Abstract

We present StrobeNet, a method for category-level 3D reconstruction of articulating objects from one or more unposed RGB images. Reconstructing general articulating object categories is challenging since objects can have wide variation in shape, articulation, appearance and topology. We address this by building on the idea of category-level articulation canonicalization – mapping observations to a canonical articulation which enables correspondence-free multiview aggregation. Our end-to-end trainable neural network estimates feature-enriched canonical 3D point clouds, articulation joints, and part segmentation from one or more unposed images of an object. These intermediate estimates are used to generate a final implicit 3D reconstruction. Our approach reconstructs objects even when they are observed in different articulations in images with large baselines, and animation of reconstructed shapes. Quantitative and qualitative evaluations on different object categories show that our method is able to achieve high reconstruction accuracy, especially as more views are added.

1. Introduction

Reconstructing the 3D shape of everyday objects is an important problem in applications such as robotics. To effectively operate in the real world, robots must not only understand the 3D shape of rigid objects, but also of articulating objects (e.g., laptops). There exists a large body of work on image-based 3D reconstruction of rigid objects [22, 24, 40] as well as on articulating human bodies [37, 59, 80]. However, reconstructing the 3D shape of previously unseen instances from common articulating object categories remains an open problem.

In this paper, our goal is to reconstruct an animatable 3D model of articulating objects observed in different poses from a sparse set of one or more unposed input images. Reconstruction from sparse unposed views is an ill-posed problem that is made even harder for articulated objects since they can have wide variation in shape and appearance, and could be in different poses in different input views. When video input with overlapping frames is available, approaches like non-rigid Structure from Motion (NR-SfM) [2, 11], or tracking [34, 82] can leverage correspondences to recover 3D geometry. However, correspondences are hard to find in images with large baselines, especially when objects articulate differently in each view. While recent methods show promising results in such cases by learning category-level shape priors [17, 24, 51, 77], they assume that objects are rigid.

StrobeNet (named after stroboscope images) is, to our knowledge, the first learning-based method that recon-
objects as a continuous, animatable implicit function model [13, 51, 56] from sparse input views (see Figure 1). Our approach is based on two key ideas. (1) We learn category-level shape priors for reconstructing not only rigid objects but also common articulating objects. To enable this, we use SAPIEN [76] and Shape2Motion [73], datasets that contain common articulating object categories, and generated a new image dataset with over 120k different views of different articulating objects. This large dataset could allow us to estimate shape by training a convolutional neural network (CNN) as a category-level shape prior. However, the space of already large object shape and appearance variation becomes even larger with articulations making learning difficult. (2) To overcome this, we use canonicalization, a tool for category-level 3D object understanding [46, 54, 64, 71] that maps object observations to a container space that is canonicalized for properties such as position, orientation and size [71]. Articulation canonicalization [46] allows correspondence-free multiview shape aggregation even when objects are observed in different articulations.

Our network takes one or more unposed input views and estimates an intermediate articulation-canonicalized point cloud enriched with learned features, location and orientation of joints, and part segmentation of the object. We use the aggregated 3D point cloud to drive a neural implicit function network [13], and output a smooth and continuous 3D shape. Our network implicitly models articulation resulting in improved reconstruction quality. Finally, our method can generate animatable 3D models at arbitrary articulations even without using any category-specific shape templates.

We evaluate our approach on a newly-created dataset of 3 common articulating object categories, with full ground truth labels. Our method compares favorably to rigid multiview aggregation methods on single view reconstruction, and significantly outperforms baselines in the articulated multiview aggregation setting. We additionally observe that joint and part segmentation estimation improves reconstruction quality. StrobeNet provides articulation manipulation functionality which can be used to animate articulated objects. To sum up our contributions:

- We introduce the problem of 3D implicit shape reconstruction of articulating objects from sparse unposed RGB views, and the first method to address this problem.
- A neural network architecture that explicitly models articulation to lift unposed RGB views of articulating objects to a 3D canonical point cloud and subsequently an implicit 3D shape.
- A method that does not require category-level templates, is fully correspondence-free, and can generate an animatable model of articulating objects.

2. Related Work

Our goal is to operate on common articulating object categories and to generalize to previously unseen instances. While we are the first to explore this particular setting, here we discuss closely related works which we split into reconstruction of rigid, non-rigid, and human shapes from sparse images. Since StrobeNet uses multiple types of 3D shape representations, we also discuss recent advances in this field in the context of learning.

