Zero-Shot 3D Shape Correspondence

Ahmed Abdelreheem
KAUST
Saudi Arabia
ahmed.abdelreheem@kaust.edu.sa

Abdelrahman Eldesokey
KAUST
Saudi Arabia
abdelrahman.eldesokey@kaust.edu.sa

Maks Ovsjanikov
LIX, École Polytechnique
France
maks@lix.polytechnique.fr

Peter Wonka
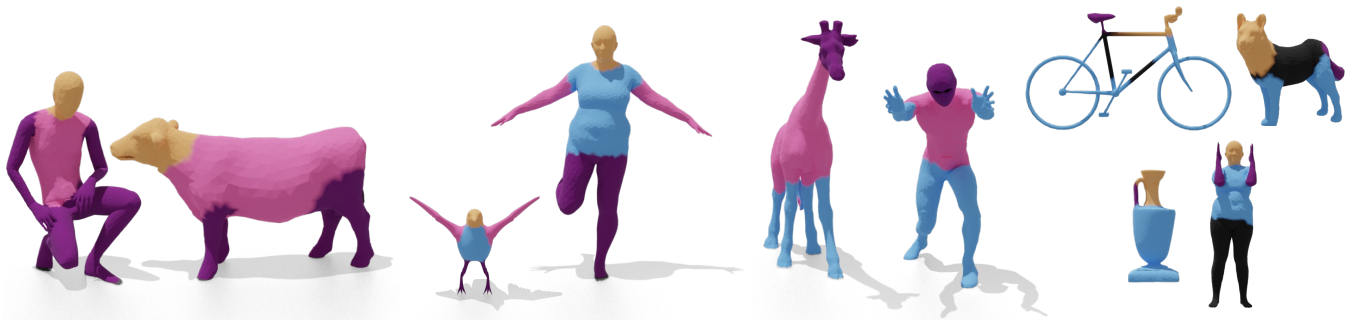KAUST
Saudi Arabia
pwonka@gmail.com

Figure 1: Our proposed approach can produce 3D shape correspondence maps for *strongly non-isometric* shapes in a zero-shot manner. Mutual semantic regions are matched and are shown in similar colors, while non-mutual regions that can not be matched are shown in black.

ABSTRACT

We propose a novel zero-shot approach to computing correspondences between 3D shapes. Existing approaches mainly focus on isometric and near-isometric shape pairs (*e.g.*, human vs. human), but less attention has been given to strongly *non-isometric* and *inter-class* shape matching (*e.g.*, human vs. cow). To this end, we introduce a fully automatic method that exploits the exceptional reasoning capabilities of recent foundation models in language and vision to tackle difficult shape correspondence problems. Our approach comprises multiple stages. First, we classify the 3D shapes in a zero-shot manner by feeding rendered shape views to a language-vision model (*e.g.*, BLIP2) to generate a list of class proposals per shape. These proposals are unified into a single class per shape by employing the reasoning capabilities of ChatGPT. Second, we attempt to segment the two shapes in a zero-shot manner, but in contrast to the co-segmentation problem, we do not require a mutual set of semantic regions. Instead, we propose to exploit the in-context learning capabilities of ChatGPT to generate two different sets of *semantic regions* for each shape and a *semantic mapping*

between them. This enables our approach to match strongly non-isometric shapes with significant differences in geometric structure. Finally, we employ the generated semantic mapping to produce coarse correspondences that can further be refined by the functional maps framework to produce dense point-to-point maps. Our approach, despite its simplicity, produces highly plausible results in a zero-shot manner, especially between *strongly non-isometric* shapes.

CCS CONCEPTS

• **Computing methodologies** → **Shape analysis**; **Neural networks**.

KEYWORDS

Zero-Shot Shape Correspondence, 3D Shape Matching, 3D Semantic Segmentation, Deep Neural Networks

1 INTRODUCTION

Shape correspondence is a fundamental task in computer vision. The objective of this task is to match two 3D shapes given some geometric representation (*e.g.*, point clouds, meshes) to produce

a region-level or point-level mapping. This mapping can be constrained based on the downstream application in terms of deformation type, density, and scope (partial or full). Examples of such downstream applications are shape interpolation, shape morphing, shape anomaly detection, 3D scan alignment, and motion capture. While early approaches for shape correspondence mainly adopted optimization-based algorithms [Van Kaick et al. 2011], the emergence of deep learning has paved the way for learning-based approaches that implicitly learn suitable representations, which can be used for efficiently solving the matching problem. These approaches can either follow a supervised [Litany et al. 2017], unsupervised [Cao et al. 2023; Halimi et al. 2019] or a self-supervised [Cao and Bernard 2023] paradigm depending on the availability and the diversity of annotated datasets.

Supervised approaches are naturally data-dependent, and they require large-scale datasets with different classes of shapes for strong generalization. On the other hand, unsupervised alternatives are class-agnostic and do not require annotated data, but they still lag behind their supervised counterparts in terms of performance [Halimi et al. 2019]. Moreover, the majority of these aforementioned approaches attempt to match near-isometric shapes of the same class, with less focus on non-isometric shapes (*e.g.*, human v.s. animal). This is mainly caused by the lack of datasets with inter-class shape pairs and the complexity of matching dissimilar shapes. To achieve a deeper understanding of 3D shapes and their relationships, it is desirable to develop methods that can generalize well both to isometric and non-isometric shape matching.

Recently, several large-scale models were introduced for different modalities such as language (*e.g.*, GPT3 [Brown et al. 2020], Bloom [Scao et al. 2022]), and vision (*e.g.*, StableDiffusion [Rombach et al. 2022], DALLE-2 [Ramesh et al. 2022]). These models are usually referred to as *foundation models*, and they have a broad knowledge of their domains since they were trained on large amounts of data. There are even ongoing efforts to connect these models to build a bridge between different modalities, such as Visual ChatGPT [Wu et al. 2023], and MiniGPT-4 [Zhu et al. 2023]. Unfortunately, it is still challenging to build similar models for modalities with limited amounts of data, such as 3D shapes. Therefore, a plausible approach is to employ existing models for language and vision to solve problems for other data-limited modalities.

