

MeshFM: 2D Features Are All You Need for 3D Shape Understanding

Jinfan Zhou^{*}, Richard Liu^{*}, Itai Lang^{*}, and Rana Hanocka^{*}

University of Chicago, Chicago IL 60637, USA
{zjf, guanzhi, itailang, ranahanocka}@uchicago.edu

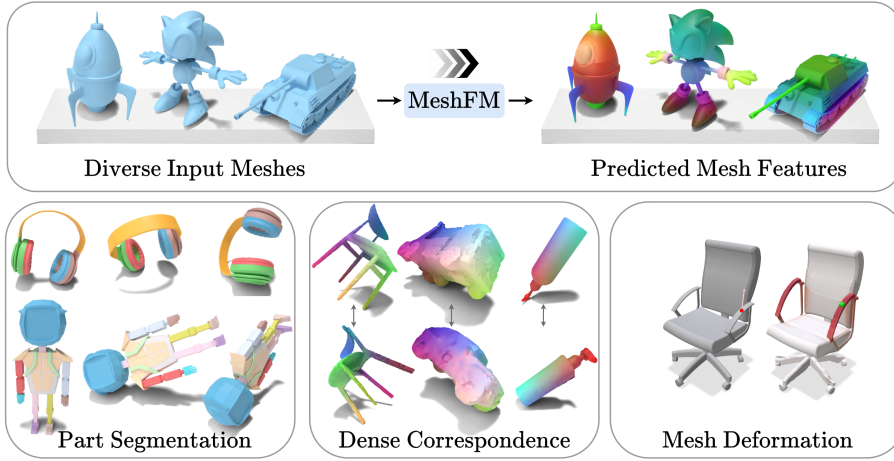


Fig. 1: MeshFM is a feedforward model that computes general-purpose features for 3D shapes. The learned features are discriminative and capture meaningful properties across diverse shapes and complex geometries. Importantly, the features are multi-purpose, and can be applied in a zero-shot manner for a variety of downstream tasks, such as segmentation, correspondence, and deformation.

Abstract. We present MeshFM, an efficient feedforward framework for extracting rich features from 3D inputs. Our method distills 2D features from visual foundation models into 3D. We train a feedforward network to directly predict 3D features without requiring optimization during inference. The approach utilizes a two-stage training strategy. First, we optimize a feature field in 3D using only 2D feature supervision. Second, we train a network to regress this feature field. The entire procedure requires no 3D annotation, instead relying on the powerful information in 2D foundation models. We demonstrate that our learned features can be immediately applied to downstream tasks, including part segmentation, dense correspondence, and mesh deformation. Extensive experiments show that MeshFM, trained solely with 2D supervision, performs on par with methods trained explicitly with 3D supervision, even without task-specific fine-tuning. Moreover, our model is trained to be robust to extreme rotations of the input objects. Project page: <https://threedle.github.io/MeshFM/>

^{*} Equal contribution.

Keywords: Foundation Models · 3D Shape Understanding · Feature Distillation

1 Introduction

The landscape of computer vision has been fundamentally transformed by the advent of large-scale 2D foundation models. Models such as DINO, CLIP, and SAM have demonstrated that representations learned from large-scale data are rich in understanding and can be applied to a wide range of tasks. Naturally, researchers have also sought to harness this rich 2D prior for 3D shape tasks and 3D representation learning.

There are currently two primary approaches for lifting 2D priors into 3D. The first is to directly render shapes, obtain the encoded render features, and backproject those features onto the shape through some optimization process. We refer to this general process as “distillation”. This procedure is exemplified by works such as Diff3F [9] and DFD [24]. Though distillation is effective at obtaining high quality 3D features, it requires per-shape optimization and is impractical for real-time applications.

On the other hand, there are works which leverage 2D priors indirectly through the training process. These works aim to learn *task-specific* 3D representations, where the network features are informed by 2D models. Exemplars in this domain are PartField [22], which uses SAM to generate clusters for a contrastive loss to learn segmentation-aware features, and DenseMatcher [44], which initializes noisy 3D features through 2D backprojection and “refines” them for the purposes of learning correspondence-aware features. Though these models achieve SOTA performance, the features are designed for a specific task and are not applicable beyond their target application. Furthermore, they ultimately rely on 3D data for supervision, which introduces a data bottleneck.

In this work, we challenge the prevailing paradigm that 3D tasks require specialized 3D networks. We posit that the reliance on task-specific heads is largely a symptom of imperfect feature distillation. Standard distillation techniques suffer from aliasing (distilling to coarse vertices) and “feature bleeding”, where image patchification in ViT architectures (e.g. DINO) causes background semantics to blur into the object. We find that when these issues are resolved, distilled 2D features are *sufficient* for achieving near-SOTA on most 3D tasks.

Championing this hypothesis, we present a feedforward **Mesh Feature Model** (MeshFM), a model which predicts rotation-robust, rich visual features over 3D inputs in a feedforward manner. Our method decouples the expensive process of feature distillation from the requirements of real-time inference, allowing us to generate “teacher-quality” feature fields in a feedforward manner. Furthermore, with direct supervision on the features, our model is trained to be robust to extreme rotations. We achieve this through two key technical innovations:

- **Refined Teacher Distillation via SAM.** We introduce a robust “teacher” generation stage that addresses the root causes of feature noise. First, we employ feature field distillation to supervise the feature field at every rasterized

pixel, decoupling the learning signal from the mesh triangulation and fully utilizing the 2D signal. Second, and most crucially, we implement a feature refinement strategy. By leveraging the precise masks from the SAM, we identify and "wipe" outlier features that bleed across part boundaries. This steers the distilled features to respect sharp geometric and visual boundaries, producing a refined ground truth.

