LEPARD: Learning Explicit Part Discovery for 3D Articulated Shape Reconstruction

Di Liu **Qilong Zhangli** **Yunhe Gao** **Dimitris N. Metaxas**
Rutgers University

Abstract

Reconstructing the 3D articulated shape of an animal from a single in-the-wild image is a challenging task. We propose LEPARD, a learning-based framework that discovers semantically meaningful 3D parts and reconstructs 3D shapes in a part-based manner. This is advantageous as 3D parts are robust to pose variations due to articulations and their shape is typically simpler than the overall shape of the object. In our framework, the parts are explicitly represented as parameterized primitive surfaces with global and local deformations in 3D that deform to match the image evidence. We propose a kinematics-inspired optimization to guide each transformation of the primitive deformation given 2D evidence. Similar to recent approaches, LEPARD is only trained using off-the-shelf deep features from DINO and does not require any form of 2D or 3D annotations. Experiments on 3D animal shape reconstruction, demonstrate significant improvement over existing alternatives in terms of both the overall reconstruction performance as well as the ability to discover semantically meaningful and consistent parts.

1 Introduction

Predicting the 3D shape and part articulation of an object from a single image is a severely under-constrained and challenging problem. It can be applied to many downstream tasks, such as shape reconstruction [45, 61, 47, 50, 54, 39], segmentation [24, 51, 32, 72, 16, 33, 6, 71], editing [65, 21], re-targeting [12, 20, 69] and medical imaging applications [22, 34, 36, 17, 37, 35, 14, 23, 18, 44]. Successful approaches [26, 29] for predicting the 3D shape of humans rely on a parametric human body model (*e.g.*, SMPL [41]) built from thousands of mocap sequences and on strong supervision from 3D joint locations. Similar breakthroughs are not seen for other articulated object categories, like animals, as 3D scanning of such categories is quite challenging. The lack of 3D annotations and an appropriate parametric animal model has led to approaches that utilize a pre-defined shape template and train with 2D supervision [27, 31, 30, 19, 28]. Assuming such a fixed shape template and supervision from 2D annotations is not optimal, but is necessary for training these systems. Recent works [67, 68, 60] have discarded both assumptions by learning a shape prior (part-based [67, 68] or holistic shape [60]) and by using deep features from an off-the-shelf vision transformer [13], DINO-ViT [5], for supervision. Although the approaches presented in [67, 68, 60] seem promising, they exhibit certain limitations. In [60] a category-specific prior mesh is obtained in a pre-training step, which can be seen as a learned shape template. Then it is trained to predict the articulation and deformation based on this fixed template. LASSIE [67] and Hi-LASSIE [68] learn generic part priors that, similar to [60], can be articulated and deformed. However, learning the deformation field *w.r.t* weak generic shape priors is a hard task. Assuming that the prior shape is far from the target shape of an object, the model has to compensate by predicting large deformations to match the image evidence.

In this paper, we propose a comprehensive framework for reconstructing the 3D shape and related articulations of an object from single-view images that has several desired properties missing from

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

https://doi.org/10.52202/075280-2359

existing works. Similar to [67, 68], the 3D shape of an object is explicitly expressed as a set of part primitives. The primitives are parameterized surfaces (*e.g.*, superquadrics) equipped with additional linear tapering and bending transformations to capture the target shape as faithfully as possible. Unlike [67, 68], the 3D primitives are not fixed across all instances and can deform to capture intra-category variations and accurately reconstruct 3D parts. To capture fine-grained shape details beyond the coverage of the primitive parametric deformations (termed global deformations), we employ a diffeomorphic mapping to estimate local non-rigid deformations of a set of points sampled from the 3D surface of each part. Our approach uses global deformations to capture the salient part of the 3D shape and uses local deformations to further improve the 3D shape reconstruction quality. As such, the local deformations are typically small, which adds to the robustness of our method. Following prior work [66, 68, 60] we use deep features from a vision transformer [13], DINO-ViT [5], as supervision to train our model. Inspired by the kinematics of 3D deformable models [57], we propose a framework to compute image-based forces based on the discrepancy of DINO features and the projected primitive parts. The image-based forces are then converted to generalized forces via kinematics modeling, which provides strong supervision for the deformation of the 3D part primitives.

We conduct extensive experiments on 3D articulated object shape reconstruction through part discovery. The Pascal-part [7] and LASSIE datasets [67] are used for training and evaluation following prior work [67, 68]. Both quantitative and qualitative evaluations demonstrate improved 3D reconstruction performance of our proposed approach compared to existing methods. Our approach even outperforms methods that rely on 3D skeletons or shape templates. In summary, our contributions are as follows:

• A new framework for reconstructing the 3D articulated shape of an object as a set of deformable 3D primitive parts given only 2D evidence. For each primitive global deformations are used to reconstruct the corresponding 3D part, while local deformations increase the fidelity of the reconstruction.

• A kinematics-inspired optimization process with perspective projection that allows converting the 3D primitive points to the generalized parameters corresponding to the transformations of each primitive.

• Extensive quantitative and qualitative evaluations showcase the superiority of our approach over the existing state-of-the-art methods.

2 Related Work

3D reconstruction of animals. Recently there have been several approaches that learn to reconstruct the 3D shape of animals from image [27, 31, 30, 19, 73, 4, 55, 60] or video inputs [28, 62–64]. Most previous methods make certain assumptions, such as pre-defined shape templates, statistical animal models, or the existence of annotated datasets. For example, some works [73, 4, 55] regress the parameters of a statistical shape model, SMAL [74]. However, this approach is only applicable to categories captured by the SMAL shape space. Some other methods rely on supervision from object silhouettes [27, 31, 30, 19], 2D keypoints [27] and the existence of a template shape [31, 30, 19]. This limits the applicability of those works to animal categories that have such annotations. In contrast, our proposed approach does not require any 3D shape template or skeleton and is trained with a self-supervised objective, which makes it easy to generalize to a wide range of animal species with no extra manual effort.