Motivated by this, we attempt to exploit existing foundation models to perform zero-shot shape correspondence with no additional training or finetuning. To address this problem, we identified the following three key problems. First, we would like to predict the class of each of the two shapes in question given only their 3D meshes. We achieve this through zero-shot shape classification by feeding rendered views of the two shapes into a language-visual model, BLIP2 [Li et al. 2023], to obtain object class proposals. Then, we use ChatGPT to merge these proposals into a single class per object. Second, we produce a semantic region set per shape, and a semantic mapping between the sets by exploiting the in-context learning [Brown et al. 2020] capabilities of ChatGPT [OpenAI 2021]. Afterward, we introduce a zero-shot 3D semantic segmentation method based on the recent large-scale models DINO [Caron et al. 2021] and Segment-Anything (SAM) [Kirillov et al. 2023], which we denote as *SAM-3D*. Our method only requires a shape mesh and its corresponding semantic region set as input. Finally, the

semantic mapping is used to provide coarse correspondences between the two shapes and a finer map can be produced, if needed, by employing the functional map framework [Ovsjanikov et al. 2012]. Remarkably, although functional maps are geared towards near-isometric shape pairs, we observe that it is possible to obtain high-quality dense maps given an initialization from SAM-3D, even across some challenging non-isometric shapes.

Since we propose a new scheme for solving the shape correspondence problem, we introduce several evaluation metrics to evaluate the performance of different intermediate tasks in our pipeline, such as zero-shot object classification, semantic region generation, and semantic segmentation. We also create a new benchmark that includes *strongly non-isometric* shape pairs (*e.g.*, humans vs. animals) that we denote as (SNIS) in order to test the generalization capabilities of our proposed approach. Experiments on the new benchmark show that our approach, despite being zero-shot, performs very well on non-isometric shape pairs.

In summary, we make the following contributions:

- We propose a novel solution to 3D shape correspondence that computes results in a zero-shot manner.
- To the best of our knowledge, we introduce a first zero-shot joint 3D semantic segmentation technique that does not start with a mutual set of semantic regions, and it requires only the shape meshes while exploiting language-vision models to generate shape-specific semantic regions.
- We introduce a benchmark for shape correspondence which includes strongly non-isometric shape pairs, as well as evaluation metrics for different stages of the proposed pipeline.

2 RELATED WORK

In this section, we give a brief overview of shape correspondence literature, large-scale models, and 3D semantic segmentation. For shape correspondence, we focus only on relevant deep learning-based approaches, and we refer the reader to [Sahillioğlu 2020; Van Kaick et al. 2011] for a comprehensive survey of earlier registration-based and similarity-based approaches.

2.1 Deep Learning-Based Shape Correspondence

Convolutional Neural Networks (CNNs) by nature are not directly applicable to non-rigid shapes due to the lack of shift-invariance property in non-Euclidean domains. Wei *et al.* [Wei et al. 2016] circumvented this by training on depth maps of shapes that are being matched, and produced pixel-wise classification maps for each point in the object. Wu *et al.* [Wu et al. 2015] generated volumetric representations from depth maps, and used 3D CNNs to process them. However, these methods do not capture all shape deformations, since they treat shapes as Euclidean structures. Alternatively, other approaches tried to generalize Convolutional Neural Networks (CNNs) to non-Euclidean manifolds. Masci *et al.* [Masci et al. 2015] introduced Geodesic CNNs that allowed constructing local geodesic polar coordinates that are analogous to patches in images. Similarly, Boscaini *et al.* [Boscaini et al. 2015] proposed localized spectral CNNs to learn class-specific local descriptors based on a generalization of windowed Fourier transform. This was followed by another generalization of CNNs in [Boscaini et al. 2016]
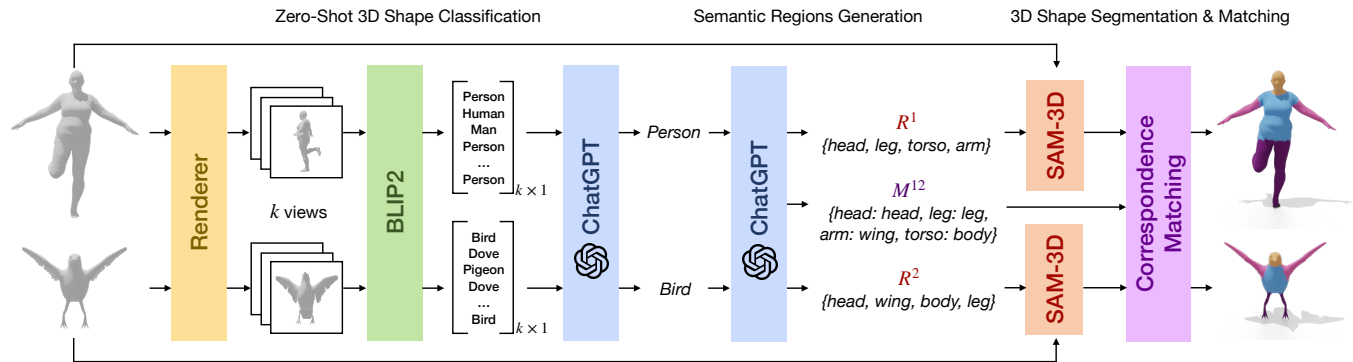
Figure 2: Our proposed approach has three main components: (1) Zero-shot 3D shape classification: By feeding rendered k views of each shape to a BLIP2 [Li et al. 2023] model to generate class proposal lists. The proposals are unified using ChatGPT to produce a single class per shape. (2) Semantic region/mapping generation: In-context learning capabilities of ChatGPT are employed to produce a semantic region set for each shape and a semantic mapping between them. (3) Zero-shot 3D semantic segmentation: our proposed SAM-3D uses the semantic regions to segment the shapes, and the mapping is used to produce a sparse correspondence map that can be densified further using the functional maps framework [Ovsjanikov et al. 2012].

denoted as Anisotropic CNNs that replace the conventional convolutions with a projection operator over a set of oriented anisotropic diffusion kernels. All these approaches allowed extracting local descriptors at each point on deformable shapes, and eventually perform shape correspondence by similarity-matching.

Another category of approaches includes the matching computations in the learning process and can find shape correspondences directly from a CNN. Litany *et al.* [Litany et al. 2017] proposed a structured prediction model in the functional maps space [Ovsjanikov et al. 2012] that takes in dense point descriptor for the two shapes, and produces a soft correspondence map. Halimi *et al.* [Halimi et al. 2019] transforms [Litany et al. 2017] into an unsupervised setting by replacing the point-wise correspondences with geometric criteria that are optimized, eliminating the need for annotated data. Donati *et al.* [Donati et al. 2020] proposed an end-to-end pipeline that computes local descriptors from the raw 3D shapes, and employs a regularized functional maps to produce dense point-to-point correspondences. Their method requires less data to train and generalizes better than its supervised counterparts. Li *et al.* [Li et al. 2022a] employed a regularized contrastive learning approach to learn robust point-wise descriptors that can be used to match shapes. We deviate from all these approaches, and we tackle the problem from a peculiar zero-shot perspective that exploits the emerging large-scale models in language and vision.