- **General-Purpose Feedforward Prediction.** Instead of training a task-specific network, we train a feedforward network to predict optimized teacher fields. We further apply $SO(3)$ rotation augmentation during training to make our network rotation-robust, allowing it to be applied to in-the-wild shapes with arbitrary orientations.

The implications of this approach are significant: MeshFM produces a general shape representation that requires no 3D annotations and no task-specific fine-tuning. We validate this by applying our predicted features zero-shot to three applications: segmentation, dense correspondence, and interactive shape deformation. In all cases, our general-purpose features perform on par with or are superior to specialized SOTA baselines. This finding suggests that when distilled properly, 2D features may be sufficient for achieving comprehensive 3D shape understanding, without the need for complex, domain-specific priors.

2 Related Work

2.1 Foundation Models

Foundation models generally refer to a class of models which predict rich visual or text features and are trained on unlabeled internet-scale data. These models are trained with various methods. Contrastive Vision-Language Models (VLMs) like CLIP [29] and BLIP [20, 21] align images and text in a shared embedding space, enabling robust open-vocabulary recognition. Recent works such as SigLIP [36, 42] has further scaled this paradigm to billions of image-text pairs, drastically improving zero-shot performance. Parallel to VLMs, self-supervised learning has produced powerful representations. The DINO series [4, 26, 33] utilizes self-distillation with Vision Transformers [8] to learn spatially aware features that exhibit emergent semantic part correspondences. Other self-supervised approaches like MAE [10] and I-JEPA [1] focus on masking and reconstruction, learning robust low-level features. The "Segment Anything" paradigm has introduced general-purpose segmentation models capable of zero-shot transfer via promptable interfaces. This field has rapidly evolved from static images [15] to unified video and concept segmentation with SAM 2 [31] and SAM 3 [3]. Concurrently, models like SAM 3D [5] have extended these capabilities to 3D. Recent agglomerative models like RADIO [11, 30] attempt to distill foundation models from multiple modalities into a single efficient backbone.

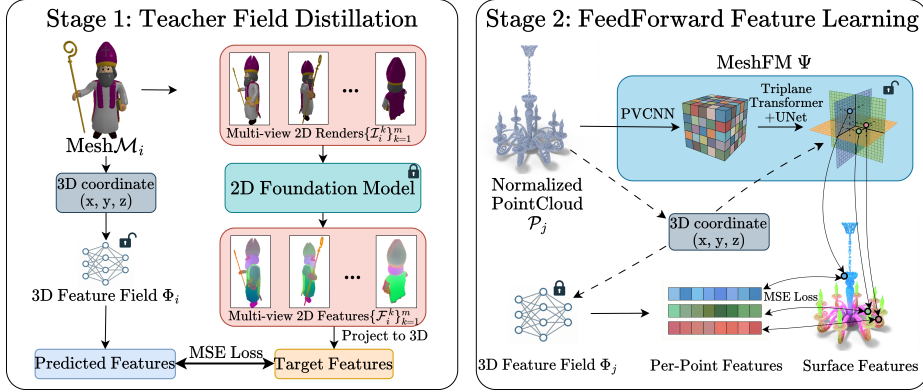


Fig. 2: Training pipeline overview. The training phase of MeshFM includes two stages. In the first stage, we distill the information from a 2D foundation model into 3D feature fields. 2D feature maps are extracted from multi-view renderings of the mesh and projected to the mesh surface, and a neural model learns to map 3D coordinates on the mesh to the reference features. In the second stage, we use the teacher feature fields to train a feedforward feature prediction model. The model takes a point cloud sampled from the mesh surface and learns to output the point features from the supervising 3D teacher fields.

2.2 2D Priors for 3D Tasks

Leveraging 2D representations for 3D understanding predates modern foundation models. Multi-View CNN [34] showed that a 3D shape can be recognized from rendered 2D views. This motivated projection-based methods that render multiple views, and back-project the 2D results onto the geometry, later extended from classification to dense segmentation: 3DMV [6] back-projects 2D feature maps into a voxel grid and Virtual Multiview Fusion [18] fuses rendered-view predictions on the surface. With VLMs, the focus shifted from transferring labels to transferring open-vocabulary features: OpenScene [28] predicts dense features co-embedded with CLIP [29], and CLIP-FO3D [43], and OpenMask3D [35] similarly lift 2D features for open-vocabulary segmentation, closely related to radiance-field distillation (LERF [13], DFF [16], N3F [37]). Diff3F [9] distills internal features from Stable Diffusion [32] and DINOv2 [26] to establish dense correspondences between diverse shapes.

