RESEARCH-ARTICLE

Surface-Aware Distilled 3D Semantic Features

LUKAS UZOLAS, Delft University of Technology, Delft, Zuid-Holland, Netherlands

ELMAR EISEMANN, Delft University of Technology, Delft, Zuid-Holland, Netherlands

PETR KELLNHOFER, Delft University of Technology, Delft, Zuid-Holland, Netherlands

Published: 15 December 2025

SA Conference Papers '25: SIGGRAPH Asia 2025 Conference Papers
December 15 - 18, 2025
Hong Kong, Hong Kong

Conference Sponsors:
SIGGRAPH

SA Conference Papers '25: Proceedings of the SIGGRAPH Asia 2025 Conference Papers (December 2025)
https://doi.org/10.1145/3757377.3763974
ISBN: 9798400721373

Surface-Aware Distilled 3D Semantic Features

LUKAS UZOLAS, Delft University of Technology, Netherlands
ELMAR EISEMANN, Delft University of Technology, Netherlands
PETR KELLNHOFER, Delft University of Technology, Netherlands

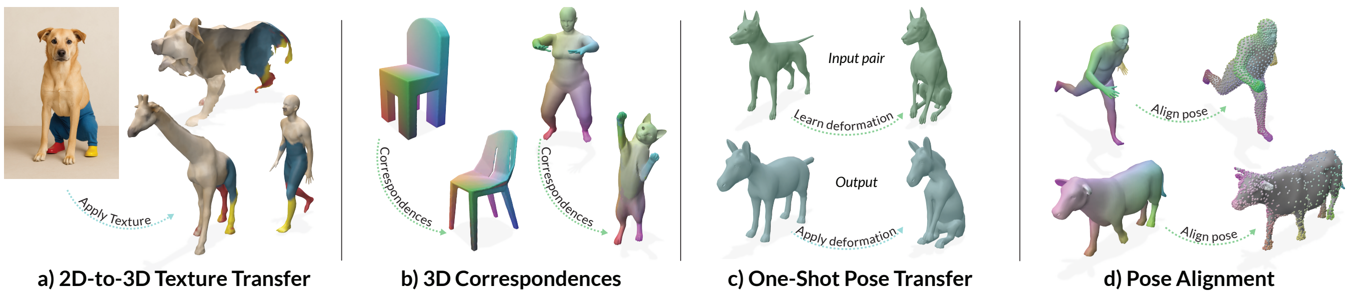

Fig. 1. We introduce a *surface-aware* feature embedding space based on 2D pre-trained foundational models. In contrast to related works, our embedding space separates instances of the same semantic class (e.g. right vs. left instances for "hand/paw") which facilitates many downstream applications: a) Texturing of even incomplete 3D shapes based on a 2D image (the input image was produced by ChatGPT), b) 3D correspondences between non-isometric shapes, c) Re-posing of meshes based on a single source and target pair, d) Pose alignment of 3D meshes with dense and sparse point correspondences.

Many 3D tasks such as pose alignment, animation, motion transfer, and 3D reconstruction rely on establishing correspondences between 3D shapes. This challenge has recently been approached by pairwise matching of semantic features from pre-trained vision models. However, despite their power, these features struggle to differentiate instances of the same semantic class such as "left hand" versus "right hand" which leads to substantial mapping errors. To solve this, we learn a surface-aware embedding space that is robust to these ambiguities while facilitating shared mapping for an entire family of 3D shapes. Importantly, our approach is self-supervised and requires only a small number of unpaired training meshes to infer features for new possibly imperfect 3D shapes at test time. We achieve this by introducing a contrastive loss that preserves the semantic content of the features distilled from foundational models while disambiguating features located far apart on the shape's surface. We observe superior performance in correspondence matching benchmarks and enable downstream applications including 2D-to-3D and 3D-to-3D texture transfer, in-part segmentation, pose alignment, and motion transfer in low-data regimes. Unlike previous pairwise approaches, our solution constructs a joint embedding space, where both seen and unseen 3D shapes are implicitly aligned without further optimization. The code is available at https://graphics.tudelft.nl/SurfaceAware3DFeatures.

CCS Concepts: • **Computing methodologies → Shape analysis**; Dimensionality reduction and manifold learning; Motion processing.

Additional Key Words and Phrases: Semantic Features, Contrastive Learning, Motion Transfer, Reposing, Shape Correspondences

Authors' Contact Information: Lukas Uzolas, Delft University of Technology, Delft, Netherlands, l.uzolas@tudelft.nl; Elmar Eisemann, Delft University of Technology, Delft, Netherlands, e.eisemann@tudelft.nl; Petr Kellnhofer, Delft University of Technology, Delft, Netherlands, p.kellnhofer@tudelft.nl.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763974

ACM Reference Format:
Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. 2025. Surface-Aware Distilled 3D Semantic Features. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757377.3763974

1 Introduction

Mapping 3D shapes into a shared space guaranteeing mutual correspondences is important for many applications, including 3D registration, pose alignment, motion transfer, as well as static and dynamic 3D reconstruction. Historically, geometric descriptors have been used to determine matches between pairs of 3D shapes under isometric deformations, but they struggle with non-isometric deformations [Aubry et al. 2011; Sun et al. 2009; Tombari et al. 2010]. In contrast, neural features, stemming from pre-trained 2D vision models, have recently achieved great success in identifying correspondences between pairs of vastly different shapes [Luo et al. 2023; Tang et al. 2023; Wimmer et al. 2024; Zhang et al. 2024b], such as mapping from cats to lions. In this paper, we make another important step by moving from one-to-one pairwise shape correspondence matching to a joint embedding space establishing many-to-many shape correspondences.

Despite their inter-class robustness, neural features often struggle to disambiguate between instances of the same class like "left hand" and "right hand" (see Fig. 3). Such mismatches can lead to substantial errors in downstream applications (see Sec. 6). Recent research has demonstrated that these features contain global pose information and that disambiguation is possible in a 2D scenario [Zhang et al. 2024a]. However, achieving the same effect on distilled 3D features is not trivial, especially in a low-data regime, which is prevalent in 3D, where data acquisition and labeling is difficult.

Our work improves 3D neural features distilled from pre-trained 2D vision models by embedding them into a space disambiguating intraclass instances. We achieve this without large annotated datasets using a self-supervised learning scheme guided by in-shape geodesic distances without the need for shape pairs. Training with a limited number of 3D meshes, our method produces space of *surface-aware features* establishing multi-faceted correspondences for diverse new shapes without any further fine-tuning. In quantitative and qualitative comparisons to prior work, we demonstrate superior suitability of these features to serve as robust descriptors for matching and as building blocks for solving other tasks. Since geodesic distances are not used during inference, our method has only a minimal overhead from its shallow neural encoder and its point-wise nature makes it robust to varying mesh complexity or shape incompleteness. Finally, the encoder preserves compatibility with per-pixel image features, and hence, also naturally establishes robust 2D-to-3D mappings.

In summary, we make the following contributions: 1. We introduce a novel contrastive loss for self-supervised distillation of 3D features. 2. We quantitatively demonstrate the effectiveness of our surface-aware features in pose transfer, correspondence matching and skinning weight regression. 3. We showcase versatility of our approach in additional downstream applications including pose alignment, instance-based part segmentation and 2D-to-3D or 3D-to-3D texture transfer. 4. We show the versatility of our features when matching many-to-many shapes of not only humanoids and animals but also other classes.

2 Related Works

Our method utilizes contrastive learning to embed semantic features from foundational models to a space enabling robust n-to-n 3D shape matching. In this section, we discuss prior work in these three areas.

2.1 Image-based features for 3D shapes

Image-based features emerge in large visual models for 2D image tasks. Self-supervised features from Vision Transformers, such as DINO-ViT [Caron et al. 2021] and DINOv2 [Oquab et al. 2024], locally encode semantic information useful for segmentation [Caron et al. 2021] or image-to-image correspondence matching [Amir et al. 2021]. SD-DINO [Zhang et al. 2023a] adds complementary features from the diffusion-based image synthesis model Stable Diffusion [Rombach et al. 2022]. Lifting these features to 3D has enabled the self-supervised construction of canonical surface maps [Shtedritski et al. 2024], transfer of appearance between 3D shapes [Fischer et al. 2024], 3D animation [Uzolas et al. 2024], keypoint detection [Wimmer et al. 2024] or matching of surface correspondences [Chen et al. 2025; Dutt et al. 2024; Morreale et al. 2024]. However, despite their semantic versatility, disambiguating between intraclass instances, such as left and right hands, remains challenging but possible, as shown in a recent 2D image study [Zhang et al. 2024a]. This motivates our 3D shape descriptors for resolving instance ambiguity.