2.2 Large-Scale Models

Several models that are trained on large-scale datasets were introduced recently for different modalities, given the advances in deep architectures design and computational capabilities. For instance, Large-Scale Language Models (LLMs) such as T5 [Raffel et al. 2020], BLOOM [Scao et al. 2022], GPT-3 [Brown et al. 2020], and Instruct-GPT [Ouyang et al. 2022]; vision models StableDiffusion [Rombach et al. 2022], and DALLE-2 [Ramesh et al. 2022]). LLMs have outstanding capabilities in understanding textual data, but they lack any understanding of natural images. Recent methods try to build cross-modal vision-language models that incorporate the capacities

of both models. Visual-ChatGPT [Wu et al. 2023] is one example that combines ChatGPT with many vision foundation models that are managed using a prompt manager that allows better combination and interaction. MiniGPT-4 [Zhu et al. 2023] pursues a similar endeavor by attempting to align frozen LLM with a visual encoder through a projection layer. To further improve language coherence, they finetune the model on a well-aligned dataset using a conversational template. BLIP2 [Li et al. 2023] bootstraps vision-language models through efficient pre-training from off-the-shelf models. We employ these models to achieve zero-shot 3D shape classification and to generate shape-specific semantic regions that can then be utilized to perform zero-shot 3D semantic segmentation to find shape correspondences.

2.3 Zero-shot 3D Semantic Segmentation

Zero-shot 3D semantic segmentation is an active research topic that attempts to segment volumes or point clouds given some textual labels or descriptors [Abdelreheem et al. 2022; Chen et al. 2022; Decatur et al. 2023; Ding et al. 2022; Liu et al. 2022; Michele et al. 2021; Naeem et al. 2021; Zhang et al. 2021; Zhu et al. 2022]. On a different note, there exist many approaches that are based on Neural Radiance Fields (NeRFs) [Lombardi et al. 2019; Mildenhall et al. 2020], which try to produce full semantic maps of 3D scenes by exploiting 3D density fields from NeRFs [Fan et al. 2022; Fu et al. 2022; Kundu et al. 2022; Siddiqui et al. 2022; Tschernezki et al. 2022; Vora et al. 2021; Zhi et al. 2021] These two categories of approaches can be combined to perform zero-shot 3D segmentation of volumetric scenes by incorporating zero-shot 2D segmentation networks (*e.g.*, [Li et al. 2022c]) into NeRFs [Goel et al. 2022; Kobayashi et al. 2022; Shafiullah et al. 2022] given some textual labels. SATR [Abdelreheem et al. 2023] showed that replacing 2D segmentation networks with 2D object detector networks yields marginally better results. Inspired by this, we propose to combine the object detector DINO [Caron et al. 2021] with Segment-Anything (SAM) [Kirillov et al. 2023] to perform zero-shot 3D shape segmentation.

3 METHOD

In zero-shot 3D shape correspondence, the input is a pair of 3D shapes (S^1, S^2), where each shape S^i is represented using triangular meshes with vertices $V^i \in \mathbb{R}^{|V^i| \times 3}$, and faces $F^i \in \mathbb{R}^{|F^i| \times 3}$. The number of vertices/faces in S^1 are not necessarily equal to that of S^2. The desired output is a point-to-point correspondence map $C \in \mathbb{R}^{|V^2| \times |V^1|}$ that contains matching scores between vertices of V^2 and V^1. Note that no other information is provided about the shape such as the shape class or semantic region names, and it is desired to perform shape correspondence in a zero-shot manner with no training or fine-tuning. To this end, we propose a new setting to tackle this problem that consists of three modules, as illustrated in Figure 2. First, we perform *zero-shot 3D object classification* on the shapes to obtain an object class per shape using a large-scale visual-language model (3.1). Afterward, a set of *semantic region names* per shape is generated using an LLM (3.2). Next, *zero-shot 3D semantic segmentation* is performed given the semantic region names (3.3). Finally, dense correspondence maps can be calculated using functional maps [Ren et al. 2018a] (3.4). We explain these components in more detail in the following sections.

3.1 Zero-Shot 3D Shape Classification

Initially, we need to identify the classes of the 3D shapes. Existing zero-shot 3D shape classification approaches [Cheraghian et al. 2020, 2019] can predict a limited set of unseen classes, but do not generalize when the unseen set is unlimited. In our case, there is no prior knowledge about the classes of the shapes, and therefore, existing approaches for zero-shot 3D classification can not be employed. To alleviate this, we propose to employ a language-vision foundation model (*e.g.*, BLIP2 [Li et al. 2023]) that exploits the generalization capabilities of Large-Scale Language Models (LLMs) to reason about 2D images.

For each shape in the pair (S^1, S^2), we render k views, where viewpoints are sampled uniformly around the shape for a wide coverage. We set the elevation angles to $\{-45°, 0°, 45°\}$, the azimuth angles to $\{0°, 90°, 180°, 270°\}$, and the radius to 2 length units, where each shape is centered around the origin and scaled to be inside a unit sphere. Then, we feed these k rendered views per shape to BLIP [Li et al. 2023] to obtain k object class proposals. A natural choice would be to perform majority voting on these predictions to get a single class type. However, it is not straightforward to achieve this for textual labels, given that the list of class proposals can include synonyms and adjectives. Figure 2 shows an example of this situation. Therefore, we exploit the reasoning capabilities of a ChatGPT agent to unify the responses and obtain a single class label per shape. We show examples of the used prompts in the supplementary materials.

3.2 Semantic Region Generation and Matching

Zero-shot 3D semantic segmentation approaches require a set of semantic labels as an input together with the 3D shape. Our problem is more difficult than traditional co-segmentation, because the two input shapes may not share the same region names. For this reason, We need to obtain two sets (R^1, R^2) of the possible names of semantic regions present in each shape in the input pair (S^1, S^2). Afterward, we attempt to match, whenever possible, between the

semantic regions defined in R^1 and R^2, where a single semantic region in R^1 can be matched to one or more semantic regions in R^2. For instance, the legs of a dog can be matched to both the arms and the legs of a person. We exploit the in-context learning [Brown et al. 2020] capabilities of LLMs to achieve this. In-context learning is the process by which a model understands a certain task and provides an adequate response to the required task. LLM models are indeed good in-context learners, allowing them to perform well on a wide range of tasks without explicit fine-tuning. The idea is when asking the model to solve a task given a certain input, we include a few (input, expected output) pairs as examples in the input prompt. We employ ChatGPT for this purpose, and we query two sets of semantic regions (R^1, R^2) for the two shape classes in question, and a mapping between the two sets $M^{12} : R^1 \leftrightarrow R^2$. Figure 2 shows an example of such a mapping. We refer the reader to the supplementary material for further details on formulating the textual prompts for ChatGPT.
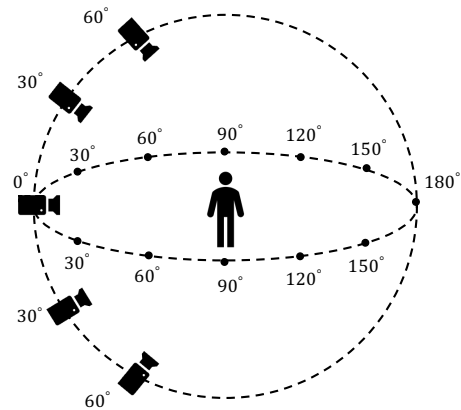