DFD [24] optimizes the feature distillation process for arbitrary resolution shapes by assigning 2D features per pixel to 3D points through barycentric coordinates. Interactive methods have also emerged to bring human intent into the loop. iSeg [19] introduces an interactive approach, employing an attention mechanism to distill SAM features into a 3D field that responds to user clicks, enabling view-consistent, user-conditioned 3D segmentation. These distillation methods rely on expensive per-scene optimization or interactive loops. In contrast, our method trains a feedforward network that predicts the distilled feature field directly from a point cloud, affording immediate access to rich 2D priors without

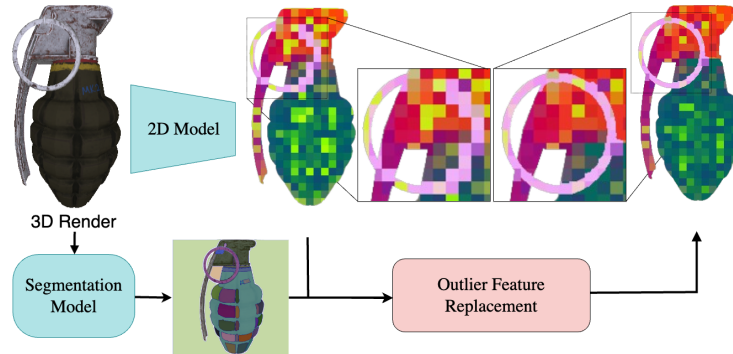


Fig.3: Feature aliasing correction. We correct the feature aliasing created by patch-based image encoder architectures through fine-grained image segmentation. The segmentation groups define pixel clusters that are assumed to have similar deep features. Pixels with features substantially different than the cluster median are replaced by the median. The resulting feature map has much less blurring at semantic boundaries and more stable features within semantic groups.

any fine-tuning or optimization. PartField [22] takes a slightly different approach to 3D task learning from 2D priors. The authors leverage SAM-predicted segmentations and backprojection to learn part-aware features through contrastive learning. PartField features have been successfully applied in downstream applications that require segmentation inputs [39], but do not generalize beyond part clustering.

3 Method

Given an input mesh, we predict general-purpose features over the mesh surface in a feedforward manner. The training of our method happens in two stages. First, we distill features from a 2D foundation model into neural fields, fitting a separate field per training shape. These feature fields share a common semantic space and serve as teacher fields for training our feedforward network in the second stage. During the second stage training, we apply $SO(3)$ rotation augmentation to the objects such that the model learns rotation-robust features.

3.1 Stage 1: Teacher Field Distillation

In the first stage, our goal is to distill rich 2D signals from visual foundation models into a continuous 3D feature field Φ . We adopt the *barycentric feature distillation* approach from DFD [24], which decouples feature supervision from the shape’s surface discretization, making it an ideal choice for robust distillation over arbitrary meshes. We render m multi-view images $\{\mathcal{I}^k\}_{k=1}^m$ from a mesh $\mathcal{M}$ and feed them to a 2D foundation model to extract the corresponding feature maps $\{\mathcal{F}^k\}_{k=1}^m$. Features are then back-projected to the 3D space and

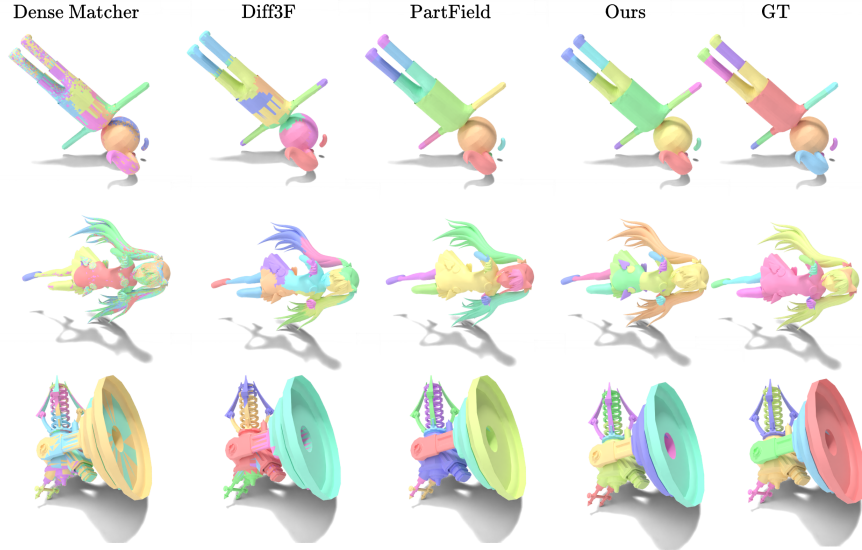


Fig. 4: Part segmentation comparison. Our method can segment the mesh without training on any 3D annotations. We can even perform on par with PartField, which is directly trained with 3D segmentation annotation. In cases where our segmentation differs from the ground-truth, they are still semantically meaningful (e.g. decorative ribbons on the dress).

aggregated from all the views. Then, a feature field Φ is optimized to fit this feature supervision by learning to map 3D coordinates over the mesh to their associated per-pixel features projected from the 2D feature maps.

Feature Aliasing Correction While ViT-based foundation models like DINOv2 [26] and RADIO [11, 30] provide rich semantic descriptors, we observe that image patchification inevitably results in “feature bleeding” at object borders and part boundaries.

We illustrate this in Fig. 3, where, due to the patch sampling of the 2D encoder, the features on the ring are blurred with adjacent unrelated features. No matter how expressive the 2D model is, this patch-level aliasing is inherent to the feature extraction at a structural level. This feature bleeding is not resolved through the distillation optimization, and makes its way into the final 3D distillation and subsequent segmentation (as shown in Fig. 3, “Original Distillation”).

We correct the feature aliasing through a feature refinement strategy that leverages the precise boundary detection capabilities of SAM [31]. For each rendered view $\mathcal{I}^k$, we generate a set $\mathcal{S}^k = \{M_1, \dots, M_{n_k}\}$ of disjoint segmentation masks using SAM. We assume that pixels within a single segment delineate a semantically consistent part, and thus should possess similar deep features.