Prior work tackled this ambiguity by mapping shapes to a spherical template [Mariotti et al. 2024], which is difficult for complex shapes including humans. Alternatively, Liu et al. [2025] recovered non-isometric correspondences from a 2D semantic flow learned from vision features. Instead, we adapt Diff3F features [Dutt et al. 2024] in 3D space and resolve the ambiguity through contrastive learning enforcing geodesic distances.

Geodesic distances have previously supported point cloud analysis [He et al. 2019] and more recently, NIE [Jiang et al. 2023] and the concurrent work DV-Matcher [Chen et al. 2025] similarly utilize geodesic distances for feature disambiguation. Yet, the previously mentioned methods rely on aligned mesh pairs or a learned alignment, while our method learns purely from intrinsic properties of individual shapes. This eases adaptation to less common classes beyond humanoids and animals (Fig. 8) and cross-class mappings (Fig. 9).

Beyond vision-only models, multimodal large language models have recently been effective in image and 3D shape analysis including keypoint labeling [Gong et al. 2024] and shape co-segmentation [Abdelreheem et al. 2023]. In our work, we focus on vision-only models because of their simplicity, but we consider a model combination a promising research direction.

2.2 Contrastive Learning

Contrastive learning embeds similar samples close to each other while keeping dissimilar samples apart. This can be achieved directly by minimizing and maximizing embedding distances for positive and negative pair samples, respectively [Chopra et al. 2005; Hadsell et al. 2006; Schroff et al. 2015; Weinberger and Saul 2009] or indirectly, such as by optimizing performance in an auto-regressive task [Oord et al. 2018]. Training pairs can be obtained by data augmentation [Chen et al. 2020], from memory banks [He et al. 2020; Wu et al. 2018], or by clustering [Caron et al. 2018, 2020]. Learning with cross-domain labels yields joint embeddings, as demonstrated by CLIP [Radford et al. 2021] for text and images. Contrastive learning was applied to learn end-to-end pose transfer from multiple unregistered meshes of the same identity in different poses [Sun et al. 2023a]. We design our contrastive loss to disambiguate intraclass instances guided by a geodesic metric, while learning from intrinsic properties of individual meshes rather than same-identity shape pairs.

2.3 Shape correspondences

Point-to-Point. Classical shape registration methods directly minimize global [Besl and McKay 1992] or local [Brown and Rusinkiewicz 2007] inter-shape distances making them susceptible to local minima [Yang et al. 2015]. This motivates the design of more informative local geometric descriptors [Aubry et al. 2011; Sun et al. 2009; Tombari et al. 2010]. These can alternatively be learned [Corman et al. 2014; Guo et al. 2015] from voxelized patches [Attaiki et al. 2023; Gojcic et al. 2019; Zeng et al. 2017] or from point clouds [Deng et al. 2018a,b, 2023; Elbaz et al. 2017; Yew and Lee 2018]. The learning can be supervised by labels [Corman et al. 2014] or achieved without them [Elbaz et al. 2017; Groueix et al. 2018; Lang et al. 2021; Zeng et al. 2021]. Our method falls into the latter category, as our contrastive loss motivates our encoder to separate instances by approximating geodesic distances [Xia et al. 2021] without training data labels. This is conceptually similar to previous methods for near-isometric shape deformations [Halimi et al. 2019; Mémoli and Sapiro 2005; Shamai and Kimmel 2017]. However, we distinctly do not measure geodesic distortions between shape pairs, and therefore we do not limit our method to isometric deformations, and we do not compute any geodesics during inference. Instead, we only use the geodesics to disambiguate information already available in the image-based features, which is critical for our results.

The correspondences can be recovered from descriptors by a matching [Fischler and Bolles 1981], directly regressed [Lu et al. 2019; Wang and Solomon 2019] or established on parametric templates [Deprelle et al. 2019; Groueix et al. 2018]. Here, we focus on the descriptors themselves, and we show several different application scenarios in Sec. 6.

Surface mapping. Functional Maps (FMs) [Ovsjanikov et al. 2012] allow for matching on a surface. FMs are real-valued surface functions in the space of Laplace-Bertrami eigenfunctions, supporting linear transformations between shapes. Constrained to match surface descriptors for each shape [Aubry et al. 2011; Sun et al. 2009; Tombari et al. 2010] they allow extracting point-wise correspondences [Ovsjanikov et al. 2012; Rodolà et al. 2015]. These functions can also be learned [Litany et al. 2017] often with little or no supervision [Donati et al. 2020; Ginzburg and Raviv 2020; Halimi et al. 2019; Roufosse et al. 2019; Sun et al. 2023b]. Extrinsic alignment can support nonisotropic deformations [Eisenberger et al. 2020a,b]. In this work, we focus on improving features for direct point-to-point matching in the spatial domain, but we later demonstrate a combination of our features with FM.

3 Preliminaries

We build upon methods that aggregate features from pre-trained 2D vision models on 3D meshes [Chen et al. 2025; Dutt et al. 2024; Morreale et al. 2024; Wimmer et al. 2024]. In this section, we give a brief overview on these methods.

3.1 Reprojection of 2D Features

We represent a 3D shape as a triangular mesh with a tuple of N vertices and M triangular faces, that is, $\mathcal{M} := (\{\mathbf{p}_n \in \mathbb{R}^3 | n = 1, ..., N\}, \{\mathbf{t}_m \in \mathbb{N}^3 | m = 1, ..., M\})$. The rendering function $R_{rgb} : (\mathcal{M}, C) \to \mathbf{I}_{rgb}$ projects $\mathcal{M}$ into a camera C and outputs an image $\mathbf{I}_{rgb} \in \mathbb{R}^{H \times W \times 3}$, with height H and width W. Optionally, texturing is possible in $R_{rgb}(.)$ or as a ControlNet [Zhang et al. 2023b] post-processing. The image is then passed to a pre-trained vision model [Caron et al. 2021; Oquab et al. 2024; Rombach et al. 2022; Zhang et al. 2023a] to obtain dense semantic feature maps $\mathbf{F} \in \mathbb{R}^{h \times w \times f}$ with h, w, f as two spatial and one feature dimension. Finally, per-vertex features $\mathbf{f}_n \in \mathbb{R}^f$ are obtained by projective texture mapping of $\mathbf{F}$ onto $\mathcal{M}$. To cover the whole surface, features are aggregated across multiple cameras, resulting in a set of features $\mathcal{F}_\mathcal{M} := \{\mathbf{f}_n \in \mathbb{R}^f | n = 1, ..., N\}$. Throughout this work, we refer to $\mathcal{F}_\mathcal{M}$ as the *base features* on which our method is built. The exact choice of $\mathcal{F}_\mathcal{M}$ is orthogonal to our contribution but must encode semantic information. To this extent, we use Diff3F [Dutt et al. 2024] in this work.

Correspondence Matching. Features $\mathcal{F}_\mathcal{M}$ have been shown to encode strong semantic information useful for correspondence matching [Dutt et al. 2024; Tang et al. 2023]. In the simplest case, the feature $\mathbf{f}_n \in \mathcal{F}_\mathcal{T}$ of a target mesh $\mathcal{T}$ that best matches the feature $\mathbf{f}_m \in \mathcal{F}_\mathcal{S}$ of a source mesh $\mathcal{S}$ is determined by maximizing the cosine similarity $\phi : \mathbb{R}^f \times \mathbb{R}^f \to \mathbb{R}$:

$$\phi(\mathbf{f}_i, \mathbf{f}_j) = \frac{\mathbf{f}_i^T \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \|\mathbf{f}_j\|_2}, \tag{1}$$

such that $\tau(\mathbf{p}_m) = \arg\max_{\mathbf{p}_n} \phi(\mathbf{p}_n \to \mathbf{f}_n, \mathbf{p}_m \to \mathbf{f}_m)$ is the best matching point. However, the features $\mathcal{F}_\mathcal{M}$ do not differentiate between semantic instances well (see Fig. 6) which we address by learning robust *surface-aware features* $\mathcal{S}_\mathcal{M}$.

4 Method

Our goal is to learn an embedding resolving instance ambiguities of the *base features* $\mathcal{F}_\mathcal{M}$ and obtain *surface-aware features* $\mathcal{S}_\mathcal{M}$ (see Fig. 2). We achieve this by training a point-based feature autoencoder with a limited set of training meshes and our contrastive loss for self-supervision. At test time, we can produce *surface-aware features* for novel unseen shapes without additional fine-tuning.