2.1. Multi-view reconstruction

Rigid Objects: Reconstruction from images of rigid objects and scenes is a long-standing problem of interest with many classical approaches like voxel coloring [60], space carving [41], or multiview stereo [18, 22]. These approaches assume that objects of interest are well segmented, camera poses are known, camera baselines are small, and do not make use of any category-level priors. More recent learning-based methods overcome some of these limitations by learning category-specific priors [17, 20, 21, 25, 29, 38, 40, 47, 48, 51, 56, 57, 72, 75, 77] to reconstruct models from a single image. When 3D supervision is not available, some methods move away from category-level priors instead using foreground masks [27, 74]. When multiple views are available at inference time, many methods explore how to exploit multiview constraints [32, 36, 39, 62]. Other approaches for combining multiview information include using recurrent neural nets [14, 69, 70], or using feature-level aggregation [42, 61, 64], or differentiable rendering [12, 35]. Different from these methods, we focus on articulated objects observed from multiple views.

Non-Rigid Objects: Work on non-rigid object reconstruction can be divided into two categories: (1) non-rigid structure from motion (NR-SfM), and (2) non-rigid tracking approaches. NR-SfM methods typically focus on reconstructing sparse points [1, 3, 15], or use basis shapes, priors, or low-rank approximations to overcome the under constrained nature of the problem [2, 10, 11, 67, 81]. Learning morphable models from images has recently received increasing attention [38, 45, 68] together with learning 4D reconstruction [53]. Tracking approaches typically operate on RGB-D videos and either assume that a template of the instance is available [23, 26, 31, 34, 43], or that neighboring frames have short baselines [44, 52, 63, 82]. Our focus is on general articulating object categories that are observed in RGB images with large camera baselines, and have varying topologies and diverse shapes that make it difficult for the above methods.

Humans: Human bodies, hands, and faces are special cases of non-rigid articulating objects and have received special attention. Like many non-rigid reconstruction methods, methods for human shape rely on template models, priors, and the fixed topology of human shapes. Extensive work
has been done on using multiple views for shape reconstruction including motion [9, 16, 65]. These methods assume studio settings with numerous synchronized cameras, a template model, and fixed kinematic skeletons. Depth sensors can alleviate the need to have multiple synchronized cameras but many methods still use templates [7, 19, 66], with help from parametric body models [6, 49]. More recently, deep learning methods have successfully shown shape reconstruction from videos [4, 5] but they assume that the person does not move or change their articulation during capture. For single views, there exist methods for high-quality reconstruction [28, 37, 58, 59, 80], but fusion of multiple views of different poses remains an open problem. Our goal of reconstructing animatable shape is close to that of ARCH [33], although they rely heavily on body priors.

Shape Representations: There exist many choices for representing 3D shape from point clouds [17], depth images [64, 79], voxel grids [60], explicit meshes [9, 16], or implicit surfaces [13, 51, 56]. Methods based on implicit functions have shown promise in faithfully representing complex 3D shapes and are widely used in learning-based reconstruction [51, 56]. Recently, [30] proposed a dual representation that can be realized as either coarse manipulable point clouds or smooth implicit surface. Similarly, in our work, we adopt two representations: a canonical depth map to represent partial object shapes from multiple views [64] that can be easily animated, and implicit functions that faithfully reconstruct articulated shape [13].

3. Background

Canonicalization: Canonicalization refers to the task of mapping object observations (pixels or points) to a canonical container that normalizes object properties like position, orientation and size. It has been used recently as a tool for solving category-level 3D object understanding tasks such as 6 DoF pose estimation [46, 71], automatic labeling [78], and 3D reconstruction [54, 55, 64]. Canonicalization can be learned using full supervision [71] or without using any supervision [54] – we follow the supervised approach in this work. Specifically, we adopt Normalized Object Coordinate Space (NOCS) [71] as our representation.

NOCS and NAOCS: NOCS is a standard 3D container in which all instances of a certain category are consistently positioned, oriented and scaled to fit within a unit cube. Figure 2 (top) illustrates NOCS and how it maps different laptop instances with the same articulation to a container that canonicalizes for position, orientation and size. NAOCS [46] is an extension of NOCS to canonicalize for articulation in addition to position, orientation and scale. Figure 2 (bottom) illustrates NAOCS and how it maps different laptop instances with different articulations to a predetermined articulation. Doing so allows us to aggregate shape even if individual observations are in different poses.

