refinement for dense alignment. DenoisingFM [21] estimates correspondences via a denoising diffusion process that reconstructs functional maps from noisy spectral embeddings.

Figure 2. **Training dataset overview.** We sample 183 base human meshes from the SMPL model using VPoser-driven body poses and random body shapes, and generate 10 topologically perturbed variants for each base mesh, yielding 1,830 training meshes in total.

For all baselines, we report the performance from their official implementations under default configurations, with two exceptions: for DenoisingFM [21], we adopt the 32×32 functional map resolution (its best on the DT4D benchmark) for consistency; and for NJF [1], we report the performance following the configuration from the recent baseline [12], as the original work does not provide an official configuration for this shape correspondence benchmark.

D. Additional Qualitative Comparison

Figure 3 shows an additional qualitative comparison under a challenging scenario with dynamic pose deformations between the source and target shapes. While functional map-based methods such as DenoisingFM and DiffuMatch handle large articulations reasonably well, they suffer from noticeable errors in locally ambiguous regions with minor deformations, such as the upper arm, hand, and thigh (highlighted in red boxes).

In contrast, SGSoft produces more stable and semantically consistent correspondences in these regions by leveraging the proposed geodesic correspondence field together with semantic cues.

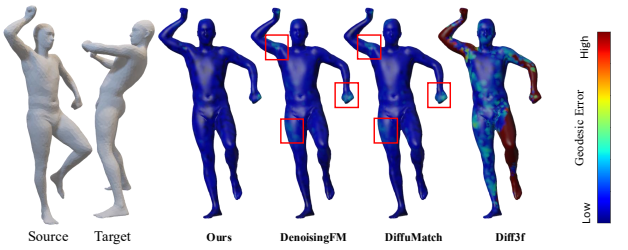


Figure 3. **Additional qualitative comparison.** Geodesic error visualization for challenging correspondence under dynamic pose deformation.

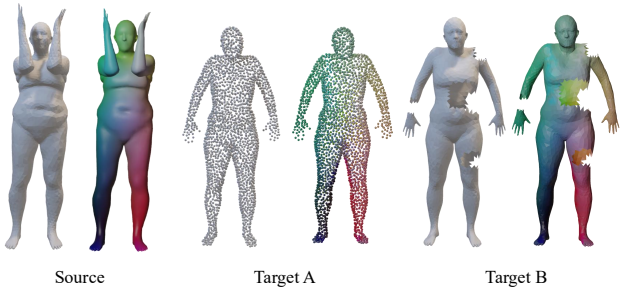


Figure 4. **Representation robustness.** Target A is a point cloud and Target B is a partial mesh. SGSoft produces consistent correspondence colors across both targets, showing representation-agnostic behavior.

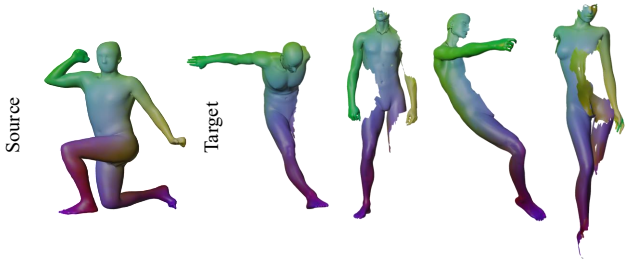


Figure 5. **Qualitative results on samples from SHREC'16.** Despite disconnected geometry, SGSoft maintains semantically consistent correspondences in visible areas.

E. Additional Results

E.1. Symmetry Disambiguation

Figure 12 visualizes SGSoft's ability to disambiguate left-right and front-back symmetries across FAUST, SCAPE, and SHREC19. For each example, we transfer correspondences from a source mesh to a target mesh under a different pose. The uncolored source and target meshes are shown in their original poses, while the colorized meshes are rotated to a canonical front view for clearer visualization. SGSoft consistently assigns stable and semantically correct colors to corresponding parts (e.g., left/right arms and legs) even under strong articulation.