Optimization from multiple views and videos. Recently, Neural Radiance Fields (NeRF) [48, 3] have gained significant attention as a robust volumetric representation for multi-view reconstruction, particularly when accurate cameras are available. In a related line of research, recent works [62–64] have focused on optimizing the 3D shapes of articulated objects using a small number of monocular videos. These approaches employ meticulously designed optimization strategies that incorporate supervision from optical flow, object silhouettes, and DensePose [49] annotations. Another line of work [67, 68], leverages supervision from DINO-ViT [5] features and optimizes a part-based model on a small collection of images (~ 30) of a particular animal category. LASSIE [67] and Hi-LASSIE [68] additionally include a test-time optimization process, in which per-instance articulation and part-refinement are employed. Our approach is closely related to LASSIE and Hi-LASSIE since we also learn to reconstruct 3D articulated shapes in a part-based manner. However, unlike those methods, our approach does not require any test-time processing.

https://doi.org/10.52202/075280-2359

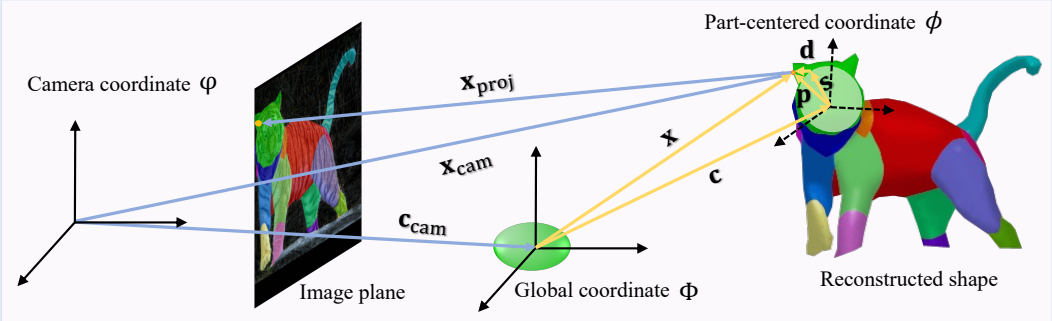

Figure 1: **LEPARD geometry.** We reconstruct an articulated animal shape by deforming a set of primitive parts with global deformation **s** and local deformations **d** that are predicted in a part-centered coordinate system ϕ. We also predict 3D translation and rotation transformations to place the reconstructed shape in a global coordinate system Φ. Global transformations are parameterized explicitly and offer an intuitive understanding of the shape (*e.g.*, bending parameters make the back and tail of the tiger bend).

Part discovery. Deep feature factorization (DFF) [9] and follow-up works [1, 8, 25, 56] show that one can automatically obtain 2D corresponding part segments by clustering deep semantic features. In the 3D domain, the object parts can be discovered by using explicit representation primitives [58, 42, 43, 52, 53, 11] (*e.g.*, cuboids, spheres, superquadrics), or learning part prior [66]. These methods mainly assume some form of supervision like 3D point clouds, keypoints or camera viewpoints. Similar to LASSIE and Hi-LASSIE, we discover parts based on deep features from DINO-ViT. However, we do not rely on a pre-defined 3D skeleton like LASSIE or an intermediate skeleton representation as Hi-LASSIE.

3 Approach

3.1 Primitive Part Representation

Geometry. Given an image of an articulated object category, LEPARD aims to learn a set of K primitive parts that compose its 3D shape. Each primitive is explicitly represented by a group of parameters that describe its 3D shape and orientation. Following [57] each individual primitive k is defined as a closed surface in a part-centered coordinate system $\phi^{(k)}$. Given a point $\mathbf{p}^{(k)}$ on the surface of primitive k, its 3D location $\mathbf{x} = (x, y, z)$ in the global coordinate system Φ can be computed as follows:

$$\mathbf{x} = \mathbf{c}^{(k)} + \mathbf{R}^{(k)}\mathbf{p}^{(k)} = \mathbf{c}^{(k)} + \mathbf{R}^{(k)}(\mathbf{s}^{(k)} + \mathbf{d}^{(k)}), \tag{1}$$

where $\delta^{(k)} \equiv (\mathbf{c}^{(k)}, \mathbf{R}^{(k)})$ represents the transformation of the part-centered coordinate system $\phi^{(k)}$ of primitive k to the global coordinate system Φ, $\mathbf{c}^{(k)} \in \mathbb{R}^3$ and $\mathbf{R}^{(k)} \in \mathbb{R}^{3\times3}$ represent the translation and rotation of $\phi^{(k)}$ *w.r.t.* Φ and $\mathbf{p}^{(k)}$ denotes the relative position of the point on the primitive surface *w.r.t.* $\phi^{(k)}$, which includes global deformation $\mathbf{s}^{(k)}$ and local deformation $\mathbf{d}^{(k)}$. The camera parameters $\pi \equiv (\mathbf{c}_{\text{cam}}, \mathbf{R}_{\text{cam}})$ are used to project **x** onto the image. In Fig. 1, we illustrate the geometry of the proposed part-based shape representation that includes global and local deformations for each part. For the sake of simplicity, we omit the superscript k for the k-th primitive in the following.

Primitive deformations. We employ superquadrics to describe the global deformations of each part primitive. Each superquadric surface **e** is explicitly defined by a set of shape-related parameters:

$$\mathbf{e} = a_0 \begin{bmatrix} a_1 \cos^{\varepsilon_1} u \cos^{\varepsilon_2} v \\ a_2 \cos^{\varepsilon_1} u \sin^{\varepsilon_2} v \\ a_3 \sin^{\varepsilon_1} u \end{bmatrix}, \text{where } -\pi/2 \leq u \leq \pi/2, -\pi \leq v \leq \pi. \tag{2}$$