4.1 Setup

Our method requires a potentially small set of training meshes $\mathbf{M}_t = \{\mathcal{M}_i | i = 1, ..., K\}$, each associated with *base features* $\mathcal{F}_\mathcal{M}$ obtained following Sec. 3 and normalized by a Euclidean norm such that $\forall \mathbf{f}_n \in \mathcal{F}_\mathcal{M}, \|\mathbf{f}_n\|_2 := 1$.

Unlike other approaches [Deng et al. 2023; Jiang et al. 2023], we do not require extrinsic canonical mesh alignment during training, because we rely only on intrinsic properties of individual meshes. Similar considerations were previously made for Functional maps [Ovsjanikov et al. 2012]. Moreover, we inherit the rotation invariance of the *base features* demonstrated by Dutt et al. [2024, Supplement "Robustness to Rotation"], although we observe a performance degradation if meshes are upside down and we avoid this in our inputs (see Appendix A.1).

Furthermore, our encoder is point-based and does not rely on shape completeness or a consistent topology. Both of these design choices favor generalization under transformations ranging from coordinate swap to shape reposing, and memory-unconstrained batch-based processing of even large shapes.

4.2 Separating Front Paw from Back Paw

Our embedding aims to separate multiple instances of the same class that are difficult to directly disambiguate in $\mathcal{F}_\mathcal{M}$. For example, consider the two surface points, $\mathbf{p}_1$ and $\mathbf{p}_2$, on the bear's paws in Fig. 2. The prevalent semantic significance of the "paw" concept hinders the separability of the corresponding *base features* $\mathbf{f}_1$ and $\mathbf{f}_2$. Fig. 3 illustrates this for human arms and Diff3F features [Dutt et al. 2024]. To solve this, we train a point-wise feature autoencoder, producing our *surface-aware features* $\mathcal{S}_\mathcal{M} \subset \mathbb{R}^s$ in its embedding space. We motivate the feasibility of separation by the prior observations that vision features additionally carry information about the global pose [Zhang et al. 2024a]. We postulate that this enables our model to distinguish between part instances when guided by their intrinsic distance. We adopt the geodesic distance $d_{1,2}$ between $\mathbf{p}_1$ and $\mathbf{p}_2$ for this purpose. Following contrastive learning, we sample point pairs on a single shape to enforce $\phi(\mathbf{s}_1, \mathbf{s}_2) \approx d_{1,2}$ for $\mathbf{s}_n \in \mathcal{S}_\mathcal{M}$.

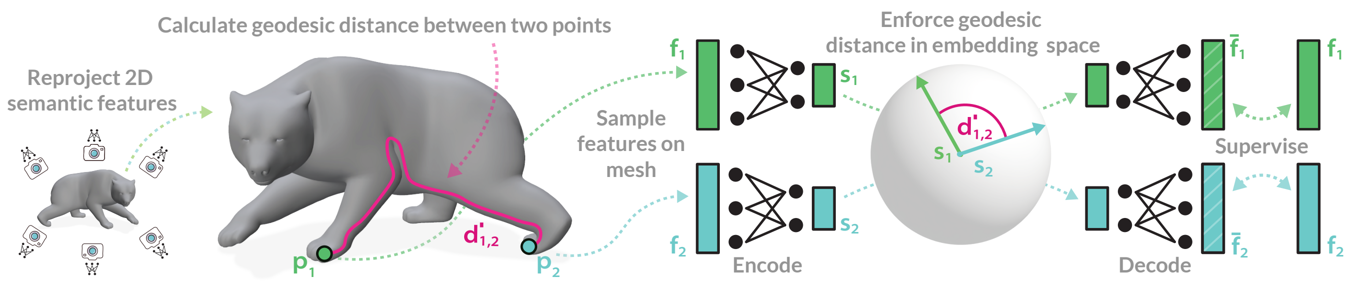

Fig. 2. Overview of our method. We feed images of a 3D shape rendered from multiple viewpoints to a pre-trained 2D vision model and extract features that are then projected back onto surface points $\mathbf{p}_i$ and aggregated into per-point features $\mathbf{f}_i$ (Sec. 3). Next, we pointwise embed the *base features* $\mathbf{f}_i$ into our *surface-aware features* $\mathbf{s}_i$ residing in a lower-dimensional space learned using our contrastive loss preserving geodesic distances $d_{i,j}$ and a reconstruction loss matching decoded features $\bar{\mathbf{f}}_i$ to $\mathbf{f}_i$ (Sec. 4). The *surface-aware features* $\mathbf{s}_i$ serve as robust descriptors for correspondence matching (Sec. 5) and base blocks for many down-stream applications (Sec. 6).

We validate our choice of hyperspherical embedding space against Euclidean space in Sec. 5.5.

Model. We train a *base feature* encoder $\mathcal{E}(.)$, such that, following a normalization, $\mathbf{s}_n = \mathcal{E}(\mathbf{f}_n)/\|\mathcal{E}(\mathbf{f}_n)\|_2$ is a *surface-aware feature* $\mathbf{s}_n \in \mathbb{R}^s$ in a hypersphere embedding. During training, we randomly sample an unpaired training mesh $\mathcal{M} \in \mathbf{M}_t$ with *base features* $\mathcal{F}_{\mathcal{M}}$, which we encode pointwise to obtain $\mathcal{S}_{\mathcal{M}}$.

In each training iteration, we use furthest-point sampling to choose a random subset of A anchor points $\mathbf{p}_a$ among the mesh vertices $\mathbf{p}_i \in \mathcal{M}$ and compute geodesic distances $d_{n,a}$ for each pair of a mesh and anchor point. We additionally rescale $d_{n,a}$ to a maximum of one, such that $d'_{n,a} := d_{n,a}/\max_{n,a}(d_{n,a})$, which removes the dependency on the scale of the mesh. We find this robust and leading to features later generalizing across morphologically equivalent shapes with different proportions (see an elephant vs. a giraffe in Fig. 8).

Subsequently, our contrastive loss preserves the rescaled geodesic distances in the embedding space:

$$\mathcal{L}_c = \frac{1}{NA}\sum_{n}^{N}\sum_{a}^{A}\left|d'_{n,a} - \left(\frac{1-\phi(\mathbf{s}_n, \mathbf{s}_a)}{2}\right)\right|. \quad (2)$$

This loss operates in a hyperspherical embedding space and utilizes cosine similarity mapped to the $[0, 1]$ range. Hereby, $\mathcal{L}_c$ penalizes features close in the embedding space but distant on the shape surface and vice versa.

Furthermore, we found it beneficial for the preservation of semantic information to train a feature decoder $\bar{\mathbf{f}}_n = \mathcal{D}(\mathbf{s}_n)/\|\mathcal{D}(\mathbf{s}_n)\|_2$ in an autoencoder fashion. To this extent, we utilize a reconstruction loss:

$$\mathcal{L}_r = \frac{1}{N}\sum_{n} 1 - \phi(\mathbf{f}_n, \bar{\mathbf{f}}_n). \quad (3)$$

We train both the encoder and the decoder end-to-end with the combined loss $\mathcal{L} = w_r\mathcal{L}_r + w_c\mathcal{L}_c$, where a choice $w_r = w_c = 1$ works well in our tests. We do not observe an increase in performance with a higher w_c.

Note that our training procedure, in contrast to related works [Chen et al. 2025; Deng et al. 2023; Lang et al. 2021], does not require target and source shape pairs.

4.3 Implementation

During preprocessing, we rasterize our triangular meshes and precompute *base features* for all vertices following Sec. 3. We implement our autoencoder in PyTorch2 [Ansel et al. 2024] and use the Polyscope renderer [Sharp et al. 2019b] for visualizations. The encoder $\mathcal{E}$ is a Multilayer Perceptron (MLP) consisting of three blocks, where each block has two linear layers, SiLU activation [Elfwing et al. 2018], and layer normalization [Ba et al. 2016]. The first layer in each block employs a skip connection [He et al. 2016], while the second reduces the dimensionality by a factor of two. With Diff3F [Dutt et al. 2024] as *base features*, $\mathcal{E}$ reduces feature dimensionality from $f = 2048$ to $s = 256$. The decoder $\mathcal{D}$ is a mirrored copy of the encoder. We train our model on NVIDIA RTX 3090 for 50k iterations with the AdamW optimizer [Loshchilov and Hutter 2017] and a learning rate of 0.0001 which takes $\approx$ 2 hours.