E.2. Zero-Shot Cross Domain Generalization

Figure 13 presents zero-shot generalization results from human-like training shapes to stylized character meshes. The *Intra* columns evaluate correspondence quality within the same character family, while the *Inter* columns demonstrate transfer across different families with large variations in appearance and body proportions. Although SGSoft is trained exclusively on SMPL-like human bodies, it preserves dense correspondences even across highly non-isometric shapes. These results indicate that our multimodal, intrinsic representation generalizes beyond the training domain and remains robust under strong geometric and stylistic distribution shifts. For clarity of comparison, the source and target meshes are shown in their original poses, while the colored correspondence results are rendered after a consistent front-facing rotation.

E.3. Representation Robustness

In Figure 4, we probe the robustness of the learned multimodal intrinsic space across different geometric representations and levels of completeness. We match a single source to two structurally different targets: *Target A*, represented as a point cloud, and *Target B*, given as a partial mesh. Despite these variations, SGSoft assigns consistent semantic correspondences across both targets, indicating stability under moderate structural changes.

To further evaluate robustness under more severe incompleteness, we additionally present qualitative results on the SHREC'16 partial shape benchmark (Fig. 5). Even with large missing regions and disconnected geometry, SGSoft maintains semantically coherent correspondences in most visible areas, although performance degrades in extreme cases in Fig. 10. These results suggest that the learned descriptor remains robust across a spectrum of geometric representations and partial observations.

E.4. Topology Robustness

To evaluate robustness under remeshing, Fig. 11 visualizes correspondence results with varying vertex resolutions and triangulations. Despite large changes in connectivity and sampling density, SGSoft preserves smooth and semantically consistent color patterns across a wide range of topologies. These results suggest that our template-guided geodesic correspondence field provides largely topology-invariant supervision that generalizes across different surface discretizations in practice. Table 2 further quantifies this behavior on 10 randomly sampled SHREC19 pairs with 50 random remeshing perturbations in total. Although the average geodesic error slightly increases from 0.0325 to 0.0386 after remeshing, the overall performance remains in a comparable range, supporting the practical robustness of

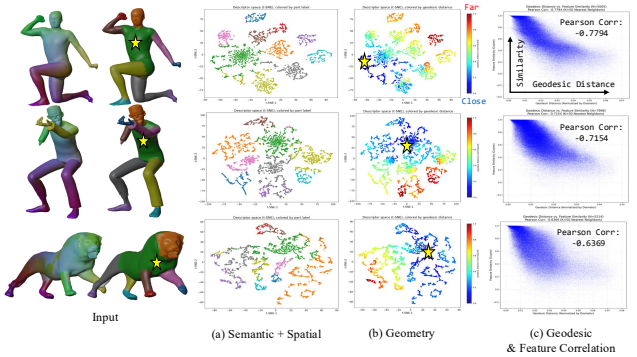


Figure 6. **t-SNE visualization and feature–geometry correlation.** (a) The descriptor space forms semantic clusters and separates symmetric regions across diverse shapes. (b),(c) Feature similarity from a reference point (star) negatively correlates with geodesic distance, consistent with Pearson trends.

SGSoft under substantial changes in mesh resolution and triangulation.

Table 2. Geodesic correspondence error under random remeshing on SHREC19. Lower is better.

Setting	Mean Error ↓
Baseline	0.0325
Avg. remeshing	0.0386

E.5. Analysis of Multimodal Descriptor

To demonstrate the distinct roles of each modality, we analyze the structure of the learned feature space using t-SNE and feature–geometry correlation. As shown in Fig. 6(a), the descriptor space exhibits clustering patterns aligned with semantic body parts, while also separating symmetric regions, indicating effective semantic and spatial discrimination. These patterns remain consistent across diverse shape categories, including humans, stylized characters, and animals, suggesting that the learned representation generalizes beyond the training domain.

Furthermore, feature similarity from a reference point (denoted by a star) shows a clear negative correlation with geodesic distance (Fig. 6(b),(c)), which is also reflected in the Pearson correlation trends. This indicates that the representation captures intrinsic geometric structure.

Overall, these observations suggest that SGSoft integrates semantic, geometric, and spatial cues into a unified feature space, where semantic priors from Uni3D contribute to global alignment, and the geodesic correspondence field encourages local consistency.