Here, a_0 is a scaling parameter, a_1, a_2, a_3 denote the aspect ratio for x-, y-, z- axes, respectively, and $\varepsilon_1, \varepsilon_2$ are squareness parameters. To enable more flexible global deformations, we further include tapering and bending parameters. These additional global deformations are defined as continuously

https://doi.org/10.52202/075280-2359

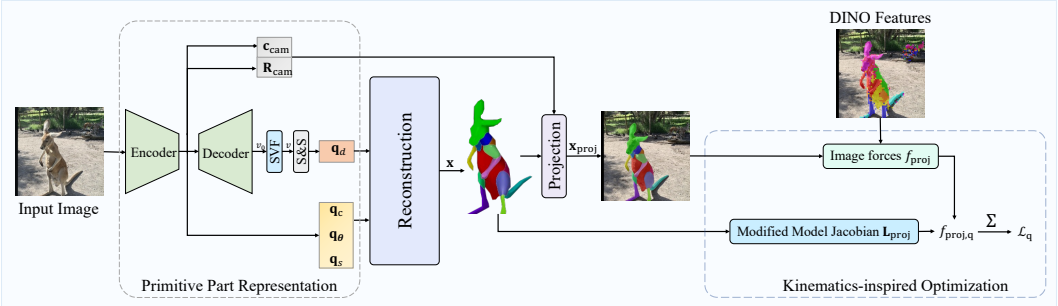

Figure 2: **LEPARD training overview.** Given an input image, we map it to a set of primitive parameters that describe K deformed parts via an encoder-decoder network. The primitive parameters, $\mathbf{q}_c$, $\mathbf{q}_\theta$, $\mathbf{q}_s$, $\mathbf{q}_d$, are then used to reconstruct the 3D articulated shape $\mathbf{x}$. During training, we project a set of points from the primitives' surface onto the image using the predicted camera parameters $\mathbf{c}_{\text{cam}}, \mathbf{R}_{\text{cam}}$. We use DINO features as supervision to compute image-based forces f_{proj} that we further convert into generalized forces $f_{\text{proj},q}$ to supervise each transformation of the primitive part.

differentiable and commutative functions following [46, 39, 38]. To capture the finer shape details beyond the coverage of global deformations, we employ diffeomorphic point flow to estimate the local non-rigid deformations $\mathbf{d}$. Since the deformation with diffeomorphism is differentiable and invertible [2, 10], it guarantees one-to-one mapping and preserves topology during the non-rigid deformations of the primitives. Please refer to [39, 38] for more details. We note that our approach is not restricted to these types of shapes and respective parameterizations. We can replace it with any differentiable type of primitive. However, for the purposes of the articulated animal shapes this type of parameterization is sufficient.

In summary, each part primitive is represented with a set of parameters $\mathbf{q} = [\mathbf{q}_c, \mathbf{q}_\theta, \mathbf{q}_s, \mathbf{q}_d]$, where $\mathbf{q}_c$ and $\mathbf{q}_\theta$ contain the parameters of the 3D translation and rotation respectively, that transform the part-centered coordinate system ϕ of the primitive to the global coordinate system Φ, $\mathbf{q}_s$ are the parameters of the global deformations, $\mathbf{q}_d$ are local deformations that are implemented as a deformation field and are added to the global shape, and $[\cdot]$ is the concatenation operator. Compared to the implicit function-based approaches such as NeRF [48] and NeRS [70], all these primitive parameters are defined explicitly for an intuitive understanding of the primitive deformation.

3.2 Primitive 3D Kinematics & Optimization

Primitive kinematics in 3D. We use kinematics to define the relationship between any point $\mathbf{x}$ on the primitive surface and the corresponding primitive parameters $\mathbf{q}$. This relationship is expressed quantitatively by the model Jacobian matrix $\mathbf{L}$, and this formulation allows us to deform the primitive shape based on its parameters $\mathbf{q}$. Specifically, the velocity of $\mathbf{x}$ is computed as follows:

$$\dot{\mathbf{x}} = \mathbf{L}\dot{\mathbf{q}}, \tag{3}$$

where $\cdot$ denotes the first-order time derivative and $\mathbf{L} = [\mathbf{I}, \mathbf{B}, \mathbf{RJ}, \mathbf{R}]$ [57] is the Model Jacobian matrix, where each of the four components transforms the 3D points $\mathbf{x}$ into the translation, rotation, global and local deformation parameters of $\mathbf{q}$. $\mathbf{R}$ is the rotation matrix that corresponds to the rotation between the part-centered coordinate system ϕ of the primitive and the global coordinate system Φ. $\mathbf{B} = \partial \mathbf{Rp}/\partial \mathbf{q}_\theta$ is related to the rotation matrix $\mathbf{R}$ and the relative position $\mathbf{p}$ of points on the primitive surface. $\mathbf{J} = \partial \mathbf{s}/\partial \mathbf{q}_s$ is the Jacobian matrix. We refer the reader to the supplementary material for more details.

Optimization in 3D. In our modeling paradigm, we minimize the energy of the primitive, defined using the principle of virtual work. The 3D forces on the primitive, $f_{3\text{D}}$, result in displacements $d\mathbf{x}$:

$$\mathcal{E}_{f_{3\text{D}}} = \int f_{3\text{D}}^\top d\mathbf{x} = \int f_{3\text{D}}^\top (\mathbf{L} d\mathbf{q}) = \int f_q d\mathbf{q}. \tag{4}$$

where $f_q = f_{3\text{D}}^\top \mathbf{L}$ [57] are the generalized forces acting on the primitive parameters $\mathbf{q}$. $f_{3\text{D}}$ are computed based on the discrepancy between points on the primitive and the target shape in 3D data space (*e.g.*, point-wise difference between the primitive surface and the target shape). The forces on

https://doi.org/10.52202/075280-2359

the primitive are proportional to the distance from the target. Minimizing the generalized forces f_q deforms the primitive to match the target shape and thus can be used as a training objective in our framework.

3.3 Primitive Part Discovery from Images

Our model is trained using a set of N in-the-wild images of articulated animals. We do not make use of any type of 2D/3D annotations or skeletons. We compute pseudo-labels using semantic clustering of self-supervised DINO [5] features similar to prior work [67, 68]. For each of the given images, we predict the parameters for the K primitive parts as defined previously. Our approach deforms a set of primitives to fit the target shape under the influence of external forces f_{3D}. However, direct primitive parameter optimization using f_{3D} is not feasible since we do not have access to any form of 3D supervision. To this end, we project the predicted 3D primitive parts onto the image space and define losses based on the corresponding 2D forces to supervise the deformation of each primitive. The training overview is given in Fig. 2.