Figure 3: Sampling strategy for the v rendered views that are used to perform zero-shot 3D semantic segmentation. Note that we use three different radii of (2, 1.75, 1.5) from the origin to produce a total of 180 views.

3.3 Zero-Shot 3D Semantic Segmentation

After generating the two sets of semantic regions (R^1, R^2), we can use them to perform zero-shot 3D semantic segmentation. Despite the fact that the recent object segmentation model Segment-Anything (SAM) [Kirillov et al. 2023] is powerful, its text-guided segmentation is still limited. The Groudning-DINO object detector [Liu et al. 2023] on the other hand, can perform 2D object detection for a large number of classes. Therefore, we propose to combine Groudning-DINO with SAM to perform zero-shot 3D semantic segmentation, and we denote this hybrid approach as *SAM-3D*. It is possible to use other object detectors such as GLIP [Li et al. 2022d] to obtain bounding boxes of parts, but we employ Gounding-DINO as it performs better than its counterparts as shown in [Liu et al. 2023].

We start by rendering a large number of viewpoints v sampled uniformly to cover the whole shape as illustrated in Figure 3. For each rendered view, we feed it to DINO to detect a bounding box for each semantic region in R^i. Afterward, we feed the detected

bounding boxes with the rendered viewpoints to SAM to provide segmentation maps for each semantic region. We define a matrix $X^i \in \mathbb{R}^{|F^i| \times |R^i|}$ that is initialized with zeros, and we use it to accumulate scores for each face in F^i for each semantic region in R^i. Finally, each face is assigned a label by selecting the highest score in each row j of X^i yielding FL^i:

$$FL^i = \arg\max_j X^i[j, :] \tag{1}$$

Once the 3D segmentation vectors FL^i are computed, the segmentation maps are matched between the two shapes using the mapping M^{12} to produce a coarse correspondence map.

3.4 Zero-Shot Dense Shape Correspondence

To produce dense point-to-point shape correspondence, we employ a functional maps-based approach. We use the overall strategy based on associating functional descriptors with region correspondences, as described in the original functional maps paper [Ovsjanikov et al. 2012] and then implemented in [Kleiman and Ovsjanikov 2019] and [Ren et al. 2018b]. Specifically, given region-wise correspondences, we formulate an optimization problem to compute a functional map. The optimization problem is obtained first by formulating functional constraints using the sum of the WKS descriptor [Aubry et al. 2011] of the points in the segment. We combine these with the Laplace-Beltrami commutativity regularization into a single system and solve it to find the optimal functional mapping matrix $\mathbf{C}$. Finally, we convert the computed functional map $\mathbf{C}$ to a point-to-point map and iteratively refine it using the BCICP refinement strategy [Ren et al. 2018b]. All parameters we use in our approach, including the way we formulate and solve the optimization problem, are exactly the same as in [Ren et al. 2018b] with the only difference being that the regions that are matched are produced by our pipeline rather than the ones produced by [Kleiman and Ovsjanikov 2019]. This gives, as output, a dense point-to-point correspondence between each 3D shape pair. Interestingly, while functional maps primarily target near-isometric shapes, our initialization with SAM-3D allows us to generate high-quality dense maps even for challenging non-isometric shape pairs, as will be shown in Figure 5. Nevertheless, artifacts in point-to-point maps can occur due to the use of the functional map framework, and we leave the development of a correspondence densification technique that can adapt to strongly non-isometric shapes, to future work.

4 EXPERIMENTS

In this section, we evaluate our proposed approach, and we introduce our new dataset for *strongly non-isometric* shape matching (SNIS). Moreover, since we propose a new strategy for solving shape correspondence problems, we introduce some evaluation metrics for different components of the pipeline.

4.1 Strongly Non-Isometric Shapes Dataset (SNIS)

Existing shape correspondence datasets [Bogo et al. 2014; Zuffi et al. 2017] usually encompass a single category of objects (*e.g.*, humans or animals), and they employ template models to derive dense correspondences to alleviate annotation workload. To facilitate the

development of approaches that can generalize to non-isometric shape matching, we introduce a new dataset with mixed shape pairs from existing isometric datasets, *e.g.*, FAUST [Bogo et al. 2014] (humans), SMAL [Zuffi et al. 2017] (animals), and DeformingThings4D [Li et al. 2021] (humanoid objects). For each pair of shapes, we annotate 34 keypoint correspondences between the shapes as well as a dense segmentation map. Figure 4 shows an illustration for these annotations. For the FAUST dataset [Bogo et al. 2014], we use a similar approach of annotation as described in [Abdelreheem et al. 2023] that includes segmentation maps for all the available 100 shapes.

Our SNIS dataset includes 250 shape pairs, where the first shape is either from FAUST or DeformingThings4D, and the second is from SMAL. The included classes are: {"cougar", "cow", "dog", "fox", "hippo", "horse", "lion", "person", "wolf"}. Note that it is desirable to include other categories of objects from diverse datasets such as SHREC09 [Godil et al. 2009]. Unfortunately, we find that this demands significant manual annotation effort, particularly with point-to-point dense annotation, which is time-consuming and labor-intensive. However, we demonstrate the generalization capabilities of our approach by showing some qualitative examples from SHREC09 in section 4.8.

4.2 Metrics

For the final dense shape correspondence map, we use the standard average geodesic error as in [Halimi et al. 2019; Litany et al. 2017]. We describe the newly proposed metrics below.

Zero-Shot Object Classification Accuracy (ZSClassAcc) To evaluate if the predicted object class in 3.1 is accurate, we compare it against the ground-truth shape label. However, since LLM-based approaches can predict several synonyms for each class (*e.g.*, human and person), the standard classification accuracy becomes infeasible.