F. Visualization of Ablation Studies

Figure 7 presents qualitative results for the ablation experiments reported in the main paper. Removing the geodesic correspondence field $\tilde{S}$ (a) leads to noticeable errors in locally structured areas such as the face, where fine geometric details are critical. Disabling contrastive supervision (b) results in severe performance degradation and inconsistent color transfer across parts.

Further removing geodesic grouping/ungrouping or geodesic encoding (c) causes significant misalignment, often producing color bleeding between adjacent semantic regions. Finally, removing the symmetry loss leads to frequent left-right flips in symmetric body parts.

In contrast, the full SGSoft model produces sharp, part-consistent correspondences with clear semantic boundaries, visually corroborating the quantitative trends reported in the ablation study of the main paper.

G. Preliminary Study on 3D Backbone Selection

G.1. Evolution of 3D Shape Representation

Recent advances in 3D shape representation learning have followed different directions depending on the input modality and output granularity. Early mesh-based methods directly operate on surface connectivity and geometry (e.g., MeshNet [5], MeshCNN [8], DiffusionNet [16]). While these approaches preserve structural information, they often require heavy preprocessing and are less scalable for large-scale semantic learning.

Mesh representations can be readily converted into point clouds, and many recent works therefore focus on point-based architectures that offer higher flexibility and better scalability. Foundational models such as PointNet [13] and PointNet++ [14] introduced permutation-invariant and locality-aware feature extraction. Transformer-based encoders (e.g., PCT [6], Point-BERT [19]) further enhanced global context modeling through self-attention, but are mostly trained with geometry-oriented objectives.

More recently, multimodal 3D foundation models such as ULIP [18], Point-Bind [7], and PointLLM [17] align 3D features with vision and language through contrastive learning, enabling more semantically informed representations.

G.2. Rationale for Choosing Uni3D as Backbone

Among recent multimodal 3D foundation models, **Uni3D** [9] is a large-scale point-cloud foundation model trained with multimodal supervision. Although it is not specifically designed for dense surface correspondence, Uni3D provides strong object-level semantic features and structured point-wise representations, as demonstrated by its performance on multiple 3D recognition and segmentation benchmarks.