Projection kinematics from 3D to 2D. We illustrate the relationship of kinematics between 3D and 2D via projective geometry. This allows us to project the primitives onto the image space and calculate the discrepancy between the projected primitives and 2D evidence. Then, we convert the image forces to their corresponding generalized forces that guide the deformation of the primitives. Specifically, using the estimated parameters for camera translation $\mathbf{c}_{\text{cam}}$ and rotation $\mathbf{R}_{\text{cam}}$, we can convert a given 3D point $\mathbf{x}$ to the camera coordinate system as follows:

$$\mathbf{x}_{\text{cam}} = \mathbf{c}_{\text{cam}} + \mathbf{R}_{\text{cam}}\mathbf{x}. \tag{5}$$

Under perspective projection, the point $\mathbf{x}_{\text{cam}} = (x_c, y_c, z_c)$ projects onto an image point $\mathbf{x}_{\text{proj}} = (x_{\text{proj}}, y_{\text{proj}})$ according to $x_{\text{proj}} = f\frac{x_c}{z_c}$, $y_{\text{proj}} = f\frac{y_c}{z_c}$, where f is the focal length. By taking the time derivative we have $d\mathbf{x}_{\text{proj}} = \mathbf{P}d\mathbf{x}_{\text{cam}}$, where

$$\mathbf{P} = \begin{bmatrix} f/z_c & 0 & -fx_c/z_c^2 \\ 0 & f/z_c & -fy_c/z_c^2 \end{bmatrix}. \tag{6}$$

Thus, from Eq. (5) we have:

$$d\mathbf{x}_{\text{proj}} = \mathbf{P}d\mathbf{x}_{\text{cam}} = \mathbf{P}d(\mathbf{c}_{\text{cam}} + \mathbf{R}_{\text{cam}}\mathbf{x}) = \mathbf{P}\mathbf{R}_{\text{cam}}d\mathbf{x}. \tag{7}$$

Given Eq. (3) we can rewrite Eq. (7) as:

$$d\mathbf{x}_{\text{proj}} = \mathbf{P}\mathbf{R}_{\text{cam}}d\mathbf{x} = \mathbf{P}\mathbf{R}_{\text{cam}}(\mathbf{L}d\mathbf{q}) = (\mathbf{P}\mathbf{R}_{\text{cam}}\mathbf{L})d\mathbf{q} = \mathbf{L}_{\text{proj}}d\mathbf{q}, \tag{8}$$

where $\mathbf{L}_{\text{proj}}$ is the modified model Jacobian matrix that converts image points $\mathbf{x}_{\text{proj}}$ into primitive parameters $\mathbf{q}$ that determine translation, rotation, global and local deformations of the primitive.

Loss components. Since we only use 2D supervision during training, the primitive energy from each image we minimize is computed based on the image forces f_{proj}:

$$\mathcal{E}_{f_{\text{proj}}} = \int f_{\text{proj}}^{\top} d\mathbf{x}_{\text{proj}} = \int f_{\text{proj}}^{\top}(\mathbf{L}_{\text{proj}}d\mathbf{q}) = \int f_{\text{proj},q}d\mathbf{q}, \tag{9}$$

where the generalized forces $f_{\text{proj},q} = f_{\text{proj}}^{\top}\mathbf{L}_{\text{proj}}$ guide the primitive deformation. For a given image i from the training dataset, we minimize the discrepancy between our primitives and the target using the image-based forces as follows:

$$\mathcal{L}_q^i = \frac{1}{K}\sum_{k=1}^{K} f_{\text{proj},q}^{(k,i)} = \frac{1}{K}\sum_{k=1}^{K} (f_{\text{proj}}^{(k,i)})^{\top}\mathbf{L}_{\text{proj}}^{(k,i)}, \tag{10}$$

where $f_{\text{proj}}^{(k,i)}$ denotes the corresponding image forces of the k-th primitive part and is a vector summation of all the forces at each sampling point on the projected primitive. The final loss used for training is computed by summing the generalized forces for all N training samples as follows:

$$\mathcal{L}_q = f_{\text{proj},q_c}^{\top} + f_{\text{proj},q_\theta}^{\top} + f_{\text{proj},q_s}^{\top} + f_{\text{proj},q_d}^{\top}, \tag{11}$$

https://doi.org/10.52202/075280-2359

where

$$f^{\top}_{\text{proj},q_c} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(f^{(k,i)}_{\text{proj}})^{\top}\mathbf{P}^{(k,i)}\mathbf{R}^{(i)}_{\text{cam}}, \qquad f^{\top}_{\text{proj},q_\theta} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(f^{(k,i)}_{\text{proj}})^{\top}\mathbf{P}^{(k,i)}\mathbf{R}^{(i)}_{\text{cam}}\mathbf{B}^{(k,i)},$$

$$f^{\top}_{\text{proj},q_s} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(f^{(k,i)}_{\text{proj}})^{\top}\mathbf{P}^{(k,i)}\mathbf{R}^{(i)}_{\text{cam}}\mathbf{R}^{(k,i)}\mathbf{J}^{(k,i)}, \quad f^{\top}_{\text{proj},q_d} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}(f^{(k,i)}_{\text{proj}})^{\top}\mathbf{P}^{(k,i)}\mathbf{R}^{(i)}_{\text{cam}}\mathbf{R}^{(k,i)}.$$

To obtain $f^{(k,i)}_{\text{proj}}$, we render the predicted 3D primitive parts with a differentiable renderer [40] and compute the distance from the pseudo-mask part annotations:

$$f^{(k,i)}_{\text{proj}} = \lambda_{\text{proj}}(\mathcal{G}^{(i)} - \mathcal{M}^{(k,i)}_{\text{proj}}), \tag{12}$$

where $\mathcal{G}^{(i)}$ and $\mathcal{M}^{(k,i)}_{\text{proj}}$ are the pseudo-mask and the k-th projected primitive for instance i, respectively. λ_{proj} is a constant modeling the strength of the image force $f^{(k,i)}_{\text{proj}}$.