We choose an exponential moving average [Polyak and Juditsky 1992] of the model with the lowest validation loss, without the need for any correspondence labels. Geodesic distances for training are calculated on the fly with the heat method [Crane et al. 2017] implemented in Geometry Central [Sharp et al. 2019a]. No geodesics are required during inference and the computational cost is determined by the Diff3F baseline with a only a negligible overhead from our shallow encoder $\mathcal{E}$. For a shape with 10k vertices, this is less than 5 milliseconds on top of $\approx$ 4 minutes from Diff3F. Moreover, downstream tasks cost benefits from the smaller feature dimensionality. Functional maps are calculated with the base algorithm [Ovsjanikov et al. 2012], provided by the Diff3F implementation [Dutt et al. 2024].

5 Experiments

Here, we first motivate the benefits of our *surface-aware feature* embedding space by visualizing its distribution. Next, we evaluate their effectiveness in tasks with quantitative benchmarks including pose transfer, skinning weight regression and 3D correspondence matching. Finally, we analyze the impact of our design choices in an ablation experiment.

Training. We train a single autoencoder on a joint dataset consisting of 49 animal samples from the *SMAL* dataset [Zuffi et al. 2017] and 49 humans from the *SURREAL* dataset [Groueix et al. 2018].

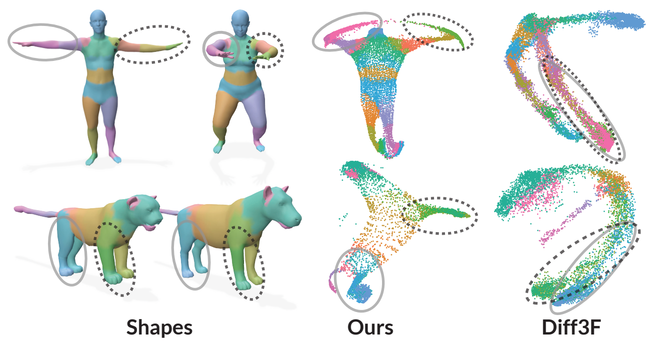

Fig. 3. Two shapes (left) and a PCA-based 2D projections of their aggregated Diff3F *base features* and our *surface-aware features* (right). Notice the separation of limbs in our result compared to Diff3F. Our features originate from the same encoder for both shapes. The animal legs appear merged along the sagittal plane due limitations of the PCA projection, but they remain disambiguated in our feature space as demonstrated in Fig. 9.

We choose 2 samples from each dataset for validation. We use this single shared model without any additional optimization for all experiments, unless stated otherwise.

5.1 Exploration of Embedding Space

To illustrate the effect of our contrastive loss on feature separation, we compare the 2D projections of the Diff3F [Dutt et al. 2024] *base features* with our *surface-aware features*.

Setup. We create two dataset $SMPL^{eval}$ and $SMAL^{eval}$ unseen during training. The former consists of 50 randomly-sampled *SMPL* [Loper et al. 2023] shapes and poses from *AMASS* [Mahmood et al. 2019], while the latter consists of 50 randomly-sampled *SMAL* [Biggs et al. 2018; Zuffi et al. 2017] shapes in canonical poses. For each sample, we obtain the *base features* and *surface-aware features* as described in Sec. 3 and Sec. 4

Embedding. We project Diff3F features aggregated from $SMPL^{eval}$ to two dimensions using principal component analysis (PCA). In Fig. 3, we visualize the projection for two selected shapes from the same dataset. We repeat this with our *surface-aware features*. To avoid bias, we derive the visualized colors from the true SMPL [Loper et al. 2023] skinning weights $\mathbf{w}_n \in \mathbb{R}^B$ for both methods, where B is the skinning weight dimension. We repeat this process for $SMAL^{eval}$. In Fig. 3, our method yields an interpretable embedding that separates the leg and hand instances for animals and humans despite not having access to extrinsic (x, y, z) point positions. This validates the suitability of our features for downstream tasks and highlights the limitations of the Diff3F *base features*.

5.2 One-shot Pose Transfer

We evaluate the performance of the *surface-aware features* in a one-shot re-posing task for arbitrary 2-manifold meshes. We use our features to fit a Neural Jacobian Field (NJF) [Aigerman et al. 2022] between poses of two input shapes and then apply it to re-pose a new target shape as described in Appendix B.

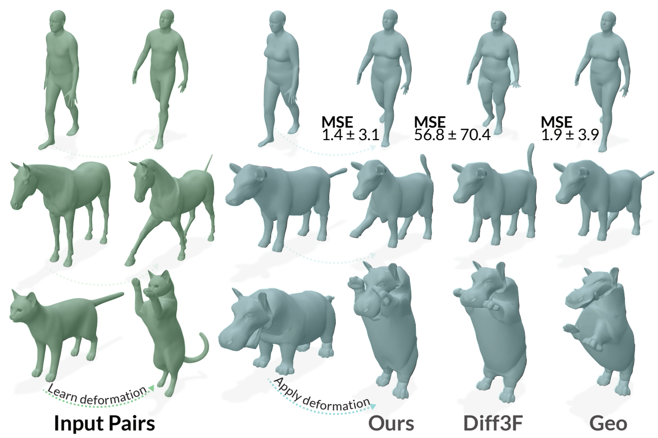

Fig. 4. One-shot pose transfer using our features, Diff3F features, or Geometric descriptors. MSE $\times 10^{-4}$ is reported for human shapes.

	$SMPL^{eval}$	$SMAL^{eval}$
Ours	**.0014±.0006**	**.0048±.0008**
Diff3F	.0058±.0024	.0086±.0028
Geo	.0210±.0031	.0287±.0017

Fig. 5. Mean Squared Error of skinning weight regression (↓ is better) and its distribution across the SMPL mesh surface.

We sample 5 input shape pairs from $SMPL^{eval}$ by choosing one challenging pose as a target and one random shape as an initial pose. The remaining shape samples serve as test inputs for pose transfer with known ground truth. We report MSE for 240 such test pairs (see upper row, Fig. 4). Next, we repeat the same procedure with the *base features* and with the Geometric descriptors (Geo) [Aigerman et al. 2022], consisting of the face centroid, face normals, and a Wave-Kernel Signature [Aubry et al. 2011]. Finally, we provide additional qualitative results for transfer of animal poses from *TOSCA* to $SMAL^{eval}$ in Fig. 4.

We observe that our *surface-aware features* outperform the *base features* both quantitatively and qualitatively by correctly distinguishing individual posed limbs. The Geometric descriptors also perform well for humans, but they struggle with larger input and output shape differences in animals. Our features perform well in both cases.

5.3 Skinning Weight Regression

We train a simple regressor to predict skinning weights of a kinematic model, based on a single training sample.

Our pointwise regressor $\mathcal{W}_\mathrm{s}(\mathbf{s}_n)$ consists of a linear layer and a Softmax activation and regresses skinning weights $\bar{\mathbf{w}}_n$ from our *surface-aware features* or, in case of $\mathcal{W}_\mathrm{f}(\mathbf{f}_n)$ and $\mathcal{W}_\mathrm{g}(\mathbf{g}_n)$, from the *base features* or the Geometric descriptors, respectively. We train all models fivefold supervised with the Mean Squared Error (MSE) and true weights separately on the $SMPL^{eval}$ and dsmalours datasets, and we report test MSE for the remaining unseen samples in the source datasets (see Fig. 5). Our features achieve lower errors and exhibit better robustness to instance ambiguities than the two alternatives. It is worth noting that the relatively lower dimensionality of the Geometric descriptors and *base features* affects the number of regressor parameters, impact of which is not studied in this experiment. However, our conclusions also hold for a 2-layer MLP regressor with an equal hidden dimensionality of 106 which matches the dimensionality of the Geometric descriptors.

5.4 Point-to-Point Correspondence Matching

Our features can be easily integrated into correspondence matching pipelines. Therefore, we replicate the evaluation setup of Diff3F [Dutt et al. 2024] and assess our *surface-aware features* in a correspondence matching task on human and animal shapes.

Data. For testing we use re-meshed versions of humans from *SHREC'19* [Donati et al. 2020; Melzi et al. 2019] and animals from both *SHREC'20* [Dyke et al. 2020] and the animal-only subset of *TOSCA* [Bronstein et al. 2008].

Baselines. We compare our method against the unsupervised image-based Diff3F method [Dutt et al. 2024], which also provides our *base features*, and against 3DCODED [Groueix et al. 2018], DPC [Lang et al. 2021] and SE-OrNet [Deng et al. 2023], which have been trained on thousands of samples, while our method is trained on less than 100 samples.