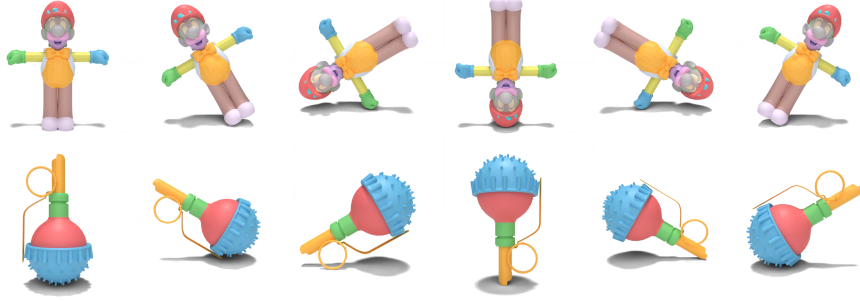


Fig. 5: Rotation robustness. We visualize our segmentation results for different rotations of the input shape. MeshFM is highly robust to rotations, where the segmentation remains consistent under large rotations in the full $SO(3)$ group.

For every segment $M_i \in \mathcal{S}^k$, we first compute the representative segment feature μ_i as the channel-wise median of all feature vectors within the mask:

$$\mu_i = \text{median}(\{\mathbf{f}_{\mathbf{u}} \mid \mathbf{u} \in M_i\}), \quad (1)$$

where $\mathbf{f}_{\mathbf{u}}$ is the raw feature vector at pixel $\mathbf{u}$. Then, we identify “outlier” pixels where the raw feature deviates significantly from the representative feature. A pixel $\mathbf{u} \in M_i$ is considered an outlier if its Euclidean distance from μ_i exceeds a scalar threshold τ :

$$\|\mathbf{f}_{\mathbf{u}} - \mu_i\|_2 > \tau. \quad (2)$$

Our features are unit-normalized by barycentric feature distillation, so we find setting $\tau = 1$ to produce reasonable results across all experiments.

For all such outlier pixels, we replace their raw feature $\mathbf{f}_{\mathbf{u}}$ with the representative μ_i . This operation effectively recovers geometric/semantic boundaries in the feature map, as demonstrated in Fig. 3, “Corrected Distillation”.

3.2 Stage 2: Rotation-Robust Feedforward Feature Prediction

In the second stage, we distill the optimized feature fields $\{\Phi\}$ from Stage 1 into a generalizable feedforward network Ψ . This allows us to predict the 3D feature field for a new shape in a single forward pass, eliminating the need for test-time optimization.

Architecture. Our feedforward network Ψ is designed to process a normalized point cloud input $\mathcal{P} \in \mathbb{R}^{C \times 3}$ and output a continuous feature field represented by triplanes. As illustrated in Fig. 2, the architecture consists of two main components:

1. **Point-Voxel Encoder:** We employ a PVCNN [25] encoder to extract per-point features from the input geometry. These features are orthogonally projected via mean-reduction onto three axis-aligned planes (XY , XZ , YZ) to form an initial triplane representation.

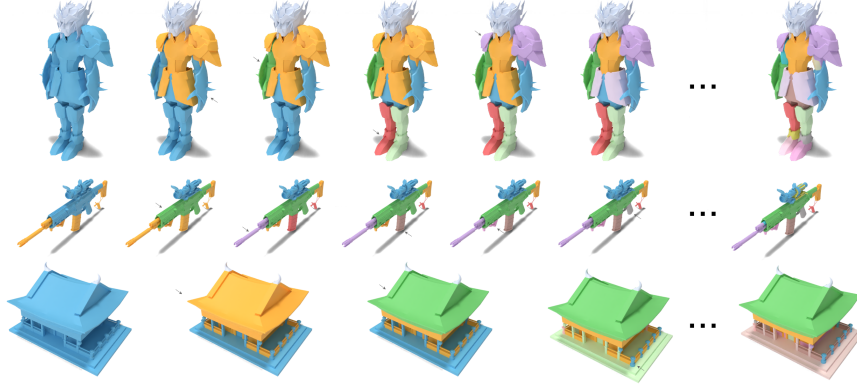


Fig. 6: Hierarchical part segmentation. Our method can segment the mesh at an increasing granularity by simply increasing the number of feature clusters (from left to right). Notably, we support a wide range of part counts, distinguish part instances (*e.g.*, the character’s boots), and capture fine-grained shape details (*e.g.*, the band on the rifle stock).

2. **Triplane Transformer:** The initial triplanes are processed by a 2D CNN for downsampling, flattened, and passed through a Transformer to capture global context. The processed features are then upsampled back via a transposed 2D CNN to yield the final triplane tensor $\mathbf{T} \in \mathbb{R}^{3 \times D \times H \times W}$, where D is the feature dimension and $H \times W$ are the spatial dimensions.

Distillation from Stage 1. Unlike Stage 1, which relies on 2D supervision from differentiable rendering (noisy and expensive), Stage 2 is fully supervised by the 3D signals generated from the frozen Stage 1 feature fields. We treat an optimized field Φ_i as a teacher and train Ψ to regress Φ_i ’s output directly.

During training, we normalize the input shape to the unit cube, same as with teacher distillation. We generate a set of query points $\mathcal{Q} = \{\mathbf{x}_k\}_{k=1}^K$ by uniformly sampling (x, y, z) coordinates over the shape surface. For each query point $\mathbf{x}_k$, we obtain the target feature vector $\mathbf{f}_k^{gt}$ by querying the frozen Stage 1 field: $\mathbf{f}_k^{gt} = \Phi(\mathbf{x}_k)$.