Therefore, we propose to generate a set of synonyms for each object class in the dataset from WordHoard [1]. Whenever a class prediction matches any of the synonyms, it is counted as a correct prediction. Eventually, the accuracy is calculated as a standard binary classification accuracy.

Semantic Regions Generation F1-Score (SRGen-F1) Similar to the previous metric, we evaluate the generated semantic regions as a multi-class classification problem. Regions that are matched with the ground truth count as True Positives (TP), ground truth regions that were not predicted count as False Negative (FN), and predicted regions that do not exist in the ground truth count as False Positives (FP). Finally, a standard F1-Score is calculated as:

$$\text{SRGen-F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{2}$$

Semantic Regions Prediction IoU (SRIoU) To evaluate the quality of semantic segmentation for different semantic regions in R^1 and R^2, we calculate the average intersection-over-union over different regions and shapes as follows:

$$I^{12} = \frac{I^1 + I^2}{2} \tag{3}$$

---

[1] https://wordhoard.readthedocs.io

Method	Acc.
Voting	44.80%
ZSM (ours)	**73.90%**

Table 1: A comparison in terms of Zero-Shot Object Classification Accuracy (ZSClassAcc) between our approach and the standard majority voting.
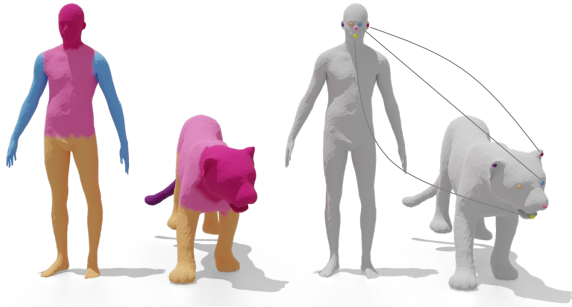

Figure 4: Keypoint correspondences and segmentation maps for the proposed Strongly Non-Isometric Shapes (SNIS) dataset. We provide 34 keypoint annotations, but we only show a few here for clarity.

$$I^i = \frac{1}{|R^i|} \sum_{r=1}^{|R^i|} IoU_r^i \qquad (4)$$

where IoU_r^i is the intersection-over-union for region r in shape i compared to the groundtruth segmentation.

Keypoint Label Matching Accuracy (KPLabelAcc) Since we provide keypoint annotations in our proposed SNIS dataset, we can evaluate the shape matching accuracy at these keypoints. For each shape i, we define keypoint indices vector $P^i \in \mathbb{R}^{34 \times 1}$, which stores vertex indices from V^i for the annotated keypoints. Given faces labels FL^i from (1), we can generate labels for vertices as well that lie on these faces, and we denote them as VL^i. The keypoint label matching accuracy is then calculated between VL^1, VL^2, and the groundtruth labels VL^{GT} as:

$$\text{KPLabelAcc} = \frac{1}{|P|} \sum_{j=1}^{|P|} VL^1[P_j^1] \ \wedge \ VL^2[P_j^2] \ \wedge \ VL^{GT}[P_j^{GT}] \quad (5)$$

where j refers to elements in the vector and $\wedge$ is the semantic AND operator, which means the three integers should share the same semantic label. This metric measures if the keypoints are matched correctly, and that they were assigned the correct label. Next, we compare our proposed approach against some baselines in terms of these metrics.

4.3 Zero-Shot 3D Shape Classification Results

Baseline We calculate a majority voting between all classification proposals generated by BLIP2. Table 1 shows that our proposed approach based on ChatGPT performs significantly better than the standard voting.

4.4 Semantic Regions Generation and Matching Results

Baseline We use BLIP2 [Li et al. 2023] model as a baseline, where we feed it with the k rendered views from section 3.1 to query semantic regions and mapping.

We report the results in Table 2 for the generated semantic regions in terms of the SRGen-F1 metric. Surprisingly, our proposed approach that employs ChatGPT outperforms BLIP2 with a huge margin despite the fact that our approach does not have access to the rendered images. This demonstrates the in-context learning capabilities of ChatGPT. We can also compare the semantic mapping M^{12} in the same fashion as the semantic regions by matching the keys and values of the generated mapping with those of the groundtruth mapping. However, we did not succeed in obtaining a valid mapping from BLIP2 as it accepts only one image at a time, and the prompt is limited to 512 tokens, which can be insufficient for in-context learning prompts. Therefore, we report only our scores in Table 2.

4.5 3D Semantic Segmentation Results

We compare against the recently released zero-shot 3D semantic segmentation approach SATR [Abdelreheem et al. 2023] that employs only 2D object detectors. SATR differs from our approach mainly in using a 2D object detector instead of SAM.

Table 3 shows that our approach outperforms SATR in terms of the SRIoU metric. We believe that is caused by employing 2D semantic segmentation masks from SAM, which are less prone to error when transferring the segmentation information to the 3D space, compared to bounding boxes. We also show selected qualitative examples in Figure 8, where it is clear that our proposed SAM-3D provides a more accurate and well-localized segmentation compared to SATR. We show in Figure 7 the generalization capability of SAM-3D on daily objects.

4.6 Keypoints Matching Results

We compare the keypoints matching results from our proposed approach to those obtained by replacing SAM-3D with SATR [Abdelreheem et al. 2023]. Table 3 shows that our approach outperforms SATR in terms of KPLabelAcc, which demonstrates that it provides better segmentation maps with more accurate labels.

4.7 Dense Shape Correspondence Comparison

Our approach generally produces sparse shape correspondences, as illustrated in Figures 1 and 6. However, dense correspondence maps can be produced by using the functional maps framework as described in Section 3.4. We provide a comparison for dense shape

Method	Semantic Regions Avg. F1-Score	Mapping Avg. F1-Score
BLIP2 [Li et al. 2023]	41.28%	-
ZSM (ours)	**80.31%**	**65.05%**

Table 2: A comparison of generated semantic regions and mapping in terms of SRGen-F1. Our proposed approach based on ChatGPT outperforms BLIP2 on semantic region generation. However, we were not able to produce valid mappings from BLIP2, so we only mention ours.

correspondence maps when using our proposed SAM-3D in comparison with the segmentation model SEG [Kleiman and Ovsjanikov 2019] to initialize the functional maps framework BCICP [Ren et al. 2018b] Table 5 shows that the use of SAM-3D outperforms SEG in terms of average geodesic error with a large margin. We also provide a qualitative comparison in Figure 5, which shows that our approach provides more accurate correspondences, and does not suffer from region discontinuities as SEG. Note that existing supervised approaches are typically trained on objects of the same category, e.g., humans or animals, where there is enough annotated training data. Therefore, they cannot be directly evaluated on SNIS without adapting and retraining these approaches, which might not always be feasible.