Metrics. We report commonly used point correspondence metrics for 1024 points per mesh [Deng et al. 2023; Dutt et al. 2024; Groueix et al. 2018; Lang et al. 2021][1]. The correspondence error measures a distance between the computed correspondence point $\tau(\mathbf{p}_n)$ (see Sec. 3) and the ground-truth correspondence point $\mathbf{t}_n^{gt}$: $err = \frac{1}{n}\sum_{\mathbf{p}_n \in \mathcal{S}} \|\tau(\mathbf{p}_n) - \mathbf{t}_n^{gt}\|_2^2$. The accuracy is the fraction of points with an error below a threshold $\epsilon \in [0,1]$: $acc(\epsilon) = \frac{1}{n}\sum_{\mathbf{p}_n \in \mathcal{S}} \mathbb{I}(\|\tau(\mathbf{p}_n) - t_n^{gt}\|_2 < \epsilon g)$, where g is the maximal Euclidean distance in the target shape and $\mathbb{I}(.)$ is the indicator function.

Results. We provide quantitative results in Tbl. 1 and qualitative comparisons in Fig. 6. We find that our method achieves the lowest error on *SHREC'19* and *TOSCA*, despite being trained on fewer samples than the supervised baselines. Furthermore, we outperform the Diff3F *base features* on both *SHREC* datasets in terms of accuracy. While Diff3F achieves a higher accuracy at 1% threshold in *TOSCA*, Fig. 7 shows that the accuracy of our model is higher for thresholds above ≈ 2%. This suggests that our method excels in the removal of outliers that can be caused by mismatched components. Additionally, adapting our features to Functional maps [Ovsjanikov et al. 2012] (*+ FM*) distributes the error towards a lower mean at the cost of accuracy, and maintains the beneficial comparison to Diff3F+FM.

Fig. 6 shows that Diff3F struggles to separate intraclass instances such as left and right legs. In contrast, the results confirm the effectiveness of our contrastive loss in mitigating this issue. We observe the same behavior for *SHREC'20* in Fig. 8 (*top*), which contains highly diverse animal shapes. Furthermore, our method generally produces visually smoother results (see Appendix D.1 and the supplementary video).

Table 1. Comparison of our 3D correspondence matching to prior works 3DC (3D-CODED [Groueix et al. 2018]), DPC [Lang et al. 2021], SEN (SE-OrNet [Deng et al. 2023]), and Diff3F [Dutt et al. 2024]. †) Numbers originate from [Dutt et al. 2024], *) Experiments were replicated, x) Omitted due to non-manifold meshes, + FM) Semantic features combined with Functional Maps. Accuracy is for the commonly used ϵ = 1 %. The per-column best results are bold and the second-to-best results are underlined.

		SHREC'19	TOSCA	SHREC'20
3DC†	err ↓	8.10	19.20	-
	acc ↑	2.10	0.50	-
DPC†	err ↓	6.26	<u>3.74</u>	<u>2.13</u>
	acc ↑	17.40	30.79	31.08
SEN†	err ↓	4.56	4.32	**1.00**
	acc ↑	21.41	**33.25**	31.70
Diff3F*	err ↓	1.69±1.44	4.51±5.48	5.34±10.22
	acc ↑	<u>26.25±9.30</u>	<u>31.00±15.73</u>	<u>69.50±24.99</u>
Diff3F + FM*	err ↓	1.51±1.65	x	4.44±7.87
	acc ↑	21.71±7.12	x	58.03±25.94
Ours	err ↓	<u>0.43±0.76</u>	**1.65±2.15**	3.89±8.90
	acc ↑	**28.78±9.30**	29.35±14.53	**73.97±26.47**
Ours + FM	err ↓	**0.24±0.64**	x	3.54±7.59
	acc ↑	24.83±6.80	x	63.61±24.34

Other Shapes. Our method is applicable beyond humanoid and animal shapes, which we show by training two additional encoders for a subset of 50 chairs and 50 airplanes from ShapeNet [Chang et al. 2015]. Here, we uniformly resample the mesh vertices for a better surface coverage [Wang et al. 2022].

For unseen shapes in Fig. 8, our *surface-aware features* again better distinguish same-class instances such as chair legs and airplane wings, supporting a wider applicability of our methodology. More examples are shown in Appendix D.2 and the supplementary videos.

5.5 Ablations

We motivate our design choices by ablation on various parts of our method in Tbl. 2 following the setup of Sec. 5.4.

Choice of Angular Space. We demonstrate the effectiveness of our hyperspherical embedding by replacing our contrastive loss $\mathcal{L}_c$ (Eq. 4.2) with three different options inspired by related work (see Appendix A.2). First, the Relative Geodesic Loss (*RGL*) [Jiang et al. 2023] optimizes relative distances in a Euclidean embedding. Similarly, the Naive Geodesic Loss (*NGL*) minimizes absolute distances. Finally, the Geometrical Similarity Loss (*GSL*) [Chen et al. 2025] enforces similarity of feature and surface distances in a local neighborhood. We remove feature and geodesic normalization wherever absolute magnitude needs to be learned. In Tbl. 2 (top), we observe that, except for correspondence accuracy for TOSCA, our contrastive loss $\mathcal{L}_c$ outperforms all of the alternatives in the correspondence matching task.

Contrastive and Reconstruction Loss. In Tbl. 2 (bottom), we individually assess our two losses. We see that the performance with only the reconstruction loss $\mathcal{L}_r$ is close to Diff3F. This indicates that

[1]We use the provided code and validate that we follow the same experimental procedure and metric definitions.

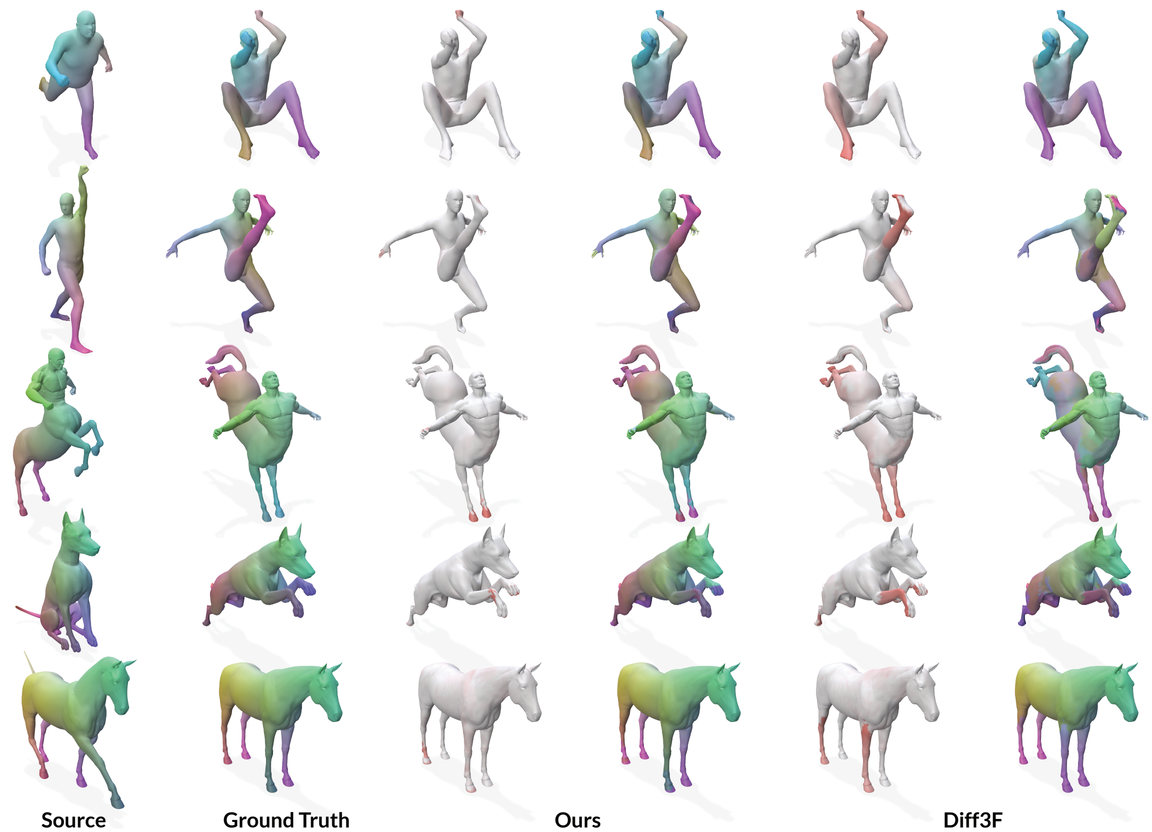

Fig. 6. Qualitative comparison on the *SHREC'19* and *TOSCA* datasets with dense true correspondence labels provided by their authors. We show the source and target meshes with their ground truth correspondence labels (the two left-most columns) in comparison to correspondences computed using our *surface-aware features* (the forth column) and Diff3F *base features* (the right-most column). We further highlight the correspondence error on the mesh surface (the third and the fifth column). The error colormap is normalized per sample by the maximal error over both methods to keep the error scale comparable across columns but not across rows. Our *surface-aware features* notably improve separation of the limb instances.