The network Ψ takes the shape’s point cloud $\mathcal{P}$ as input and predicts the triplane representation $\mathbf{T} = \Psi(\mathcal{P})$. We then compute the predicted feature $\mathbf{f}_k^{pred}$ at $\mathbf{x}_k$ by bilinearly interpolating the features from the three planes of $\mathbf{T}$ and summing them: $\mathbf{f}_k^{pred} = \text{Sample}(\mathbf{T}, \mathbf{x}_k) = \sum_{p \in \{XY, XZ, YZ\}} \mathbf{T}_p(\text{proj}_p(\mathbf{x}_k))$.

The training objective is simply the L2 loss between the predicted features and the target features:

$$\mathcal{L}_{\text{stage2}} = \frac{1}{K} \sum_{k=1}^K \|\mathbf{f}_k^{pred} - \mathbf{f}_k^{gt}\|^2. \quad (3)$$

Category	Original			SO(3) Rotated		
	Diff3F [9]	PartField [22]	Ours	Diff3F [9]	PartField [22]	Ours
Animals	0.507	0.561	<u>0.555</u>	<u>0.462</u>	0.322	0.570
Buildings & Outdoor	<u>0.515</u>	0.481	0.572	<u>0.445</u>	0.308	0.533
Daily-Use	0.548	0.562	<u>0.556</u>	<u>0.512</u>	0.456	0.522
Electronics	0.539	<u>0.572</u>	0.595	<u>0.524</u>	0.437	0.605
Food	<u>0.544</u>	0.635	0.494	0.565	0.461	<u>0.520</u>
Human-Shape	0.487	<u>0.508</u>	0.519	<u>0.469</u>	0.440	0.514
Plants	0.515	<u>0.570</u>	0.613	<u>0.496</u>	0.361	0.574
Transportation	0.464	<u>0.443</u>	0.491	<u>0.429</u>	0.260	0.498
Average	0.515	<u>0.542</u>	0.549	<u>0.488</u>	0.393	0.539

Table 1: Class-agnostic semantic segmentation on the PartObjaverse-Tiny dataset. We report the mean IoU for evaluating on original and SO(3) rotated shapes. Higher is better. The best result is **bolded**, and the second best is underlined. Our method performs better than the state-of-the-art baselines both with and without SO(3) rotation.

This two-stage approach allows MeshFM to combine the semantic richness of 2D foundation models (captured in Stage 1) with the inference speed of feedforward 3D networks (achieved in Stage 2).

SO(3) Rotation Augmentation. To enable application to in-the-wild shapes, which frequently exhibit arbitrary orientation, we apply SO(3) rotation augmentation during training. Specifically, we first train the model without rotation until convergence. We then train for another 365k iterations, and randomly sample rotation angles along all 3 axes for each training shape, linearly increasing the maximum rotation angle θ_{max} until $\theta_{max} = 2\pi$ in the first 125k iterations.

4 Experiments

To demonstrate the general-purpose utility of our learned features, we apply them to common 3D tasks and evaluate our method against state-of-the-art (SOTA) baselines for each task. For all the evaluation datasets, we additionally generate a “rotation-augmented” version by sampling 5 random SO(3) rotations for each shape, and further evaluate the baselines for rotation robustness. Our training is performed on 4 L40S GPUs, using a batch size of 4 objects on each GPU. We train with the Adam optimizer with a learning rate of 1×10^{-4} . Our model has ~ 126 M parameters.

Our goal is not to demonstrate SOTA performance on any given task, but rather to show that our *general-purpose* features are competitive with SOTA *task-specific* features. Importantly, we demonstrate that our features generalize across different tasks, whereas the task-specific features do not. To this end, we evaluate all baseline features on all tasks zero-shot, taking care to apply the same postprocessing functions to all features for the same task. This ensures we are fairly comparing features in terms of their representation quality.

Baselines. We compare against PartField [22] (segmentation), DenseMatcher [44] (textured correspondence), and Diff3F [9] (correspondence), which are all SOTA models in their respective domains. These baselines all provide point features per shape, allowing for general evaluation across all our shape tasks.

PartField Comparison Discussion While we use PartField as a baseline, PartField itself notes that its PVCNN and triplane architecture is inherently extrinsic and requires consistently oriented shapes. Since our model uses a similar backbone and gains rotation robustness solely through $SO(3)$ augmentation, PartField could in principle be trained the same way, and we acknowledge that comparing our augmented model against the un-augmented PartField may give us some advantage on rotated inputs, though its reliance on 2D SAM-mask supervision would make such augmentation harder and likely degrade its mask quality. Even on non-rotated shapes, however, our model performs on par with PartField without using any annotation for supervision.

Category	Original			SO(3) Rotated		
	Diff3F	PartField	Ours	Diff3F	PartField	Ours
Electro. & Comput.	0.495	0.521	<u>0.502</u>	0.494	0.503	<u>0.495</u>
Home Appliances	0.373	0.425	<u>0.382</u>	0.372	0.393	<u>0.377</u>
Kitchen & Food	0.513	0.573	<u>0.542</u>	0.533	0.570	<u>0.545</u>
Furnit. & Househo.	0.451	<u>0.471</u>	0.486	0.454	<u>0.444</u>	0.467
Tools, Office, & Misc.	<u>0.633</u>	0.600	0.651	<u>0.636</u>	0.592	0.647
Average	0.496	<u>0.517</u>	0.520	<u>0.501</u>	0.499	0.510

Table 2: Class-agnostic semantic segmentation on the PartNetE [23] test set. We use labels based on the original PartNet semantic labels and report the mean IoU on both the original and $SO(3)$ rotated dataset. Higher is better. The best result is **bolded**, and the second best is underlined. Our method performs better than the state-of-the-art baselines both with and without $SO(3)$ rotation.