Force regularization. To enable more robust primitive fitting strategy, we incorporate regularization to the forces during training. We follow the physics-based deformable models (DMs) [46] and avoid the collisions between primitives by checking for primitive inter-penetration in each training iteration. If two primitives penetrate each other, we allocate two equivalent and opposite collision forces f_n and $-f_n$ that are proportional to the distance between each pair of selected points on the two primitives. These two forces are added to the respective points on the two inter-penetrating primitives, respectively, to adjust the external forces f_{proj} and thus push the primitives to separate from each other.

4 Experiments

Datasets and baselines. We emphasize that our aim is to reconstruct an articulated shape by discovering semantically meaningful parts in 3D. Thus, we mainly compare our approach with LASSIE [67] and Hi-LASSIE [68] who share the same goal. We conduct extensive experiments on a few animal categories, including horses, giraffes, zebras, etc. For horses, we train our model using 11k images collected from Pascal [15], LASSIE [67] and DOVE [59] datasets. For the other categories, we only use the images from Pascal [15] and LASSIE [67] datasets to fine-tune our pre-trained model. To put our results into perspective, we also quantitatively compare with A-CSM [30] and 3D Safari [73], which learn to reconstruct the 3D articulated shape of animals holistically. Finally, we also provide qualitative comparisons with LASSIE using their released code. We only compare quantitatively with Hi-LASSIE, using the reported results in [68], since their code is not available at the time of submission.

Qualitative results. In Fig. 3 we provide qualitative comparisons with LASSIE. We observe that LEPARD yields more geometrically accurate reconstructions with finer details compared to LASSIE. This is particularly evident in the cases of tiger and elephant, where the primitives used in LASSIE do not accurately correspond to the animal parts with complex shapes such as the tiger tail and the elephant nose. Additionally, LASSIE fails in the presence of additional objects in the background (see the predicted shape for the kangaroo in row 4). For the case of the kangaroo, our model accurately locates the head and correctly recovers the pose under the interference of the inaccurate clustering appearing on the top right of the DINO features. The superiority of our model in estimating accurate poses can also be observed in the case of the penguin (row 5 of Fig. 3).

Consistency visualization. In the first two rows of Fig. 4 we show that LEPARD predicts consistent parts across different samples of the same animal category, where LASSIE struggles (*e.g.*, in the 5-th column, LASSIE predicts the elephant's nose as a leg). In addition, we visualize the semantic consistency of our model across different animal categories. In the last two rows of Fig. 4, we compare the reconstruction results of kangaroos and penguins. We observe that our approach employs the same primitive to consistently represent corresponding parts of different species, *e.g.*, the heads of both penguins and kangaroos are represented with the same primitive part (highlighted in green). But LASSIE employs this same green part to represent the mouth of penguins and the head of kangaroos, respectively.

https://doi.org/10.52202/075280-2359

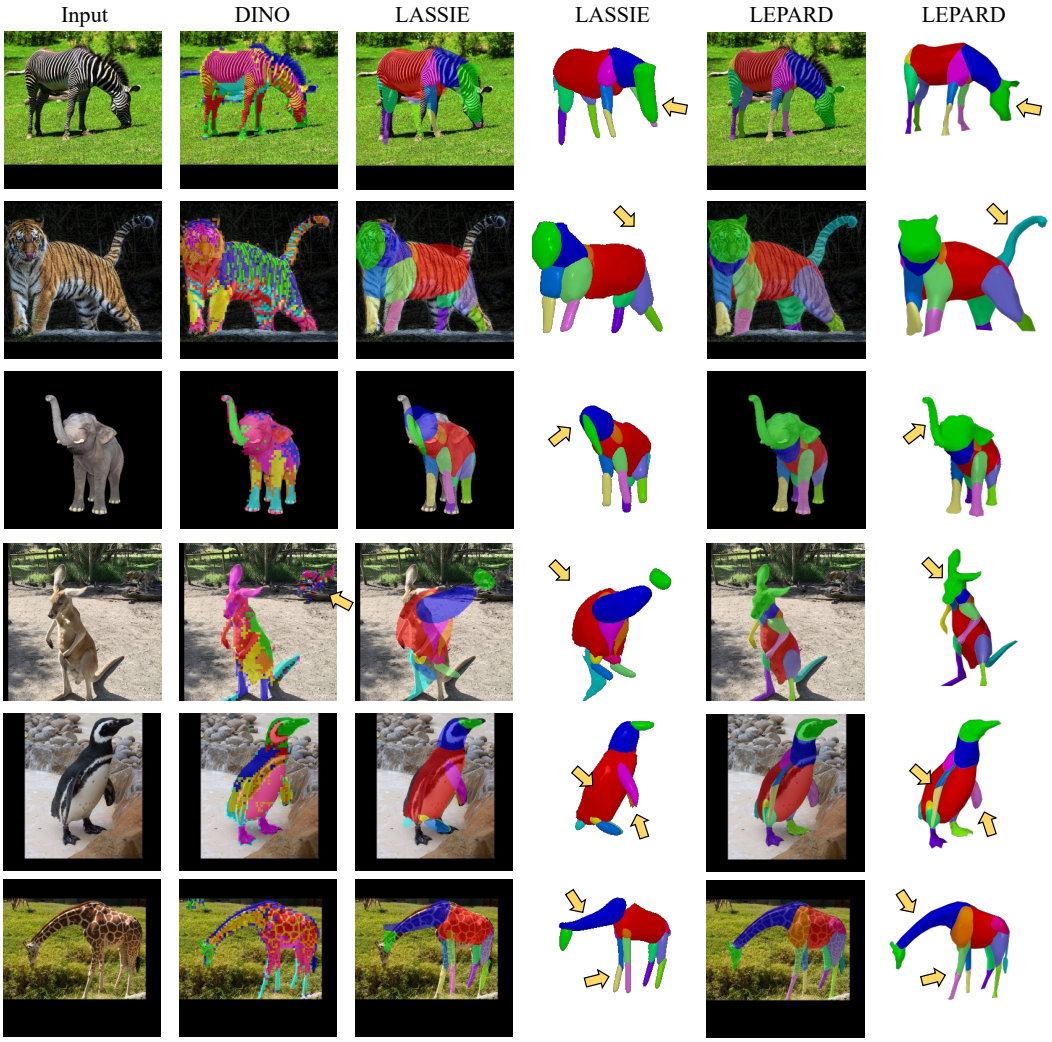

Figure 3: **Qualitative results.** We compare the recovered parts by LASSIE [67] and LEPARD (ours). We observe that LEPARD discovers semantically meaningful and consistent parts, offering more faithful reconstructions than LASSIE. Unlike LASSIE, our approach is robust to the presence of additional objects in the background (see the kangaroo in row 4). Failures in LASSIE's prediction are indicated by arrows, whereas LEPARD avoids such failures.