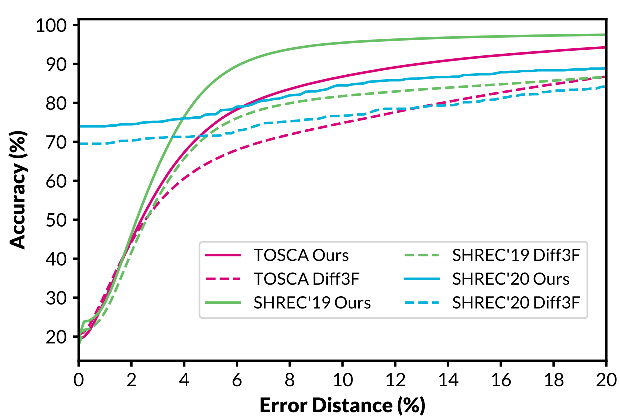

Fig. 7. Correspondence accuracy (↑ is better) at different error distance thresholds for our own and Diff3F features.

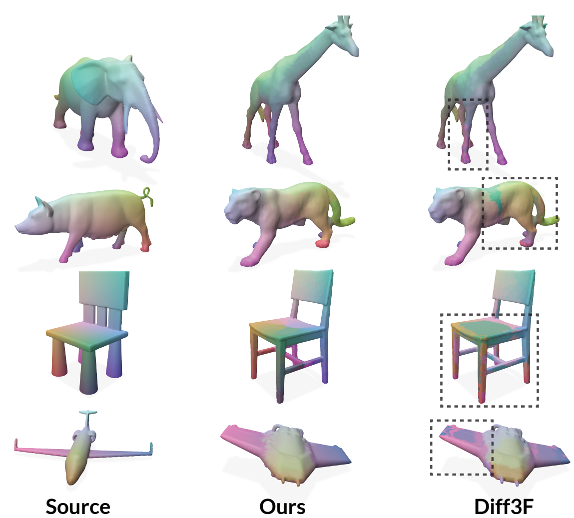

Fig. 8. Qualitative comparison of correspondence matching on *TOSCA* and ShapeNet [Chang et al. 2015] (dense ground truth labels not available). Source shape (left) matched to target (right) using our and Diff3F features.

Table 2. Ablation on our method. *Above the bar:* Ablation on alternative losses inspired by related work [Chen et al. 2025; Jiang et al. 2023] compared to the unmodified Diff3F features. *Below the bar:* Our full method compared to its reduced variant omitting losses $\mathcal{L}_c$ or $\mathcal{L}_r$. The per-column best results are bold and the second-to-best results are underlined.

		SHREC'19	TOSCA	SHREC'20
RGL	err ↓	0.80±1.08	2.74±2.53	5.29±9.85
	acc ↑	20.16±10.03	16.53±10.41	55.23±21.93
NGL	err ↓	0.54±0.90	2.11±2.02	4.86±9.49
	acc ↑	18.84±9.59	18.86±11.50	58.97±23.12
GSL	err ↓	1.72±1.45	4.17±5.21	4.34±9.23
	acc ↑	<u>26.89±9.07</u>	29.77±14.92	<u>73.39±26.31</u>
Diff3F	err ↓	1.69±1.44	4.51±5.48	5.34±10.22
	acc ↑	26.25±9.30	**31.00±15.73**	69.50±24.99
only $\mathcal{L}_r$	err ↓	1.65±1.44	4.70±5.64	4.87±9.38
	acc ↑	26.53±9.19	<u>30.27±15.17</u>	72.94±26.21
only $\mathcal{L}_c$	err ↓	**0.38±0.61**	<u>1.67±2.29</u>	<u>4.30±9.31</u>
	acc ↑	26.21±8.78	25.58±13.88	70.08±25.17
Ours	err ↓	<u>0.43±0.76</u>	**1.65±2.15**	**3.89±8.90**
	acc ↑	**28.78±9.30**	29.35±14.53	**73.97±26.47**

the gain in performance does not originate predominantly from a smaller embedding space or from access to training data. Similarly, the contrastive loss $\mathcal{L}_c$ alone results in an accuracy drop compared to our full model. This justifies our autoencoder approach with both losses playing an import role. Ablations on the number of anchors can be found in Appendix A.2.

6 Applications

We present additional downstream tasks that benefit from our *surface-aware features* learned in Sec. 5.

6.1 Instance-based Part Segmentation

Following the prior work [Dutt et al. 2024], we segment a target shape by clustering features around centroids from K-means clustering of source-shape features. In the top row of Fig. 9, we demonstrate a transfer from a big cat to a human and see that, unlike the Diff3F features, our *surface-aware features* disambiguate the limbs. In the bottom two rows, we repeat this experiment with a shared encoder trained on human, animals, and a subset of *ShapeNet* (see Appendix D.4) where a true mapping cannot be defined but our method finds reasonable analogies between the classes.

In Fig. 11, we repeat this with centroids obtained jointly from all $SMPL^{eval}$ and *TOSCA* samples. In contrast to Diff3F, our method successfully matches features across diverse shapes, which demonstrates our embedding's capability of many-to-many shape matching without any additional pairwise optimization. Finally, we show similar results for chairs and airplanes in Fig. 12.

6.2 Pose Alignment

Our *surface-aware features* are also useful for pose alignment of a kinematic model to another 3D shape. To this end, we establish point correspondences between shape pairs as in Sec. 5.4 and optimize the kinematic pose parameters to minimize point-to-point distances (see Appendix C).

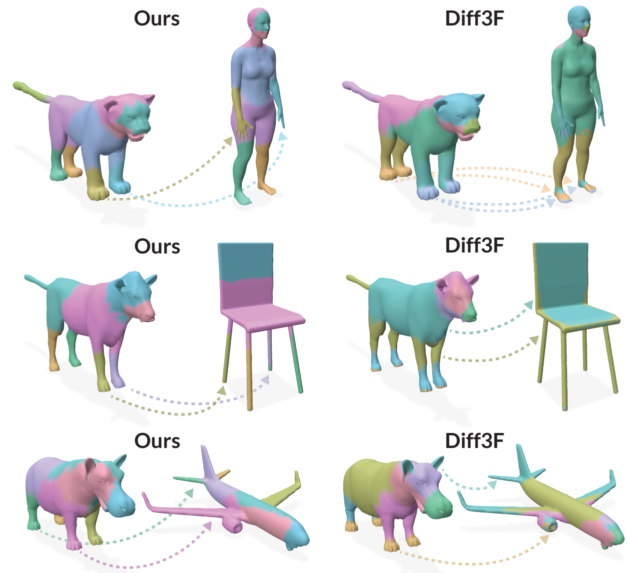

Fig. 9. In the top row, 10 k-means cluster centers from the big cat were used to segment the human. In the bottom two rows, 8 k-means cluster centers from the animals were used to segment the chairs and airplanes with a shared encoder. Unlike Diff3F, our method successfully separates all limbs for a plausible mapping from animal limbs to human limbs, chair legs, or airplane wings.

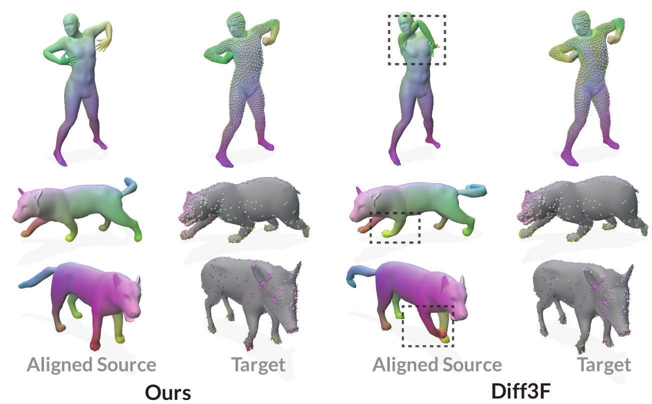

Fig. 10. Pose alignment of a source shape (color) into the pose of a target (gray). The boxes highlight challenging areas handled well by our method. For humans, we densely fit all the target vertices, while for animals, we only fit 5 % of the vertices as highlighted.