4.1 Part Segmentation

Following PartField [22], we evaluate our method on class-agnostic 3D part segmentation on two datasets, PartObjaverse-Tiny [41] and PartNetE [23]. The PartObjaverse-Tiny dataset contains 200 shapes spanning 8 categories with human-annotated segmentations. The PartNetE test set contains 1,906 shapes across 45 categories, which are consolidated into 5 categories defined by PartField (see supplemental for the mapping copied over from PartField). We do not report DenseMatcher results as the model fails on the non-manifold inputs in these datasets.

We report results on both semantic segmentation and instance segmentation labels for both datasets (instance segmentation results in supplemental). We apply the same agglomerative clustering algorithm as PartField for all baseline features, and we set clusters k to be the number of unique ground truth labels.

Method	Original				SO(3) Rotated			
	All		Held-out		All		Held-out	
	AUC $\uparrow$	Err $\downarrow$	AUC	Err	AUC	Err	AUC	Err
Diff3F [9]	0.46	11.8	0.48	7.6	0.42	13.8	0.41	10.1
DenseMatcher [44]	<u>0.50</u>	<u>10.2</u>	0.52	6.8	0.46	11.8	<u>0.50</u>	7.8
PartField [22]	0.35	17.1	0.37	11.5	0.24	21.0	0.38	11.2
MeshFM (ours)	0.52	9.6	<u>0.51</u>	<u>7.2</u>	0.50	10.5	0.51	7.0

Table 3: DenseCorr3D correspondence. We report normalized geodesic error (“Err”) and AUC with a 10% threshold on all test categories and held-out categories. The best result is **bolded**, and the second best is underlined. Our method is on par with state-of-the-art baselines.

Method	Original		SO(3) Rotated	
	AUC $\uparrow$	Err $\downarrow$	AUC $\uparrow$	Err $\downarrow$
Diff3F [9]	0.48	17.1	<u>0.28</u>	<u>24.8</u>
DenseMatcher [44]	0.11	31.5	0.02	41.7
PartField [22]	0.25	19.3	0.05	39.5
MeshFM (ours)	<u>0.34</u>	<u>18.8</u>	0.33	19.3

Table 4: TOSCA correspondence. We report normalized geodesic error (Err) and AUC with a 10% threshold. The best result is **bolded**, and the second best is underlined. Our method is on par with state-of-the-art baselines.

Metric Calculation. Following previous work, we compute mIoU between the predicted segmentations and ground truth. To handle label ambiguity, we use the Hungarian algorithm [17] to obtain the best 1-1 assignment between the different label sets.

Results. We report semantic segmentation results in Tables 1-2. MeshFM marginally outperforms PartField on both datasets, and with rotation augmentation, our method handily outperforms all baselines across all categories. We provide a qualitative comparison in Fig. 4, which clearly highlights the failure of both PartField and Diff3F to adapt to rotations. PartField features have a strong canonical coordinate bias, as discussed earlier, whereas Diff3F interestingly tends to be slightly more robust (perhaps due to some learned rotation robustness from the backbone image models).

We visualize our rotation robustness in Fig. 5 and show additional hierarchical segmentation results in Fig. 6. MeshFM features produce clean, boundary-aligned segmentations which are consistent across the full rotation spectrum.

4.2 Dense Correspondence

We evaluate dense correspondence on the TOSCA [2] and DenseCorr3D [44] datasets. The TOSCA dataset consists of 80 human/animal shapes deformed

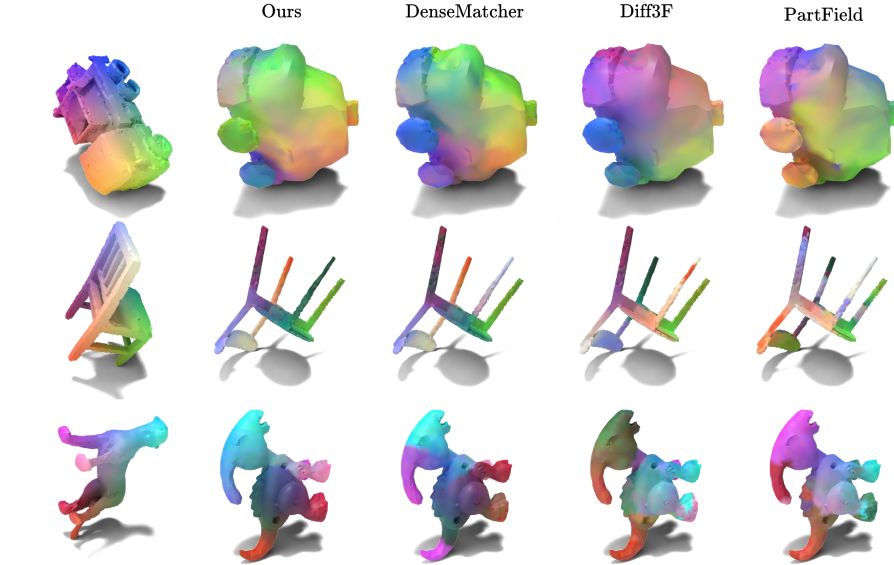


Fig. 7: Dense correspondence comparison. While other methods fail on these examples, our mesh features are consistent across shapes and enable dense matching between shapes even when shapes are not in a canonical pose and are not aligned. The correspondence is visualized by color transfer over each pair. Notably, the correspondence applies to various domains and different geometries, including household objects and vehicles with sharp edges, as well as (toy) animals with smooth surfaces.

from template meshes. We compute correspondence maps for each pair of shapes in each class, resulting in a total of 420 pairs. The DenseCorr3D dataset is a recent correspondence dataset which leverages textured shape data from Objaverse-XL [7] and OmniObject3D [40]. The DenseCorr3D test set consists of 24 shape classes with 3 held-out classes, with 306 total test pairs. Following DenseMatcher, we use the regularized functional map [27] designed for deep features to compute the correspondence for all methods.