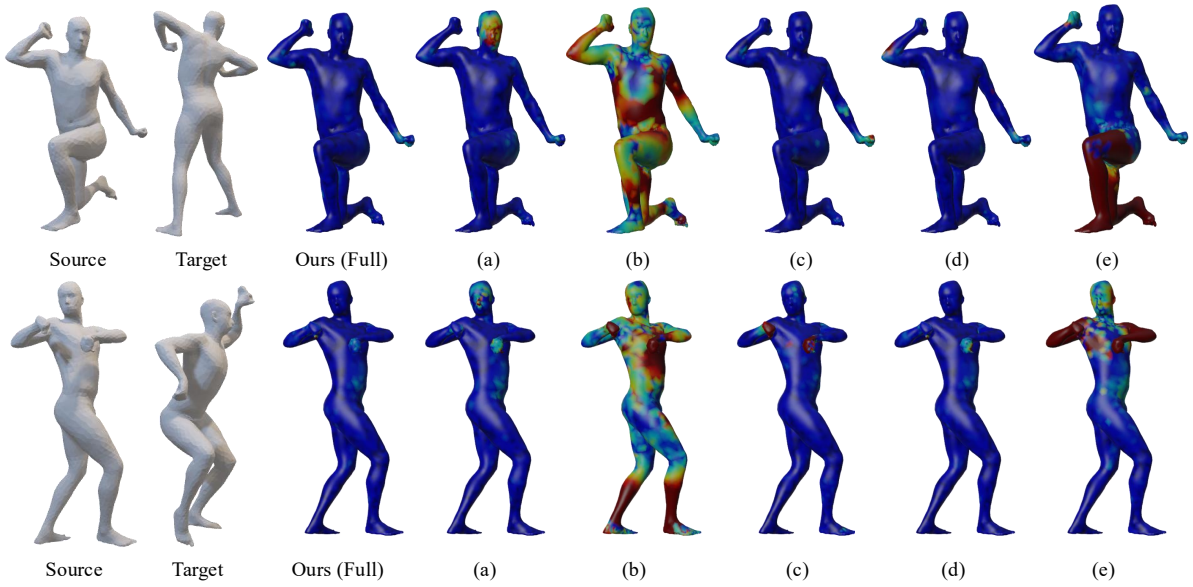


Figure 7. **Visualization of ablation studies.** From left to right: source mesh, target mesh, SGSoft (full), and ablated models (a) w/o geodesic correspondence field $\tilde{S}$, (b) w/o contrastive loss on $\tilde{S}$, (c) w/o geodesic grouping & ungrouping, (d) w/o geodesic encoding, and (e) w/o symmetry loss. Removing each component produces visible artifacts such as color bleeding across parts, symmetry flips, and misalignment near joints, whereas the full model yields sharp and part-consistent correspondences.

Table 3. **Rationale for choosing Uni3D as the semantic backbone.** Comparison between SGSoft design requirements and the properties supported by Uni3D.

Design requirement for SGSoft	Required	Supported by Uni3D
Object-level semantic priors	✓	✓
Point-wise part-level semantics	✓	✓
Direct encoding from raw 3D geometry	✓	✓
Generalization across object categories and domains	✓	✓

Moreover, its representations are learned directly from 3D geometry without relying on rendered views, which helps avoid viewpoint-dependent artifacts and preserves local neighborhood relationships in the feature space. This spatial consistency is particularly important for correspondence learning, where stable geometric neighborhoods must be preserved under large deformations.

Therefore, we adopt Uni3D as the semantic backbone of SGSoft, given its multimodal pretraining on large-scale 3D data, spatially discriminative point-wise feature representations, and stable performance across diverse 3D benchmarks. As summarized in Table 3, these properties align well with the design requirements of SGSoft for semantically consistent and geometry-aware correspondence learning.

H. Additional Comparison and Uni3D Ablation

H.1. Additional Comparison with Other Methods

We further provide qualitative comparisons with ULRSSM [2] and NICP [11], which represent classical unsupervised spectral matching and non-rigid registration approaches, respectively. ULRSSM is an unsupervised spectral matching method, while NICP is designed under relatively strong assumptions on shape similarity and initialization.

	DT4D-Inter (↓)
ULRSSM	15.34 *(Faust)*
	22.89 *(Scape)*
	12.10 *(faust+Scape)*
Ours	**8.3**

(a) ULRSSM vs Ours (quantitative)

(b) NICP Qualitative Failure

Figure 8. **Comparison with ULRSSM (left) and NICP (right).** SGSoft achieves lower error and more consistent correspondences under cross-domain and large deformation settings. ULRSSM is not designed for zero-shot correspondence, and NICP follows a different registration objective.

As shown in Fig. 8, SGSoft produces more stable and semantically consistent correspondences under large deformations and cross-category settings. In contrast, ULRSSM shows limited generalization across domains, and NICP often struggles when the input shapes deviate from its underlying assumptions.

We note that these methods are not specifically designed for zero-shot cross-category correspondence. Nevertheless, we include these comparisons to provide a broader perspective on the behavior of existing approaches under such challenging conditions.

H.2. Effect of Uni3D Semantic Features

We further analyze the contribution of semantic features from Uni3D. Ablating the semantic features from Uni3D (w/o Uni3D) leads to a drastic performance drop across all benchmarks (e.g., 2.9→68.0 on SCAPE), indicating that semantic anchoring is essential for resolving global ambiguities and enabling stable correspondence.

Table 4. Ablation study on Uni3D semantic features.

Method	SCAPE	SHREC19	DT4D-Inter
w/o Uni3D	68.0	70.6	69.2
Full (SGSoft)	2.9	4.0	8.3

I. Downstream Applications

Semantic Segmentation. We evaluate semantic label transfer in the SGSoft descriptor space. As shown in Fig. 1 of the main paper, our descriptors preserve part-level coherence and disambiguate left-right symmetry, enabling reliable cross-instance label propagation across different poses and shapes. These results indicate that SGSoft provides a stable and transferable semantic representation that can serve as a generic backbone for downstream segmentation tasks.

Deformation Transfer. Using our predicted correspondences, we transfer source deformations to unseen targets. SGSoft yields smoother deformation across joints and avoids typical artifacts such as stretching. Figure 9 shows two example deformations (A and B) applied to a source and then transferred to a target. The results retain the intended motion while respecting the target's shape and proportions, indicating that our correspondences are precise enough to drive high-quality deformation transfer.