Figure 2: (Top) NOCS [71] is a canonical container that maps different object instances to a unit cube with consistent position, orientation, and size. However, the articulation is left unchanged (laptop lid is articulated). (Bottom) NAOCS [46] is a container that additionally canonicalizes for articulation (laptop lid is set to a pre-determined pose). This allows the aggregation of shape even when they are in different articulations. The RGB colors denote the (x,y,z) position of a point within the container.

NOCS/NAOCS Map: A NOCS/NAOCS map is a projection of a 3D shape in the container onto a 2D image where each projected pixel carries the (x,y,z) position of the corresponding 3D point (denoted by RGB colors). These maps encode the partial shape of an object and are equivalent to a dense point cloud. NOCS/NAOCS maps provide a way to estimate 3D shape as 2D images within neural networks without dealing with 3D network designs. Additionally, if joints and part segments are available, NOCS maps can be reposed to create NAOCS maps.

We treat the 3D shape estimation task as finding a function $F: \mathcal{R}^{H \times W \times 3} \rightarrow \{\mathcal{R}^{H \times W \times 3}, \{0,1\}^{H \times W}\}$ defined on RGB images that predicts a per-pixel 3D coordinate as well as a binary foreground-background mask. Here $W$ and $H$ denote the width and height of the image. Like previous work, we realize $F$ via a 2D CNN [43, 64, 71], and train it using ground truth supervision. We also predict the position and orientation of joints and parts which allows us to repose NOCS maps to NAOCS maps. When multiple views are available, we can aggregate object shape into the same NAOCS container irrespective of the original articulation.

Alternative Formulations: Different from us, [46] used per-part NOCS containers (referred to as NPCS) for articulated object pose estimation. While this is important in their method, there are several reasons why we choose a different formulation. First, their approach solves a different problem

![Figure 2: (Top) NOCS [71] is a canonical container that maps different object instances to a unit cube with consistent position, orientation, and size. However, the articulation is left unchanged (laptop lid is articulated). (Bottom) NAOCS [46] is a container that additionally canonicalizes for articulation (laptop lid is set to a pre-determined pose). This allows the aggregation of shape even when they are in different articulations. The RGB colors denote the (x,y,z) position of a point within the container.](image)
Figure 3: Our method takes as input a set of sparse images $I_q$ and aims to reconstruct an animatable 3D shape $\hat{S}$. In stage 1, we use a SegNet-like architecture to predict: (1) the point cloud $x_q$ in NOCS as a NOCS map and mask $(M_q, B_q)$, (2) the part segmentation $P_q$, and (3) joint location and orientations in NOCS $J_q$. The network also outputs a high-dimensional unsupervised feature $f_q$. With the predicted part segmentation and joints, we can transform the point cloud into canonical articulation in NAOCS where inputs from arbitrary pose are aggregated ($\bar{x}$) along with enriched features $f$. ($\bar{x}, f$) are then fed into stage 2, an IF-Net, to output an reconstruction at any resolution, or articulation (stage 3).

4. Method

Given a sparse set of RGB views of an object $\{I_q \in R^{H\times W\times 3}\}^Q_{q=1}$, possibly at different articulations in each view, our goal is to reconstruct the 3D shape $\hat{S}$ of that object at arbitrary articulations. Achieving this requires not just estimating shape but also joints positions and orientations, and part segmentation for reposing objects. We also need to handle different articulations in input views – we use a category-level canonical 3D space to which the input views are lifted while stripping away the view-specific scale, pose, and articulation information. Our approach, depicted in Figure 3, has 3 stages: (1) lifting 2D views to a canonical 3D space, (2) surface reconstruction, and (3) reposing. We describe each of these stages in detail below.

4.1. Stage 1: Lifting to a Canonical 3D Space

To lift the RGB input images to 3D, we realize the function $F_{NOCS}(I_q) = \{M_q, B_q\}$ described in Section 3 using a CNN and estimate NOCS maps $M_q$ with per-pixel 3D coordinates along with a binary object mask $B_q$. These grant us an equivalent representation in the form of a point cloud $x_q = \{M_q(w,h)|B_q(w,h) = 1\} \in R^{N_q\times 3}$, where $N_q = \sum_{(w,h)} B_q(w,h)$ is the number of valid foreground pixels of the $q$-th view, and $(w, h)$ are the image width and height indices. While $\{x_q\}^Q_{q=1}$ are canonicalized for position, orientation, and size, we still cannot use them to perform multiview aggregation directly as each view presents the object in a different articulation. We therefore adopt the articulation-canonicalized maps (NAOCS) [46].