Keypoint transfer. Due to the lack of ground-truth 3D annotations in our datasets, we follow a common practice [30, 67, 68] and quantitatively evaluate our model using 2D keypoint transfer between each pair of images in the test set. In particular, given a set of keypoints on a source image, we map them onto the 3D primitive parts and then project them to the target image. We then compute the percentage of correct keypoints (PCK) under a tight threshold $0.05 \times \max(h, w)$ (*i.e.*, PCK@0.05), where h and w are the image height and width respectively. For a successful 2D→3D→2D mapping, accurate 3D reconstructions for both the source and target images are necessary. We report results for the keypoint transfer evaluation in Table 1. The results show that LEPARD achieves higher PCK compared to the baselines without performing test-time per-instance optimization.

Part transfer. Next, we evaluate our approach on part transfer using the ground-truth part segmentation masks from the Pascal-Part dataset. Similar to 2D keypoint transfer, we transfer part annotations from source to target images through a 2D→3D→2D mapping utilizing the predicted 3D part primitives for the sources and target images. In this setting, we measure the performance with the percentage of correct pixels (PCP) metric, where a pixel from the source image is considered to be transferred correctly if it is mapped to the same semantic part in the target image. The results

https://doi.org/10.52202/075280-2359

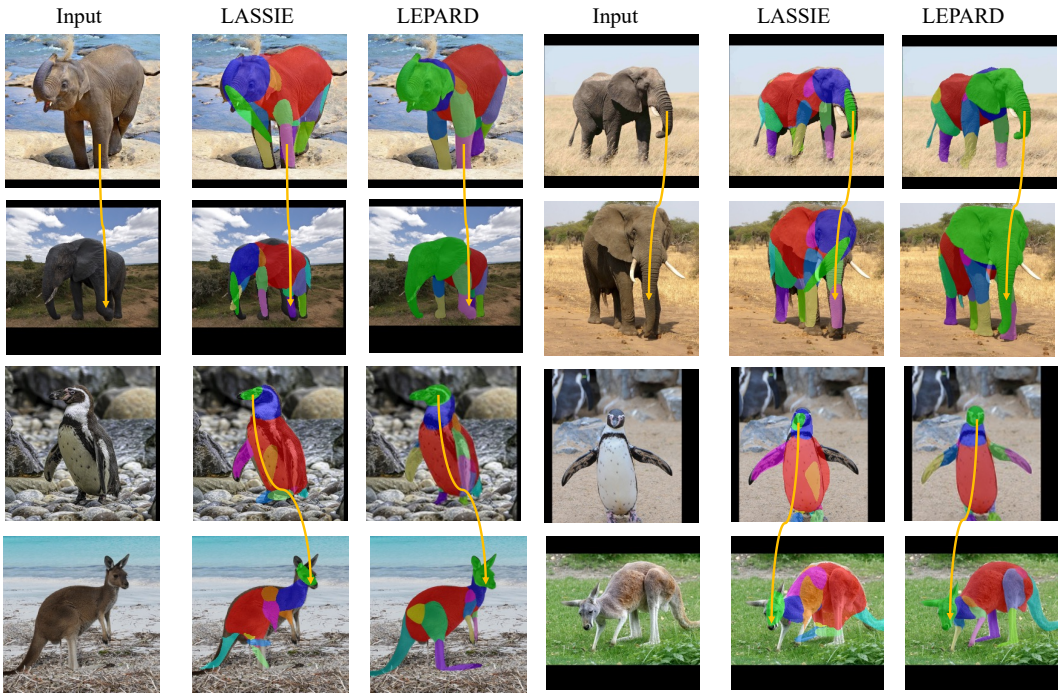

Figure 4: **Consistency visualization.** We compare the discovered parts by LEPARD and LASSIE [67]. The arrows indicate the correct correspondence of the parts among instances (*e.g.*, the left front legs of the elephants should be discovered by primitive parts in the same color) from the same (first two rows) and different (last two rows) animal categories. We observe that LEPARD discovers 3D parts with better semantic consistency (*i.e.*, reconstructs the same parts using the same primitives).

Table 1: **Keypoint transfer evaluation.** Results on the Pascal-Part and LASSIE datasets using PCK@0.05. LEPARD outperforms all methods for all animal categories by a significant margin.

	Pascal-Part dataset			LASSIE dataset					
	Horse	Cow	Sheep	Zebra	Tiger	Giraffe	Elephant	Kangaroo	Penguin
3D Safari [73]	57.1	50.3	50.5	62.1	50.3	32.5	29.9	20.7	28.9
A-CSM [30]	55.3	60.5	54.7	60.3	55.7	52.2	39.5	26.9	33.0
LASSIE [67]	58.0	62.4	55.5	63.3	62.4	60.5	40.3	31.5	40.6
Hi-LASSIE [68]	59.6	63.1	56.2	64.2	63.1	61.6	42.7	35.0	44.4
LEPARD	**61.0**	**63.7**	**56.9**	**64.7**	**63.8**	**62.1**	**43.2**	**35.4**	**44.6**

are reported on the right column of Table 2, demonstrating that LEPARD compares favorably to all baseline methods.