J. Limitations and Future Work

We identify three representative failure modes, as illustrated in Fig. 10. First, semantic conflicts arise in shapes with extreme proportions or additional structures (e.g., accessories), where semantic priors become ambiguous. Second, spatial ambiguity occurs when multiple parts are in close proximity, leading to confusion despite correct global context. Third, in severely partial shapes, missing connectivity degrades the reliability of geodesic distances, reducing global alignment accuracy. Overall, SGSoft remains robust under moderate deformation and partiality, with failures primarily occurring when both semantic and geometric cues are unreliable.

While SGSoft demonstrates strong cross-domain generalization, it is still influenced by the choice of canonical template used during training. In particular, when input shapes exhibit significantly different proportions or structural characteristics from the training domain, the shared reference space induced by the template becomes less optimal.

Extending the framework to support multiple or more diverse templates, or learning a template-agnostic correspondence space, would further improve robustness to large structural variations. We leave this as an important direction for future work.

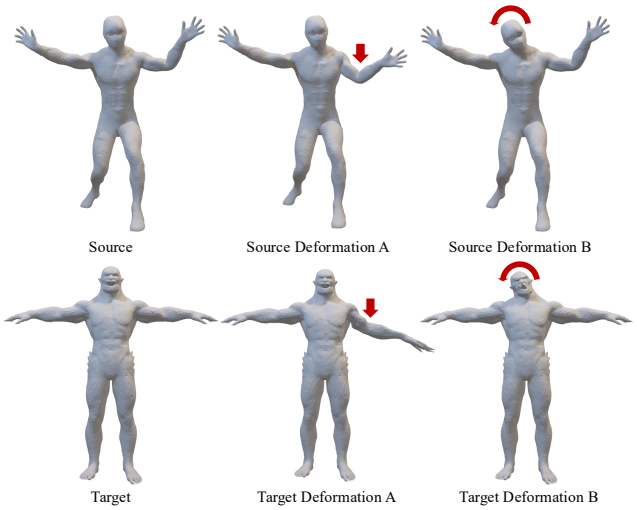


Figure 9. **Deformation transfer with SGSoft correspondences.** Given a source mesh and two source deformations (A and B), we propagate the deformations to an unseen target using our predicted correspondences. The transferred motions preserve the intended pose changes while adapting smoothly to the target's body shape.

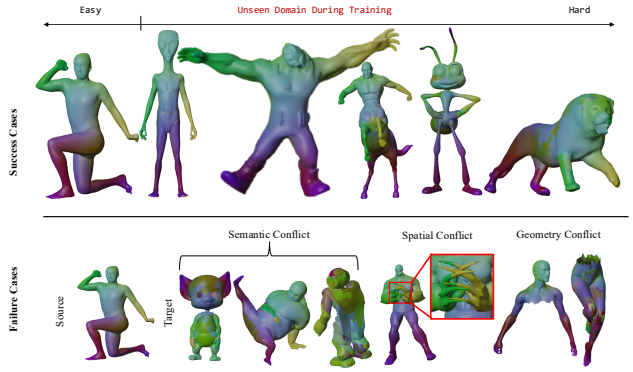


Figure 10. **Success and failure cases of SGSoft across increasing domain shift.** Top: Successful correspondences from seen to unseen domains with progressively increasing difficulty. Bottom: Representative failure modes. (1) Semantic conflict under extreme proportions or additional structures, (2) spatial ambiguity when parts are in close proximity (zoomed), (3) geometric degradation on severely partial shapes due to unreliable geodesic distances.