Instead of directly estimating NAOCS we structure the prediction by explicitly modeling the notion of articulation into the network. We do this by making the network estimate object parts and joints in addition to NOCS maps. We then use the parts and joints to repose the predicted NOCS map into a NAOCS map. Our structured prediction has two advantages. First, it enables an explicit access to the object pose thus allowing to conveniently animate it. Second, by supervising the reconstructions at all input articulations we avoid overfitting to the canonical one and strengthen the generalization capabilities of our network, as we show in Table 2.

Joint and Part Estimation: To facilitate articulation canonicalization and implicit modeling of articulations, we extend our map to $F_{NAOCS}(I_q) = \{M_q, B_q, P_q^1, \ldots, P_q^N, J_q^1, \ldots, J_q^N\}$ where we further predict for each view $N_p$ binary part segmentation masks $P_q \in \{0, 1\}^{H\times W}$ as well as a prediction of each joint location and its rotation angles with respect to a pre-defined rest pose $J_q \in R^6$ for each joint $J^j$ in an object category, and foreground pixel $p$ we predict a 6D vote $v_q^j(p) = (c_q^j(p), r_q^j(p))$ along with a 1D confidence score $w_q^j(p)$. The first 3 dimensions $c_q^j$ correspond to the joint center location, followed by 3 dimensions $r_q^j$ for joints orientation in axis angle form. We additionally output $C$-dimensional per pixel features (unsupervised) $f_q(p)$ which will be used in stage 2. Our approach only requires the number of joints and parts to be
uses a loss function our dataset (see Section 5.1). Stage 1 of our architecture mentation, all of which are fully supervised using labels in
The network estimates NOCS maps, joints, and part seg-
bar as well: $f = \bigcup_{i=1}^{Q} f_q^i$. Aside from these features, all other estimated parameters are fully supervised as will be
explained next.
Network, Supervision and Losses: We use a SegNet [8] based encoder-decoder architecture with skip connections. The network estimates NOCS maps, joints, and part seg-
mentall of which are fully supervised using labels in our dataset (see Section 5.1). Stage 1 of our architecture uses a loss function $L = \sum_{i=1}^{7} \lambda_i L_i$ consisting of 7 terms. The per-view NOCS map loss $L_1$ minimizes the L2 distance between the predicted 3D coordinates of foreground pixels. The predicted foreground binary mask $B$ is super-
vised using a binary cross entropy loss $L_2$. For the dense 6D joints prediction, we estimate a $N_J \times H \times W \times 6$ channel. The joint position loss $L_3$ minimizes the per-pixel L2 distance between the first 3 channels, where the joint orient-
tion loss $L_4$ minimizes the L2 distance for the last 3 channels. An additional joint consistency loss $L_5$ minimizes the distance between the per-view joint estimates in a pair-wise fashion. This provides sufficient supervision to learn meaningful confidence maps $w_q^i$ indirectly. For part segmentation, we use a standard multiclass cross entropy loss $L_6$. Finally, we also have a loss $L_7$ between the estimated per-
view canonical point cloud $x_q^{can}$ and the ground truth. All losses have a weight of 1.0, except the BCE loss which has a weight of 0.1.

4.2. Stage 2: Reconstruction

The output from stage 1 is an aggregated 3D point cloud enriched with features. However, this is insufficient for many downstream applications in animation and render-
ing for which we seek an implicit (occupancy) or explicit (mesh) 3D reconstruction. Such reconstruction has the added benefit of smoothing out outliers and noise in recon-
struction point clouds. Therefore, the second stage of our method takes an input point cloud (with unsupervised features) and produces an smooth implicit reconstruction from which we can extract a mesh. We build on top of IF-Net [13] which was recently proposed for reconstruction from point clouds in the form of an occupancy function. A compelling property of IF-Net is its hierarchical structure which, differ-
ently from other solutions in this category of neural implic-
ts, allows reasoning about object parts. This is especially important when dealing with articulations since while parts maintain the same geometry, the global structure changes entirely and may lead to poor generalization.

We modify IF-Net to support our problem using addi-
tional input channels. Specifically, the input to IF-Net is the aggregated articulation-canonicalized point cloud enriched with features $(\tilde{x}, \tilde{f}) \in \mathcal{R}^{|N_q| \times (3+C)}$. We discretize our unit-length canonical container into a $R \times R \times R$ voxel-grid, where each voxel contains a C-dimensional fea-
ture that is the average of all the points contained in that voxel. Following IF-Net, a hierarchical point encoding $(F_1(\tilde{x}_1), ..., F_K(\tilde{x}_i))$ is performed via a 3D CNN and fed to a point decoder $f(x_i) : F_1(\tilde{x}_1), ..., F_K(\tilde{x}_i) \rightarrow [0, 1]$. All reconstructions shown in the paper use the output occupancy processed to a mesh using marching cubes [50].