2D IoU. In addition, we evaluate LEPARD using overall and part IoU between the ground-truth and rendered masks. The results are reported in Table 2. We follow the baselines [68, 67, 25, 5] and manually assign the discovered parts to the best-matched parts in the Pascal-Part segmentation masks. From Table 2 we observe that LEPARD outperforms all other methods in terms of overall IoU as well as Part IoU by a large margin.

Effect of local deformations. We investigate the effect of local deformations in terms of reconstruction accuracy and present qualitative results in Fig. 5. We observe that local deformations can capture fine-grained shape details and significantly improve the visual quality of the reconstructed shapes.

Model robustness and limitation. In Fig. 6 we evaluate on some held-out images from the Objaverse dataset for known categories such as elephants to show the robustness of our method. For the reconstruction of unknown categories, our method may require fine tuning on additional images, and will be included in our future work.

https://doi.org/10.52202/075280-2359

Table 2: **Quantitative results on the Pascal-Part [7] dataset.** We report the overall IoU, part mask IoU, as well as part transfer results measured by the percentage of correct pixels (PCP).

	Overall IoU			Part IoU			Part Transfer (PCP)		
	Horse	Cow	Sheep	Horse	Cow	Sheep	Horse	Cow	Sheep
SCOPS [25]	62.9	67.7	63.2	23.0	19.1	26.8	-	-	-
DINO clustering [1]	81.3	85.1	83.9	26.3	21.8	30.8	-	-	-
3D Safari [73]	72.2	71.3	70.8	-	-	-	71.7	69.0	69.3
A-CSM [30]	72.5	73.4	71.9	-	-	-	73.8	71.1	72.5
LASSIE [67]	81.9	87.1	85.5	38.2	35.1	43.7	78.5	77.0	74.3
Hi-LASSIE [68]	83.4	88.1	86.3	39.0	35.3	43.4	79.9	**77.8**	75.5
LEPARD	**83.7**	**88.3**	**86.7**	**39.5**	**35.4**	**43.6**	**80.6**	77.6	**75.8**

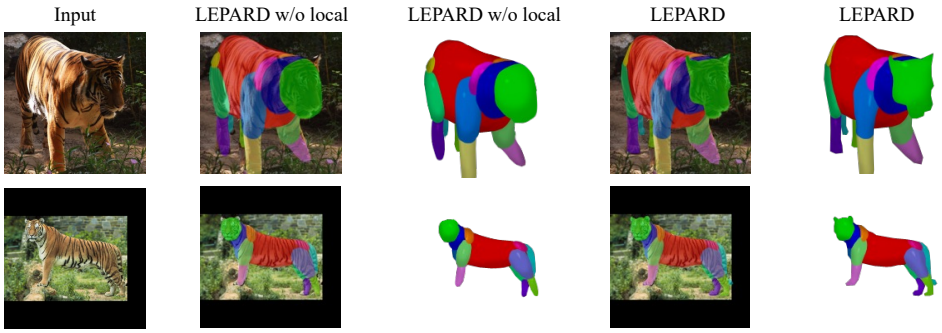

Figure 5: **Effect of local deformations.** We test the effectiveness of the local deformations on the Pascal-Part and the LASSIE dataset. We observe that local deformations have a significant impact on the visual quality of the reconstructed shapes.

Figure 6: Testing on held-out images from known category. We show some qualitative results on the elephant category of Objaverse dataset.

5 Conclusion

We present LEPARD, a learning-based framework for discovering the 3D parts and shape of an articulated animal from an image without any 2D/3D annotations or skeletons. Our approach jointly optimizes a set of primitives which are explicitly defined by a few shape-related parameters, and provides an intuitive understanding of the primitive deformation. To obtain high-fidelity reconstruction that matches the 2D evidence provided by clustering DINO features, we propose a kinematics mapping between 3D and 2D, and convert the 2D forces defined *w.r.t.* images to the generalized forces that supervise the motions and deformations of the primitive parts in 3D. Our approach does not require any test-time processing and discovers semantically meaningful and consistent 3D parts. We demonstrate the effectiveness of our approach through quantitative and qualitative evaluations where we achieve significant improvements over existing methods.

Acknowledgments

This research has been partially funded by research grants to D. Metaxas through NSF: IUCRC CARTA 1747778, 2235405, 2212301, 1951890, 2003874, and NIH-5R01HL127661.

https://doi.org/10.52202/075280-2359

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 3, 9

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 924–931. Springer, 2006. 4

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[4] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 2, 5, 8

[6] Qi Chang, Zhennan Yan, Mu Zhou, Di Liu, Khalid Sawalha, Meng Ye, Qilong Zhangli, Mikael Kanski, Subhi Al'Aref, Leon Axel, et al. Deeprecon: Joint 2d cardiac segmentation and 3d volume reconstruction via a structure-specific generative method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2022. 1

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2, 9

[8] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *NeurIPS*, 2021. 3

[9] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *ECCV*, 2018. 3

[10] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019. 4

[11] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *CVPR*, 2020. 3

[12] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021. 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2

[14] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu Dietlinde Tizabi, Fabian Isensee, Tim J Adler, Patrick Godau, Veronika Cheplygina, Michal Kozubek, Sharib Ali, et al. Biomedical image analysis competitions: The state of current participation practice. *arXiv preprint arXiv:2212.08568*, 2022. 1

[15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[16] Yunhe Gao, Zhuowei Li, Di Liu, Mu Zhou, Shaoting Zhang, and Dimitris N Meta. Training like a medical resident: Universal medical image segmentation via context prior learning. *arXiv preprint arXiv:2306.02416*, 2023. 1

[17] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022. 1

[18] Chuanbin Ge, Di Liu, Juan Liu, Bingshuai Liu, and Yi Xin. Automated recognition of arrhythmia using deep neural networks for 12-lead electrocardiograms with fractional time–frequency domain extension. *Journal of Medical Imaging and Health Informatics*, 10(11):2764–2767, 2020. 1

[19] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 1, 2

[20] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019. 1

[21] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Yuxiao Chen, Di Liu, Qilong Zhangli, et al. Improving negative-prompt inversion via proximal guidance. *arXiv preprint arXiv:2306.05414*, 2023. 1