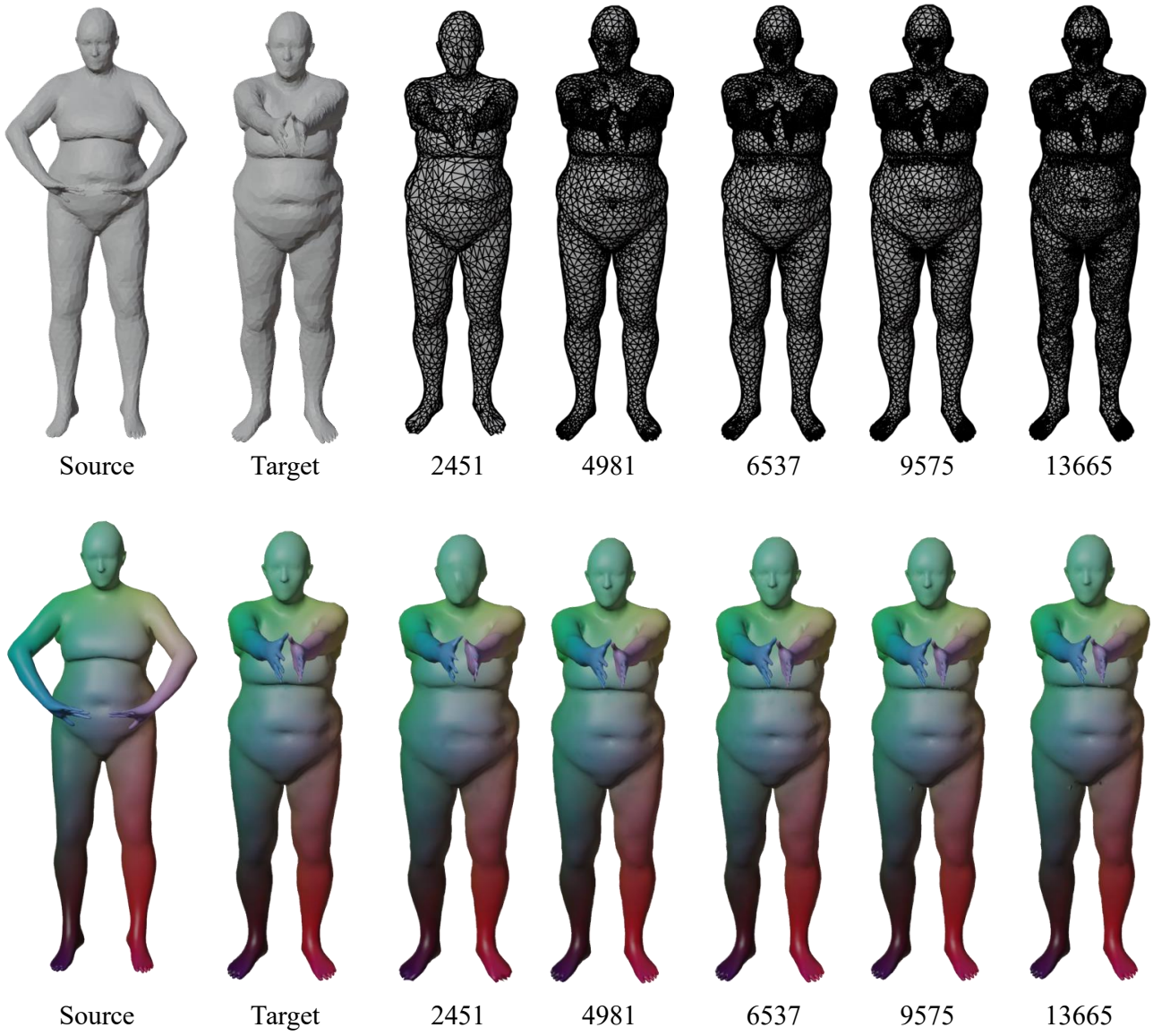


Figure 11. **Topology robustness.** Colors are transferred from a source to targets with different mesh resolutions. Numbers denote vertex counts. SGSoft preserves smooth and semantically consistent correspondences across topologies.