Metric Calculation. We compute normalized geodesic error (“Err”) [14] of the predicted correspondences and area-under-curve (“AUC”) at 10% threshold.

Results. We report quantitative results in Tables 3 and 4. Similar to the segmentation results, MeshFM features are on par or slightly underperform the SOTA models on the standard datasets, and exhibit impressive stability on the rotation-augmented shapes, whereas the other baselines largely fail under arbitrary orientation. PartField features do particularly poorly in this setting, which is understandable given that they are optimized for segmentation. This suggests, nevertheless, that generalizable 3D features cannot be learned from task-specific training.

We present qualitative comparison results in Fig. 7, where we show method predictions on canonically oriented examples alongside the same pairs with random rotation applied. Observe that all methods apart from ours exhibit strong canonical coordinate bias in the computed correspondence, which supports our quantitative evaluation and suggests a common failure mode across modern feature learning techniques.

	Original	SO(3) Rotated
Diff3F [9]	0.88	0.33
PartField [22]	0.85	0.41
DenseMatcher [44]	0.82	0.28
Ours	0.92	0.85

Table 5: Manifold40 Classification. We evaluate the class-discriminateness of the features from each method by training and testing a lightweight (3-layer) MLP on mean-pooled features over the Manifold40 dataset [12]. We report average accuracy over the Manifold40 test set, both with and without SO(3) rotation augmentations.

4.3 Classification

We evaluate the class-discriminateness of the features from each method by training and testing a lightweight (3-layer) MLP on mean-pooled features over a subset of the Manifold40 dataset [12]. We report average accuracy over the original dataset and a rotation-augmented version in Table 5. Our features are more discriminative than baseline features when used for classification by the same lightweight architecture, and naturally exhibit a more consistent signal under rotation augmentation.

4.4 Mesh Deformation

We produce deformations by predicting features over mesh vertices, then applying affine transformations to control handles through the DFD framework [24] over the Part-Objaverse dataset in Fig. 8. The deformation weights derived from our features enable smooth, semantic-aware deformations of the input shape, regardless of its initial orientation. Observe that our deformations can robustly isolate varying semantic features of shapes and are smooth enough to interpolate deformations in a reasonable manner.

4.5 Ablation Study

Feature Aliasing Correction Ablation. We perform an ablation experiment on the SAM-based feature aliasing correction procedure by comparing the performance of teacher fields trained with and without SAM on PartObjaverse-Tiny segmentation. Specifically, we report mIOU on instance segmentation on

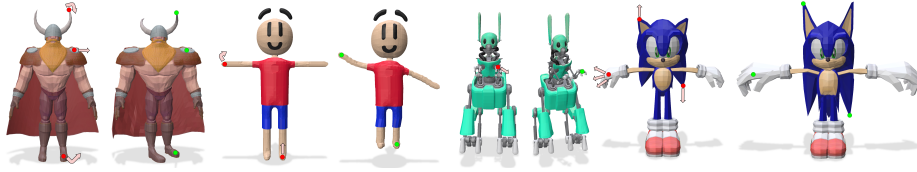


Fig. 8: Handle-Based Deformation. We show multi-handle deformations using our features and the Deep Feature Deformation framework [24] on shapes from the Part-Objaverse dataset. Initial control handles and affine transformations are shown in red. Final handle destinations are shown in green.

PartObjaverse-Tiny in Table 6. Observe that the feature aliasing correction of ours offers a substantial boost to the mIOU performance of the teacher fields across all categories, indicating the feature antialiasing consistently improves feature quality across shape classes.

Category	Teacher w/o SAM	Teacher w/ SAM
Animals	0.600	0.618
Buildings & Outdoor	0.572	0.587
Daily-Use	0.653	0.647
Electronics	0.701	0.716
Food	0.620	0.778
Human-Shape	0.680	0.714
Plants	0.592	0.633
Transportation	0.577	0.595
Average	0.630	0.661

Table 6: Ablation experiment: instance segmentation on the PartObjaverse-Tiny dataset. We report the mean IoU between teacher feature fields optimized with and without SAM aliasing correction (see 3.1 in paper). Higher is better. The best result is **bolded**. Observe that SAM feature aliasing correction improves segmentation performance across nearly all categories.

Foundation Model Feature Ablation. We compared the distilled teacher fields extracted from RADIO, SAM, DINOv2, and DINOv3 directly on PartNetE semantic segmentation (Table 7). Both RADIO and DINO variants yield comparable, high-performance results, demonstrating our method’s robustness to different semantically rich features. SAM features perform slightly worse due to high view-inconsistency (though our feature field optimization still consolidates them reasonably well).