In Fig. 10, we align $SMPL^{eval}$ to *SHREC'19*, and $SMAL^{eval}$ to DeformingThings4D [Li et al. 2021] animals. Benefiting from the robust instance separation, our method produces poses closer to the targets for both dense and sparse correspondences. See our video for a 3D shape animation obtained by aligning to a target shape sequence.

6.3 Texturing

Since the *base features* are obtained from image models (see Sec. 3) and our pointwise encoder can process points sampled from a mesh

Fig. 11. Results when clustering features across all samples in $SMPL^{eval}$ and *TOSCA*. Our method implicitly aligns semantically-related regions (shown as the same colors) across diverse 3D shapes in a self-supervised manner (the top row). Diff3F produces inconsistent labeling across different shape categories as well as lack of separability between instanced components such as individual limbs (the bottom row).

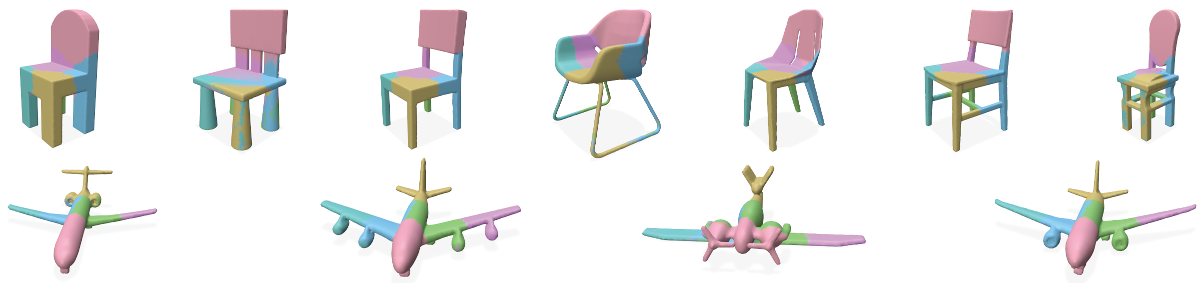

Fig. 12. Results when clustering the *surface-aware features* across chairs (top row) and airplanes (bottom row) from ShapeNet [Chang et al. 2015]. Note, that while we use a single shared encoder for all humanoid and animal shapes, we train a separate encoder for each ShapeNet class due to the large domain gap.

as easily as pixels sampled from an image, we can establish correspondences between a 2D image and a 3D mesh. We demonstrate this by texturing 3D meshes from a masked target image and individually assign each vertex a color from the image pixel that maximizes the mutual feature similarity (Eq. 3.1) (see Fig. 13 and Appendix D.3). We observe that our features produce a more coherent mapping leading to a better preservation of the source appearance when compared to the Diff3F *base features*. In Fig. 14, we further show that textures can also be effectively transferred between two 3D shapes using a combination our *surface-aware features* with Functional Maps like in Sec. 5.4.

7 Discussion

Limitations and Future Work. Our method inherits limitations connected to the extraction of the *base features*. Specifically, the extraction of Diff3F features [Dutt et al. 2024] takes several minutes per mesh and its vision model is sensitive to rendering artifacts or upside-down mesh orientations. We expect that advances in rendering of point representations increase the applicability across representations [Kerbl et al. 2023]. Furthermore, our method cannot establish a consistent partitioning for objects that are both geometrically and semantically isotropic (e.g., a round table). Hence, while our embedding separates human legs following the body's notion of front and rear, it cannot do so for table legs. However, this is not an issue for applications such as shape morphing [Sun et al. 2024]. Lastly, our method relies on consistency of geodesic distances between semantically distinct parts, and therefore it will be affected by geodesic shortcuts for partially blended parts in noisy 3D reconstructions (e.g., touching hands).

Beyond 3D alignment, our methodology could inspire 3D-to-2D pose estimation [Kanazawa et al. 2018; Peng et al. 2019], articulated 3D reconstruction [Uzolas et al. 2023; Yao et al. 2022], automated rigging [Xu et al. 2020] or 2D-to-3D uplifting [Liu et al. 2023; Poole et al. 2022], where our features could support more view-consistent representations. Finally, an interesting topic for future research is the development of foundational features using massive datasets, such as Objaverse [Deitke et al. 2023].

Conclusion. We have introduced novel *surface-aware features* for 3D shape matching that disambiguate intra-class instances among semantic features derived from pre-trained 2D vision models. Our descriptors have proven effective in distinguishing instances of the same semantic class and they generalize even when trained on a limited number of 3D shapes. Furthermore, our contrastive loss facilitates easy integration in future unsupervised methods which reduces data labeling effort. Consequently, our method is a promising building block toward adapting pre-trained 2D models to 3D tasks.

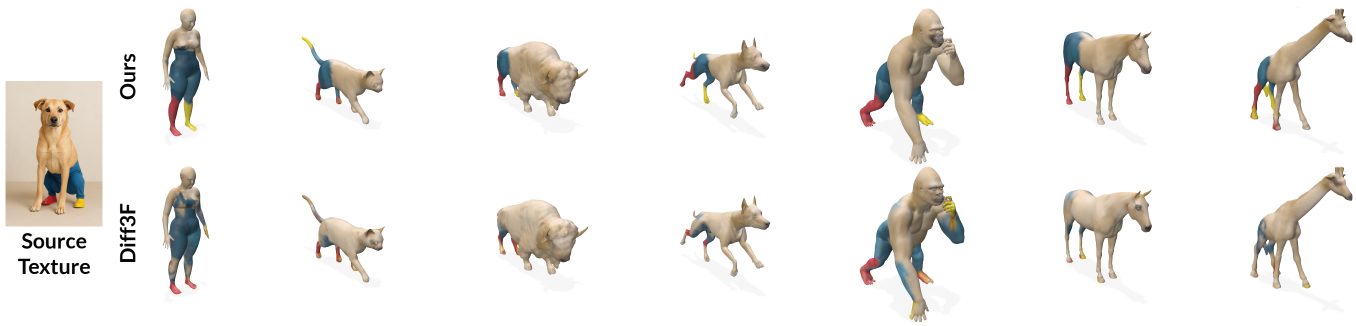

Fig. 13. Texturing of 3D meshes from *SMPL[eval]* and *TOSCA*, based on a 2D image generated with ChatGPT. The appearance is transferred by establishing correspondence between the image features and 3D mesh features. In contrast to Diff3F, our *surface-aware features* represent the input image more faithfully.

Fig. 14. Texturing of 3D meshes based on a source 3D mesh. The appearance is transferred based on correspondences established by combining our *surface-aware features* with Functional Maps. The data originate from *SMPL[eval]*, SMPLitex [Casas and Comino-Trinidad 2023], and DeformingThings4D [Li et al. 2021].

Acknowledgments

This work was partially supported by the Convergence AI Immersive Tech Lab at TU Delft.

References

Ahmed Abdelreheem, Abdelrahman Eldesokey, Maks Ovsjanikov, and Peter Wonka. 2023. Zero-shot 3d shape correspondence. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Noam Aigerman, Kunal Gupta, Vladimir G Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. 2022. Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–17.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. 2021. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814* 2, 3 (2021), 4.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 929–947.

Souhaib Attaiki, Lei Li, and Maks Ovsjanikov. 2023. Generalizable local feature pre-training for deformable shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13650–13661.

Mathieu Aubry, Ulrich Schlickewei, and Daniel Cremers. 2011. The wave kernel signature: A quantum mechanical approach to shape analysis. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 1626–1633.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611. Spie, 586–606.

Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. 2018. Creatures great and SMAL: Recovering the shape and motion of animals from video. In *ACCV*.

Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. 2008. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media.

Benedict J Brown and Szymon Rusinkiewicz. 2007. Global non-rigid alignment of 3-D scans. *ACM Transactions on Graphics (TOG)* 26, 3 (2007), 21–es.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*. 132–149.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

Dan Casas and Marc Comino-Trinidad. 2023. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

Zhangquan Chen, Puhua Jiang, and Ruqi Huang. 2025. DV-Matcher: Deformation-based Non-Rigid Point Cloud Matching Guided by Pre-trained Visual Features. *arXiv preprint arXiv:2408.08568v2* (2025).

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

Etienne Corman, Maks Ovsjanikov, and Antonin Chambolle. 2014. Supervised descriptor learning for non-rigid shape matching. In *European conference on computer vision*. Springer, 283–298.

Keenan Crane, Clarisse Weischedel, and Max Wardetzky. 2017. The heat method for distance computation. *Commun. ACM* 60, 11 (2017), 90–99.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2023. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems* 36 (2023), 35799–35813.