References

[1] Noam Aigerman, Kunal Gupta, Vladimir G Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. *arXiv preprint arXiv:2205.02904*, 2022. 3

[2] Dongliang Cao, Paul Roetzer, and Florian Bernard. Unsupervised learning of robust spectral shape matching. *arXiv preprint arXiv:2304.14419*, 2023. 6

[3] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):1–11, 2013. 1

[4] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4494–4504, 2024. 3

[5] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: Mesh neural network for 3d shape representation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8279–8286, 2019. 5

[6] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational visual media*, 7(2):187–199, 2021. 5

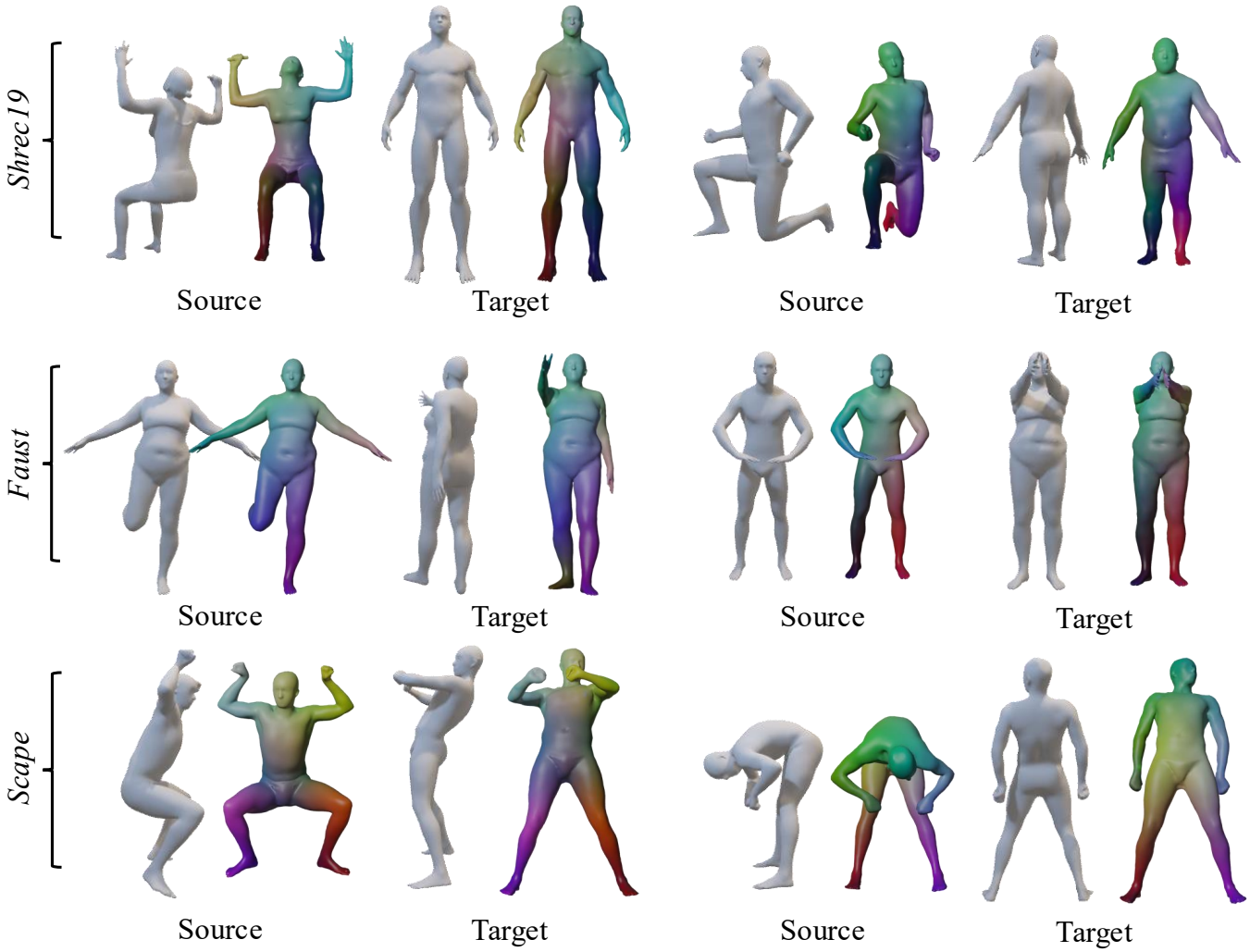


Figure 12. **Symmetry disambiguation.** We visualize correspondence transfer from a source mesh to target meshes on FAUST, SCAPE, and SHREC'19. SGSoft consistently aligns left/right limbs and resolves front-back ambiguity under large articulations, avoiding symmetric flips. The uncolored source and target meshes are shown in their original poses, while the colorized meshes are rotated to a front view for clearer visualization of correspondence quality.