[22] Ali Hatamizadeh, Debleena Sengupta, and Demetri Terzopoulos. End-to-end trainable deep active contour models for automated image segmentation: Delineating buildings in aerial imagery. In *ECCV*, 2020. 1

[23] Xiaoxiao He, Chaowei Tan, Bo Liu, Liping Si, Weiwu Yao, Liang Zhao, Di Liu, Qilong Zhangli, Qi Chang, Kang Li, et al. Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation. *arXiv preprint arXiv:2303.14357*, 2023. 1

[24] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *CVPR*, 2017. 1

[25] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, 2019. 3, 8, 9

[26] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1

[27] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1, 2

[28] Filippos Kokkinos and Iasonas Kokkinos. Learning monocular 3d reconstruction of articulated categories from motion. In *CVPR*, 2021. 1, 2

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1

[30] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, 2020. 1, 2, 6, 7, 8, 9

[31] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, 2019. 1, 2

[32] Zhuowei Li, Long Zhao, Zizhao Zhang, Han Zhang, Di Liu, Ting Liu, and Dimitris N Metaxas. Steering prototype with prompt-tuning for rehearsal-free continual learning. *arXiv preprint arXiv:2303.09447*, 2023. 1

[33] Di Liu, Yunhe Gao, Qilong Zhangli, Zhennan Yan, Mu Zhou, and Dimitris Metaxas. Transfusion: Multi-view divergent fusion for medical image segmentation with transformers. *arXiv preprint arXiv:2203.10726*, 2022. 1

[34] Di Liu, Chuanbin Ge, Yi Xin, Qin Li, and Ran Tao. Dispersion correction for optical coherence tomography by the stepped detection algorithm in the fractional fourier domain. *Optics express*, 28(5):5919–5935, 2020. 1

[35] Di Liu, Jiang Liu, Yihao Liu, Ran Tao, Jerry L Prince, and Aaron Carass. Label super resolution for 3d magnetic resonance images using deformable u-net. In *Medical Imaging 2021: Image Processing*, volume 11596, page 1159628. International Society for Optics and Photonics, 2021. 1

[36] Di Liu, Yi Xin, Qin Li, and Ran Tao. Dispersion correction for optical coherence tomography by parameter estimation in fractional fourier domain. In *2019 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 674–678. IEEE, 2019. 1

[37] Di Liu, Zhennan Yan, Qi Chang, Leon Axel, and Dimitris N Metaxas. Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac mr segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 315–322. Springer, 2021. 1

[38] Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, and Dimitris N Metaxas. Deformer: Integrating transformers with deformable models for 3d shape abstraction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14236–14246, 2023. 4

[39] Di Liu, Long Zhao, Qilong Zhangli, Yunhe Gao, Ting Liu, and Dimitris N Metaxas. Deep deformable models: Learning 3d shape abstractions with part consistency. *arXiv preprint arXiv:2309.01035*, 2023. 1, 4

[40] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 6

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 1

[42] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. In *ICLR*, 2020. 3

[43] Priyanka Mandikal, Navaneet KL, and R Venkatesh Babu. 3d-psrnet: Part segmented 3d point cloud reconstruction from a single image. In *ECCV*, 2018. 3

[44] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: The m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. 1

[45] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[46] Dimitris N Metaxas. *Physics-based deformable models: applications to computer vision, graphics and medical imaging*, volume 389. Springer Science & Business Media, 2012. 4, 6

[47] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *ICCV*, 2019. 1

[48] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 4

[49] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *NeurIPS*, 2020. 2

[50] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019. 1

[51] Chengjie Niu, Jun Li, and Kai Xu. Im2struct: Recovering 3d shape structure from a single rgb image. In *CVPR*, 2018. 1

[52] Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *CVPR*, 2020. 3

[53] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *CVPR*, 2021. 3

https://doi.org/10.52202/075280-2359

[54] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *CVPR*, 2019. 1
[55] Nadine Rüegg, Silvia Zuffi, Konrad Schindler, and Michael J Black. Barc: Learning to regress 3d dog shape from images by exploiting breed information. In *CVPR*, 2022. 2
[56] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Motion-supervised co-part segmentation. In *ICPR*, 2021. 3
[57] Demetri Terzopoulos and Dimitri Metaxas. Dynamic 3d models with local and global deformations: deformable superquadrics. *IEEE TPAMI*, 13(7):703–714, 1991. 2, 3, 4
[58] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 3
[59] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, pages 1–12, 2023. 6
[60] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *CVPR*, 2023. 1, 2
[61] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *NeurIPS*, 2019. 1
[62] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 2
[63] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Ce Liu, and Deva Ramanan. Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. *NeurIPS*, 2021.
[64] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 2
[65] Kaizhi Yang, Xiaoshuai Zhang, Zhiao Huang, Xuejin Chen, Zexiang Xu, and Hao Su. Movingparts: Motion-based 3d part discovery in dynamic radiance field. *arXiv preprint arXiv:2303.05703*, 2023. 1
[66] Chun-Han Yao, Wei-Chih Hung, Varun Jampani, and Ming-Hsuan Yang. Discovering 3d parts from image collections. In *ICCV*, 2021. 2, 3
[67] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Lassie: Learning articulated shapes from sparse image ensemble via 3d part discovery. In *NeurIPS*, 2022. 1, 2, 5, 6, 7, 8, 9
[68] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *CVPR*, 2023. 1, 2, 5, 6, 7, 8, 9
[69] Meng Ye, Mikael Kanski, Dong Yang, Qi Chang, Zhennan Yan, Qiaoying Huang, Leon Axel, and Dimitris Metaxas. Deeptag: An unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7261–7271, 2021. 1
[70] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *NeurIPS*, 2021. 4
[71] Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, Zhaoyang Xia, Haiming Tang, He Wang, Mu Zhou, and Dimitris Metaxas. Region proposal rectification towards robust instance segmentation of biological images. *arXiv preprint arXiv:2203.02846*, 2022. 1
[72] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *ICCV*, 2017. 1
[73] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *ICCV*, 2019. 2, 6, 8, 9
[74] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, 2017. 2