4.8 Generalization to Other Datasets

To examine if our proposed approach generalizes to other datasets with highly unrelated shapes, we include some objects from the SHREC09 [Godil et al. 2009], and 3D-CoMPaT [Li et al. 2022b] datasets. We form pairs of shapes where the first item is from SNIS, and the second is from SHREC09 or 3D-CoMPaT. Figures 10 and 11 show these examples. Our method was able to produce plausible results when matching a human with a chair, where the legs were matched correctly, and the seat was matched to the rest of the human body. A horse was also matched to a tricycle, where the horse limbs were matched to the wheels, the head to the handle, and the tail to the seat. We show in Figure 9 examples where the pairs are from daily objects [Chang et al. 2015; Xiang et al. 2020]. These examples demonstrate the reasoning capabilities of our approach, even when the shape pairs are not related. We provide more detailed qualitative examples in Figure 11.

4.9 Impact of Varying Number of Viewpoints

We examine the effect of changing the number of rendered views k, and v in the proposed zero-shot 3D object classifier and in SAM-3D, respectively. Table 4 shows that both the classification and segmentation accuracy improve when increasing the number of views. We do not consider higher values for computational efficiency.

5 CONCLUSION

We proposed a novel zero-shot approach for 3D shape correspondence. Our approach exploited the capabilities of recently emerged language and vision foundation models to match challenging nonisometric shape pairs. There are two key differences in our work to traditional co-segmentation. First, we do not require the region
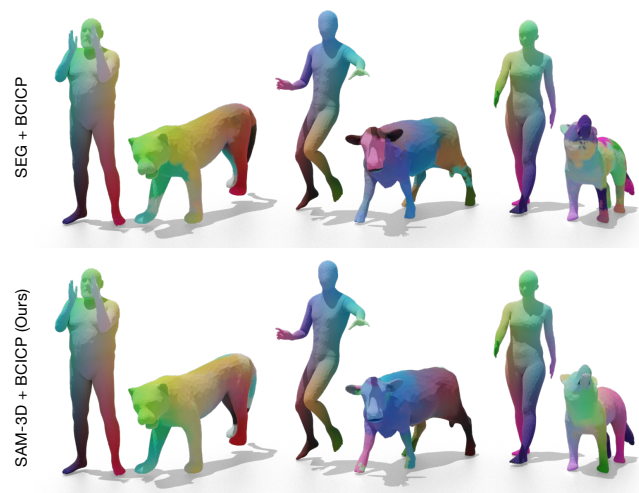

Figure 5: Dense shape correspondences generated by the functional maps framework BCICP [Ren et al. 2018b] when initialized with our proposed SAM-3D in comparison to SEG [Kleiman and Ovsjanikov 2019].

Method	SRIoU	KPLabelAcc
SATR [Abdelreheem et al. 2023]	69.98%	56.60%
SAM-3D (ours)	**73.55%**	**59.72%**

Table 3: A comparison with an existing zero-shot 3D semantic segmentation approach in terms of SRIoU and KPLabelAcc
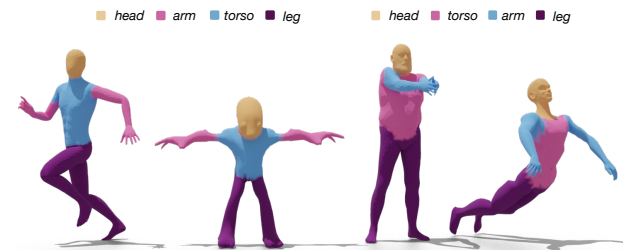

Figure 6: Qualitative results for two input pairs of shapes within the same class (left and right columns). The shapes are from DeformingThings4D, FAUST, and SHREC09 datasets.

names to be known in advance. Second, our approach does not require a mutual set of semantic regions and generates shape-specific sets, and a semantic mapping between them instead, enabling it to match diverse shape pairs. We also introduced a new dataset for strongly non-isometric shapes (SNIS) as well as evaluation metrics for each stage in our pipeline to facilitate the development and evaluation of future methods.

Limitations and Future Work Our approach can match coarse semantic regions such as main body parts (*e.g.*, head, torso, and legs). In future work, it would be desirable to produce finer regions in such as eyes, mouth, and hands. This is challenging because the current image-based segmentation models are not able to provide

k	ZSClassAcc		v	SRIoU	KPLabelAcc
1	60.10%		30	69.84%	54.97%
5	62.76%		60	72.06%	57.63%
24	**73.90%**		120	73.06%	58.89%
			180	**73.55%**	**59.72%**

Table 4: Ablation study on the effect of changing the number of viewpoints on the zero-shot object classification in terms of ZSClassAcc, and the zero-shot 3D semantic segmentation in terms of SRIoU and KPLabelAcc.

Method	Avg geodesic error
SEG + BCICP [Ren et al. 2018b]	0.41
SATR + BCICP (ours)	0.37
SAM-3D + BCICP (ours)	**0.36**

Table 5: A comparison for dense shape correspondence in terms of average geodesic error. Initializing the BCICP framework with segmentation from SAM-3D yields significantly better results.

fine-grained segmentation for renderings of meshes without textures. Foundation models in machine learning are helpful for a wide range of tasks. In the future, it would also be interesting to design foundation models that can map 3D shapes, images, and text to a common latent space. Finally, adapting functional maps to handle strongly non-isometric shape pairs, starting from high-quality segment matches, is another interesting problem for future work.

REFERENCES

Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. 2022. ScanEnts3D: Exploiting Phrase-to-3D-Object Correspondences for Improved Visio-Linguistic Models in 3D Scenes. *ArXiv* abs/2212.06250 (2022).

Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. 2023. SATR: Zero-Shot Semantic Segmentation of 3D Shapes. arXiv:2304.04909 [cs.CV]

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.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. 2014. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3794–3801.

D. Boscaini, J. Masci, S. Melzi, M. M. Bronstein, U. Castellani, and P. Vandergheynst. 2015. Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks. *Computer Graphics Forum* 34, 5 (2015), 13–23. https://doi.org/10.1111/cgf.12693 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12693.

Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. 2016. Learning shape correspondence with anisotropic convolutional neural networks. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/228499b55310264a8ea0e27b6e7c6ab6-Abstract.html

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

Dongliang Cao and Florian Bernard. 2023. Self-Supervised Learning for Multimodal Non-Rigid 3D Shape Matching. https://doi.org/10.48550/arXiv.2303.10971 arXiv:2303.10971 [cs].