[7] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 5

[8] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 5

[9] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. *arXiv preprint arXiv:2402.03327*, 2024. 5

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 3

[11] Riccardo Marin, Enric Corona, and Gerard Pons-Moll. Nicp: neural icp for 3d human registration at scale. In *European Conference on Computer Vision*, pages 265–285. Springer, 2024. 6

[12] Emery Pierson, Lei Li, Angela Dai, and Maks Ovsjanikov. Diffumatch: Category-agnostic spectral diffusion priors for robust non-rigid shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5745–5756, 2025. 3

[13] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5

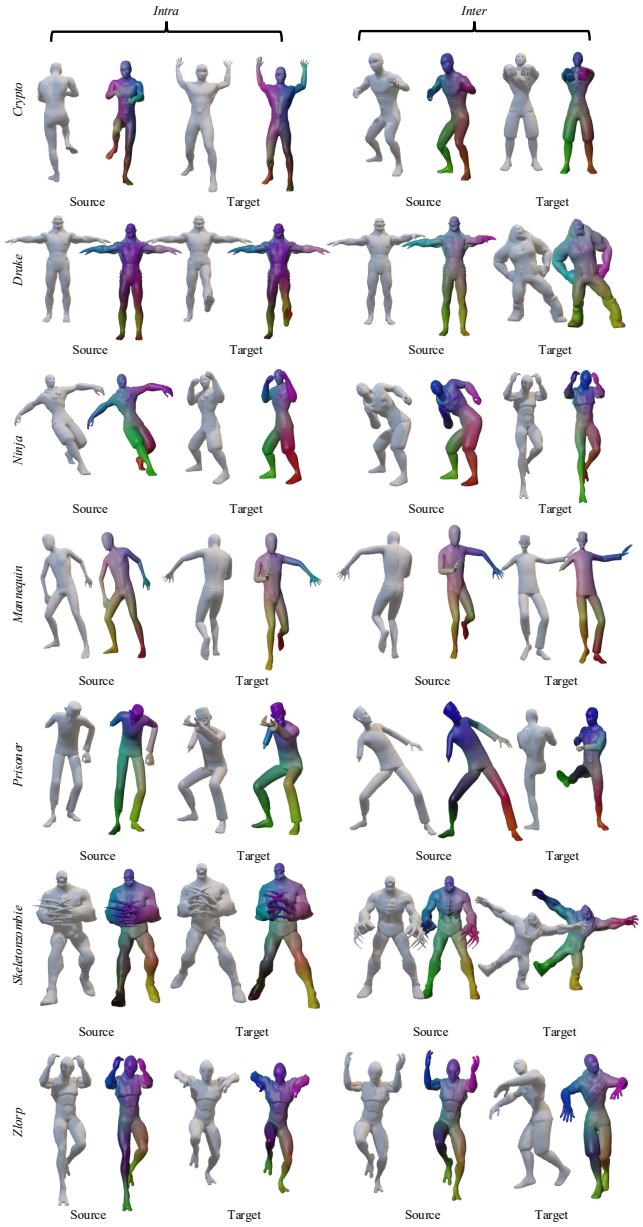


Figure 13. **Zero-shot cross-domain generalization.** We train SGSoft only on SMPL-like human bodies and evaluate on stylized rigs from the Mannequin, Drake, Ninja, Prisoner, Skeleton, Zombie, and Crypto Zlorp families. *Intra* (left) shows correspondences within each character family, while *Inter* (right) transfers correspondences across different families. Despite substantial differences in shape and articulation, SGSoft preserves consistent part-wise color patterns without any finetuning.

[14] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[15] Szymon Rusinkiewicz. Estimating curvatures and their derivatives on triangle meshes. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 486–493. IEEE, 2004. 1

[16] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3): 1–16, 2022. 5

[17] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiang-

miao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024. 5

[18] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1179–1189, 2023. 5

[19] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 5

[20] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 2

[21] Aleksei Zhuravlev, Zorah Lähner, and Vladislav Golyanik. Denoising functional maps: Diffusion models for shape correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26899–26909, 2025. 3