5 Limitations

Despite the generality of our predicted mesh features, we still inherit the shortcomings of the supervising 2D foundation model. In particular, in our work, we

Category	DINOv2 (Original)	RADIO	SAM2	DINOv3
Electro. & Comput.	0.52	0.50	0.53	0.52
Home Appliances	0.39	0.37	0.38	0.41
Kitchen & Food	0.57	0.59	0.52	0.58
Furnit. & Househo.	0.51	0.52	0.49	0.53
Tools, Office, & Misc.	0.66	0.65	0.61	0.67
Average	0.54	0.54	0.51	0.54

Table 7: Ablation of 2D foundation models on PartNetE dataset.**Fig. 9: Left-right mismatch limitation.** Our mesh features sometimes do not distinguish left and right symmetric shape parts, such as chair legs and animal limbs, leading to correspondence errors.

used the popular DINOv2 model [26], which is known to struggle distinguishing left from right [38], as it mainly reasons about semantic meaning and less about relative position. This limitation is evident in the correspondence task, where left and right matches are occasionally swapped, as seen in Fig. 9. Nonetheless, our feature distillation framework is not tailored to a specific 2D model. Thus, our MeshFM capabilities can be improved with the advancement of 2D foundation models.

6 Conclusion

We presented MeshFM, a feedforward model that predicts general-purpose features for 3D shapes. We distill 2D features from a foundation model into 3D teacher fields, using our novel aliasing correction to combat feature bleeding between parts and use these clean, boundary-aware fields to supervise a feedforward model trained with full-range rotation augmentation. The resulting features enable zero-shot, rotation-robust inference and perform competitively with task-specific baselines on segmentation, correspondence, and deformation.

7 Acknowledgement

This work was supported by NSF grants #2304481, #2335493, #2402894, #2542757, BSF grant #2022363, gifts from Adobe, Snap, and Google, and the Bennett Family AI + Science Collaborative Research Program. We appreciate the insightful discussions and feedback from 3DL lab members. We also thank the computing resources, support and staff expertise at the University of Chicago.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15619–15629 (2023)
2. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: *Numerical Geometry of Non-Rigid Shapes*. Monographs in Computer Science, Springer Science & Business Media (2008)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9650–9660 (2021)
5. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624* (2025)
6. Dai, A., Nießner, M.: 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 452–468 (2018)
7. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663* (2023)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
9. Dutt, N.S., Muralikrishnan, S., Mitra, N.J.: Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4494–4504 (2024)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16000–16009 (2022)
11. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2. 5: Improved baselines for agglomerative vision foundation models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 22487–22497 (2025)
12. Hu, S., Liu, Z., Guo, M., Cai, J., Huang, J., Mu, T., Martin, R.R.: Subdivision-based mesh convolution networks. *ACM Trans. Graph.* **41**(3), 25:1–25:16 (2022), <https://doi.org/10.1145/3506694>
13. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 19729–19739 (2023)
14. Kim, V., Lipman, Y., Funkhouser, T.: Blended intrinsic maps. *ACM Transactions on Graphics (Proc. SIGGRAPH)* **30**(4) (Jul 2011)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)

16. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems* **35**, 23311–23330 (2022)
17. Kuhn, H.W.: The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* **2**(1–2), 83–97 (March 1955). <https://doi.org/10.1002/nav.3800020109>
18. Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., Pantofaru, C.: Virtual multi-view fusion for 3d semantic segmentation. In: *European conference on computer vision*. pp. 518–535. Springer (2020)
19. Lang, I., Xu, F., Decatur, D., Babu, S., Hanocka, R.: iseg: Interactive 3d segmentation via interactive attention. In: *SIGGRAPH Asia 2024 Conference Papers*. pp. 1–11 (2024)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*. pp. 19730–19742. PMLR (2023)
21. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *International conference on machine learning*. pp. 12888–12900. PMLR (2022)
22. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9704–9715 (2025)
23. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H.: Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21736–21746 (2023)
24. Liu, R., Lang, I., Hanocka, R.: Deep Feature Deformation Weights (Jan 2026). <https://doi.org/10.48550/arXiv.2601.12527>, <http://arxiv.org/abs/2601.12527>, arXiv:2601.12527 [cs]
25. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems* **32** (2019)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
27. Ovsjanikov, M., Ben-Chen, M., Solomon, J., Butscher, A., Guibas, L.: Functional maps: a flexible representation of maps between shapes. *ACM Trans. Graph.* **31**(4) (Jul 2012). <https://doi.org/10.1145/2185520.2185526>, <https://doi.org/10.1145/2185520.2185526>
28. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 815–824 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
30. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model reduce all domains into one. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12490–12500 (2024)

31. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
33. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
34. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
35. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)
36. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
37. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In: 2022 International Conference on 3D Vision (3DV). pp. 443–453. IEEE (2022)
38. Uzolas, L., Eisemann, E., Kellnhöfer, P.: Surface-aware distilled 3d semantic features. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. SA Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3757377.3763974>, <https://doi.org/10.1145/3757377.3763974>
39. Wang, Z., Wei, X., Shi, R., Zhang, X., Su, H., Liu, M.: Partuv: Part-based uv unwrapping of 3d meshes. In: ACM SIGGRAPH Asia Conference and Exhibition on Computer Graphics and Interactive Techniques (2025)
40. Wu, T., Zhang, J., Fu, X., Wang, Y., Jiawei Ren, L.P., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
41. Yang, Y., Huang, Y., Guo, Y.C., Lu, L., Wu, X., Lam, E.Y., Cao, Y.P., Liu, X.: Sampart3d: Segment any part in 3d objects. arXiv preprint arXiv:2411.07184 (2024)
42. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
43. Zhang, J., Dong, R., Ma, K.: Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2048–2059 (2023)
44. Zhu, J., Ju, Y., Zhang, J., Wang, M., Yuan, Z., Hu, K., Xu, H.: Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. arXiv preprint arXiv:2412.05268 (2024)