Dongliang Cao, Paul Roetzer, and Florian Bernard. 2023. Unsupervised Learning of Robust Spectral Shape Matching. *arXiv preprint arXiv:2304.14419* (2023).

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.

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).

Runnan Chen, Xinge Zhu, Nenglun Chen, Wei Li, Yuexin Ma, Ruigang Yang, and Wenping Wang. 2022. Zero-shot Point Cloud Segmentation by Transferring Geometric Primitives. *arXiv preprint arXiv:2210.09923* (2022).

Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. 2020. Transductive zero-shot learning for 3d point cloud classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 923–933.

Ali Cheraghian, Shafin Rahman, and Lars Petersson. 2019. Zero-shot learning of 3d point cloud objects. In *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 1–6.

Dale Decatur, Itai Lang, and Rana Hanocka. 2023. 3d highlighter: Localizing regions on 3d shapes via text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20930–20939.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. 2022. Language-driven Open-Vocabulary 3D Scene Understanding. *arXiv preprint arXiv:2211.16312* (2022).

Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. 2020. Deep Geometric Functional Maps: Robust Feature Learning for Shape Correspondence. 8592–8601. https://openaccess.thecvf.com/content_CVPR_2020/html/Donati_Deep_Geometric_Functional_Maps_Robust_Feature_Learning_for_Shape_Correspondence_CVPR_2020_paper.html

Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. 2022. NeRF-SOS: Any-View Self-supervised Object Segmentation on Complex Scenes. *arXiv preprint arXiv:2209.08776* (2022).

Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. 2022. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224* (2022).

Afzal Godil, Helin Dutagaci, Ceyhun Burak Akgül, Apostolos Axenopoulos, Benjamín Bustos, Mohamed Chaouch, Petros Daras, Takahiko Furuya, Sebastian Kreft, Zhouhui Lian, Thibault Napoléon, Athanasios Mademlis, Ryutarou Ohbuchi, Paul L. Rosin, Bülent Sankur, Tobias Schreck, Xianfang Sun, Masaki Tezuka, Anne Verroust-Blondet, Michael Walter, and Yücel Yemez. 2009. SHREC 2009 - Generic Shape Retrieval Contest.

Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. 2022. Interactive Segmentation of Radiance Fields. *arXiv preprint arXiv:2212.13545* (2022).

Oshri Halimi, Or Litany, Emanuele Rodolà Rodolà, Alex M. Bronstein, and Ron Kimmel. 2019. Unsupervised Learning of Dense Shape Correspondence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4365–4374. https://doi.org/10.1109/CVPR.2019.00450 ISSN: 2575-7075.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

Yanir Kleiman and Maks Ovsjanikov. 2019. Robust structure-based shape correspondence. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 7–20.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. 2022. Decomposing NeRF for Editing via Feature Field Distillation. *arXiv preprint arXiv:2205.15585* (2022).

Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. 2022. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12871–12881.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. 2022c. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546* (2022).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]

Lei Li, Souhaib Attaiki, and Maks Ovsjanikov. 2022a. SRFeat: Learning Locally Accurate and Globally Consistent Non-Rigid Shape Correspondence. In *International Conference on 3D Vision (3DV)*. IEEE.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022d. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 2021. 4DComplete: Non-Rigid Motion Estimation Beyond the Observable Surface. arXiv:2105.01905 [cs.CV]

Yuchen Li, Ujjwal Upadhyay, Habib Slim, Ahmed Abdelreheem, Arpita Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 2022b. 3D CoMPaT: Composition of Materials on Parts of 3D Things. In *European Conference on Computer Vision*.

Or Litany, Tal Remez, Emanuele Rodolà, Alex M. Bronstein, and Michael M. Bronstein. 2017. Deep Functional Maps: Structured Prediction for Dense Shape Correspondence. https://doi.org/10.48550/arXiv.1704.08686 arXiv:1704.08686 [cs].

Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. 2022. PartSLIP: Low-Shot Part Segmentation for 3D Point Clouds via Pretrained Image-Language Models. *arXiv preprint arXiv:2212.01558* (2022).

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751* (2019).

Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. 2015. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*. 37–45.

Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. 2021. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 992–1002.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.

Muhammad Ferjad Naeem, Evin Pinar Ornek, Yongqin Xian, Luc Van Gool, and Federico Tombari. 2021. 3D Compositional Zero-shot Learning with DeCompositional Consensus. *ArXiv* abs/2111.14673 (2021). https://api.semanticscholar.org/CorpusID:244714384

OpenAI. 2021. *GPT-3.5 Language Model*. OpenAI. https://www.openai.com/research/gpt-3 Accessed: May 21, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

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* 31, 4 (July 2012), 30:1–30:11. https://doi.org/10.1145/2185520.2185526

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Jing Ren, Adrien Poulenard, Peter Wonka, and Maks Ovsjanikov. 2018a. Continuous and orientation-preserving correspondences via functional maps. *ACM Transactions on Graphics (TOG)* 37 (2018), 1 – 16.

Jing Ren, Adrien Poulenard, Peter Wonka, and Maks Ovsjanikov. 2018b. Continuous and Orientation-preserving Correspondences via Functional Maps. arXiv:1806.04455 [cs.GR]

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.

Yusuf Sahillioğlu. 2020. Recent advances in shape correspondence. *The Visual Computer* 36, 8 (Aug. 2020), 1705–1721. https://doi.org/10.1007/s00371-019-01760-0

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. 2022. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663* (2022).

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. 2022. Panoptic Lifting for 3D Scene Understanding with Neural Fields. arXiv:arXiv:2212.09802

Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. 2022. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494* (2022).

Oliver Van Kaick, Hao Zhang, Ghassan Hamarneh, and Daniel Cohen-Or. 2011. A survey on shape correspondence. In *Computer graphics forum*, Vol. 30. Wiley Online Library, 1681–1707.

Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. 2021. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260* (2021).

Lingyu Wei, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li. 2016. Dense Human Body Correspondences Using Convolutional Networks. https://doi.org/10.48550/arXiv.1511.05904 arXiv:1511.05904 [cs]

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671* (2023).

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3D ShapeNets: A Deep Representation for Volumetric Shapes. https://doi.org/10.48550/arXiv.1406.5670 arXiv:1406.5670 [cs]

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. 2020. SAPIEN: A SimulAted Part-based Interactive ENvironment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2021. PointCLIP: Point Cloud Understanding by CLIP. *arXiv preprint arXiv:2112.02413* (2021).

Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. 2021. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *Proceedings of theIEEE International Conference on Computer Vision*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. 2022. PointCLIP V2: Adapting CLIP for Powerful 3D Open-world Learning. *ArXiv* abs/2211.11682 (2022). https://api.semanticscholar.org/CorpusID:253735373

Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 2017. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
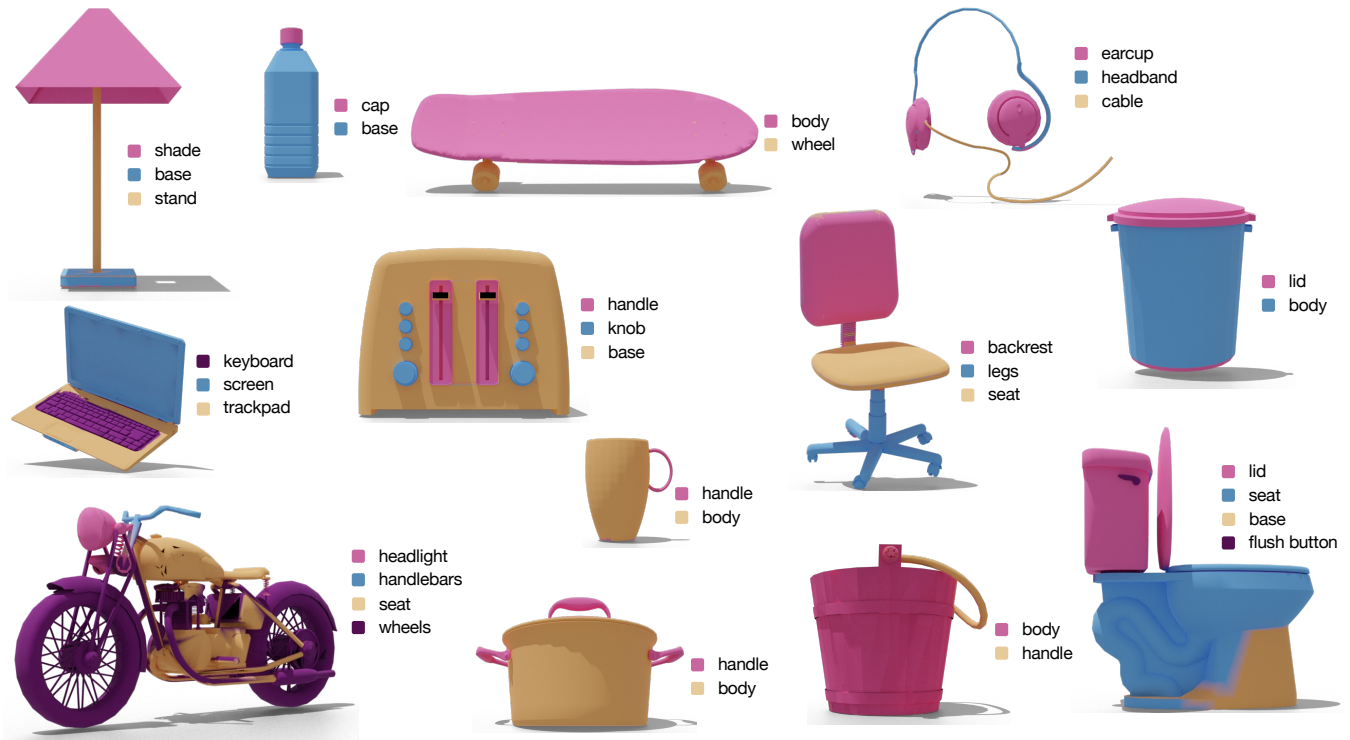
Figure 7: Qualitative examples of SAM-3D for zero-shot semantic segmentation of daily objects. The input textual prompts are provided by ChatGPT. SAM-3D can predict fine-grained parts such as the knobs of a toaster, the cap of a bottle, the flush button of a toilet, or the cable in the headset.
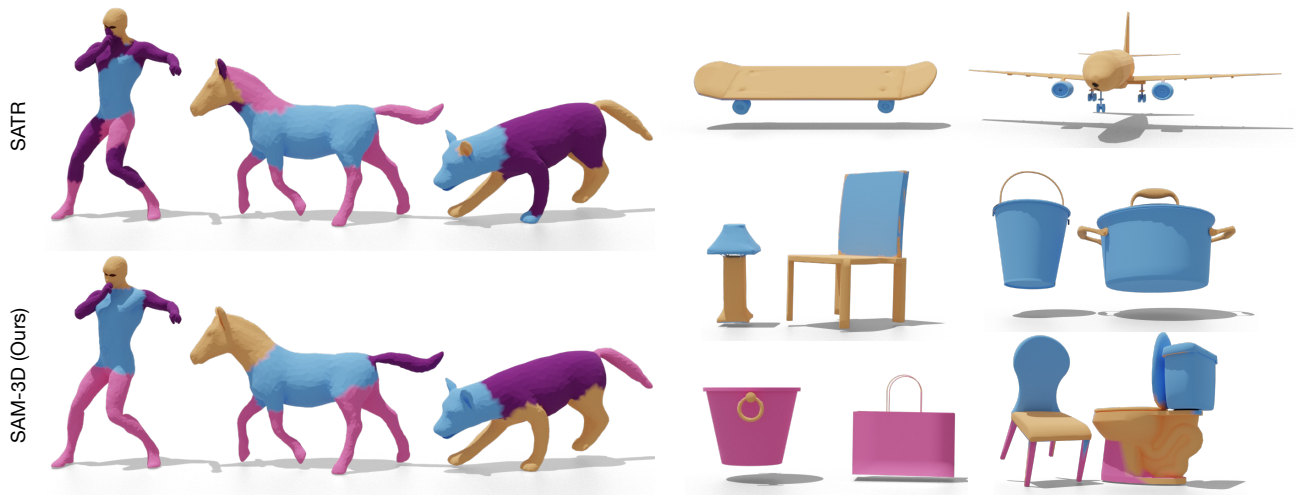

Figure 8: Qualitative comparison between our proposed SAM-3D in comparison with SATR [Abdelreheem et al. 2023]. SAM-3D provides more accurate and consistent segmentation compared to SATR.

Figure 9: Qualitative examples when matching unrelated daily objects. Our approach produces plausible correspondences demonstrating its reasoning and generalization capabilities.

Figure 10: Qualitative examples when matching unrelated shapes. Our approach produces plausible correspondences demonstrating its reasoning and generalization capabilities.

Figure 11: Detailed qualitative results for strongly non-isometric pairs of shapes. We show the intermediate predictions from different components of our proposed approach, including the predicted class labels, proposed semantic regions and mapping, and the coarse shape-matching output.