Haowen Deng, Tolga Birdal, and Slobodan Ilic. 2018a. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision (ECCV)*. 602–618.

Haowen Deng, Tolga Birdal, and Slobodan Ilic. 2018b. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 195–205.

Jiacheng Deng, Chuxin Wang, Jiahao Lu, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Zhe Zhang. 2023. Se-ornet: Self-ensembling orientation-aware network for unsupervised point cloud shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5364–5373.

Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. 2019. Learning elementary structures for 3d shape generation and matching. *Advances in Neural Information Processing Systems* 32 (2019).

Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. 2020. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8592–8601.

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

Roberto M Dyke, Yu-Kun Lai, Paul L Rosin, Stefano Zappalà, Seana Dykes, Daoliang Guo, Kun Li, Riccardo Marin, Simone Melzi, and Jingyu Yang. 2020. SHREC'20: Shape correspondence with non-isometric deformations. *Computers & Graphics* 92 (2020), 28–43.

Marvin Eisenberger, Zorah Lahner, and Daniel Cremers. 2020a. Smooth shells: Multiscale shape registration with functional maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12265–12274.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, and Daniel Cremers. 2020b. Deep shells: Unsupervised shape correspondence with optimal transport. *Advances in Neural information processing systems* 33 (2020), 10491–10502.

Gil Elbaz, Tamar Avraham, and Anath Fischer. 2017. 3D point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4631–4640.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks* 107 (2018), 3–11.

Michael Fischer, Zhengqin Li, Thu Nguyen-Phuoc, Aljaz Bozic, Zhao Dong, Carl Marshall, and Tobias Ritschel. 2024. NeRF Analogies: Example-Based Visual Attribute Transfer for NeRFs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4640–4650.

Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

Dvir Ginzburg and Dan Raviv. 2020. Cyclic functional mapping: Self-supervised correspondence between non-isometric deformable shapes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16.* Springer, 36–52.

Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. 2019. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 5545–5554.

Bingchen Gong, Diego Gomez, Abdullah Hamdi, Abdelrahman Eldesokey, Ahmed Abdelreheem, Peter Wonka, and Maks Ovsjanikov. 2024. ZeroKey: Point-Level Reasoning and Zero-Shot 3D Keypoint Detection from Large Language Models. *arXiv preprint arXiv:2412.06292* (2024).

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2018. 3d-coded: 3d correspondences by deep deformation. In *Proceedings of the european conference on computer vision (ECCV).* 230–246.

Kan Guo, Dongqing Zou, and Xiaowu Chen. 2015. 3D mesh labeling via deep convolutional neural networks. *ACM Transactions on Graphics (TOG)* 35, 1 (2015), 1–12.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2. IEEE, 1735–1742.

Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. 2019. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4370–4379.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

Tong He, Haibin Huang, Li Yi, Yuqian Zhou, Chihao Wu, Jue Wang, and Stefano Soatto. 2019. Geonet: Deep geodesic networks for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6888–6897.

Puhua Jiang, Mingze Sun, and Ruqi Huang. 2023. Neural intrinsic embedding for non-rigid point cloud matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 21835–21845.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 7122–7131.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

Itai Lang, Dvir Ginzburg, Shai Avidan, and Dan Raviv. 2021. Dpc: Unsupervised deep point correspondence via cross and self construction. In *2021 International Conference on 3D Vision (3DV).* IEEE, 1442–1451.

Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 2021. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 12706–12716.

Or Litany, Tal Remez, Emanuele Rodola, Alex Bronstein, and Michael Bronstein. 2017. Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE international conference on computer vision.* 5659–5667.

Haolin Liu, Xiaohang Zhan, Zizheng Yan, Zhongjin Luo, Yuxin Wen, and Xiaoguang Han. 2025. Stable-SCore: A Stable Registration-based Framework for 3D Shape Correspondence. *arXiv preprint arXiv:2503.21766* (2025).

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision.* 9298–9309.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2.* 851–866.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, and Shiyu Song. 2019. Deepvcp: An end-to-end deep neural network for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision.* 12–21.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2023), 47500–47510.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision.* 5442–5451.

Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. 2024. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19521–19530.

Simone Melzi, Riccardo Marin, Emanuele Rodolà, Umberto Castellani, Jing Ren, Adrien Poulenard, P Ovsjanikov, et al. 2019. SHREC'19: matching humans with different connectivity. In *Eurographics Workshop on 3D Object Retrieval.* The Eurographics Association, 1–8.

Facundo Mémoli and Guillermo Sapiro. 2005. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics* 5 (2005), 313–347.

Luca Morreale, Noam Aigerman, Vladimir G Kim, and Niloy J Mitra. 2024. Neural semantic surface maps. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15005.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=a68SUt6zFt Featured Certification.

Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. 2012. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG)* 31, 4 (2012), 1–11.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. 2019. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 4561–4570.

Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* 30, 4 (1992), 838–855.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PmLR, 8748–8763.

Emanuele Rodolà, Michael Moeller, and Daniel Cremers. 2015. Point-wise map recovery and refinement from functional correspondence. *arXiv preprint arXiv:1506.05603* (2015).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684–10695.

Jean-Michel Roufosse, Abhishek Sharma, and Maks Ovsjanikov. 2019. Unsupervised deep learning for structured shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 1617–1627.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference* *on computer vision and pattern recognition*. 815–823.

Gil Shamai and Ron Kimmel. 2017. Geodesic distance descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6410–6418.

Nicholas Sharp et al. 2019b. Polyscope. www.polyscope.run.

Nicholas Sharp, Keenan Crane, et al. 2019a. GeometryCentral: A modern C++ library of data structures and algorithms for geometry processing. https://geometry-central.net/. (2019).

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2024. SHIC: Shape-Image Correspondences with No Keypoint Supervision. In *European Conference on Computer Vision*. Springer, 129–145.

Jiaze Sun, Zhixiang Chen, and Tae-Kyun Kim. 2023a. Mapconnet: Self-supervised 3d pose transfer with mesh and point contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14452–14462.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. 2009. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, Vol. 28. Wiley Online Library, 1383–1392.

Mingze Sun, Chen Guo, Puhua Jiang, Shiwei Mao, Yurun Chen, and Ruqi Huang. 2024. SRIF: Semantic Shape Registration Empowered by Diffusion-based Image Morphing and Flow Estimation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Mingze Sun, Shiwei Mao, Puhua Jiang, Maks Ovsjanikov, and Ruqi Huang. 2023b. Spatially and spectrally consistent deep functional maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14497–14507.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems* 36 (2023), 1363–1389.

Federico Tombari, Samuele Salti, and Luigi Di Stefano. 2010. Unique signatures of histograms for local surface description. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*. Springer, 356–369.

Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. 2023. Template-free articulated neural point clouds for reposable view synthesis. *Advances in Neural Information Processing Systems* 36 (2023), 31621–31637.

Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. 2024. Motiondreamer: Zero-shot 3d mesh animation from video diffusion models. *arXiv preprint arXiv:2405.20155* (2024).

Peng-Shuai Wang, Yang Liu, and Xin Tong. 2022. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Yue Wang and Justin M Solomon. 2019. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3523–3532.

Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research* 10, 2 (2009).

Thomas Wimmer, Peter Wonka, and Maks Ovsjanikov. 2024. Back to 3D: Few-Shot 3D Keypoint Detection with Back-Projected 2D Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4154–4164.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.

Qianwei Xia, Juyong Zhang, Zheng Fang, Jin Li, Mingyue Zhang, Bailin Deng, and Ying He. 2021. GeodesicEmbedding (GE): a high-dimensional embedding approach for fast geodesic distance queries. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 4930–4939.

Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. 2020. Rignet: Neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559* (2020).

Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. 2015. Go-ICP: A globally optimal solution to 3D ICP point-set registration. *IEEE transactions on pattern analysis and machine intelligence* 38, 11 (2015), 2241–2254.

Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. 2022. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. *Advances in Neural Information Processing Systems* 35 (2022), 15296–15308.

Zi Jian Yew and Gim Hee Lee. 2018. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *Proceedings of the European conference on computer vision (ECCV)*. 607–623.

Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 2017. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1802–1811.

Yiming Zeng, Yue Qian, Zhiyu Zhu, Junhui Hou, Hui Yuan, and Ying He. 2021. Corrnet3d: Unsupervised end-to-end learning of dense correspondence for 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6052–6061.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2024a. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3076–3085.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2023a. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2023), 45533–45547.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. 2024b. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems* 36 (2024).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 2017. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6365–6373.