Now the goal of the additional features becomes clear. Different from the setting in IF-Net where the (possibly partial) point-cloud is accurately located, our point locations may suffer from inaccuracies due to errors in either of the
We train both stage 1 and 2 of our method end-to-end which allows these additional features to be used to improve 3D shape and acts as an implicit denoising mechanism.

**Network, Supervision and Losses**: The second stage uses a single-view reconstruction version IF-Net with 11 3D convolutional layers with different granularity and 3 fully-connected layers (more details are in the supplemental). The occupancy is fully supervised by labels our dataset (see Section 5.1). We use a binary cross-entropy loss between the ground truth occupancy and our estimate.

### 4.3. Stage 3: Reposing and Animation

Using the above two stages, we can obtain both the aggregated articulation-canonicalized point clouds as well as a complete 3D reconstruction in implicit or explicit form. However, our approach provides even more flexibility – we can **repose and animate** the reconstructed object to an articulation of our choice. We train stage 2 of our network to reconstruct the original ground truth articulation shape of the input views in addition to the articulation-canonicalized shape. At inference time, we allow the user to freely choose a per-joint rotation angle as a natural reposing interface (see Stage 3 in Figure 3). When this rotation angle is input to stage 2, we can reconstruct the object in arbitrary articulations. We also propose an alternative strategy for more consistent animation that takes the predicted joints and parts to repose the canonical mesh using a standard rigid skinning procedure (more details in the supplementary materials).

## 5. Results & Experiments

We evaluate StrobeNet on our newly generated benchmark of articulated shapes and their multi-view images. After describing the curation of the dataset and supervision signals (5.1) we describe the baseline methods in comparison (5.2). We then present the performance of our StrobeNet under different settings (5.3), followed by an ablation study of key ingredients of our method (5.4). Finally, we demonstrate the reposing of reconstructed objects in 5.5.

### 5.1. Datasets

We utilize two data sources of manipulable 3D piecewise rigid objects to build our dataset: SAPIEN [76] and shape2motion [73]. Both provide a medium-sized collection of objects from different categories with controllable part articulation. We generated a dataset of 3 object classes (common in related work [56, 64]): “eyeglasses”, “laptop” and “oven” at various combinations of 6D pose and articulation. We note that in each category, all objects share the same number of parts and joints. We randomly sample 20 part-pose combinations per object and render it from 8 different view angles and ranges. We split our dataset between (48, 47, 26) training shapes of glasses, laptops, and ovens, respectively, and a test set of unseen (5, 5, 3) shapes. Due to the relatively fewer shapes, we further augment the 3D models using non-uniform scaling, and images gamma correction resulting in a total of (42400, 33280, 46400) samples per category.

### 5.2. Baseline methods

To the best of our knowledge, there are no methods that are directly applicable to this new problem setting. We therefore resorted to the closest related setting and chose OccNet [51] as a representative state-of-the-art single-view reconstruction method. To modify it for our setting, we trained two variants of OccNet on the articulated objects dataset. **OccNet-F** is trained to reconstruct the objects in the articulation presented at the input view. When more than one view is provided, we stack the input views and supervise using the 3D shape in the articulation of the first view in the stack. **OccNet-C** is instead trained to predict the 3D object in its articulation **canonical** state. We train 3 models of each network variant on 1, 2 and 4 input views respectively and compare their reconstruction with our StrobeNet. Note that while OccNet-C and OccNet-F are trained to aggregate geometry from multiple views, their architecture is inherently limited compared with StrobeNet in two ways: (a) they cannot generalize to an arbitrary number of views at inference (we demonstrate this in Figure 6), and (b) their output articulation is not animatable.

![Figure 5: Qualitative results with 2 input views. From left to right: 2 input views, OccNet-C (baseline), StrobeNet (Ours) and ground truth. StrobeNet produces better reconstructions with fine-grained details.](image-url)
5.3. Articulated Shape Aggregation

We evaluate our method on the articulated objects test split. In Table 1 we report the mesh reconstruction quality w.r.t the articulation-canonicalized mesh using 1, 2 and 4 input views. We use two standard metrics: 3D Intersection over Union (IoU), and Chamfer Distance (CD) between points sampled from the meshes, and note that CD is a better measure of the detail precision than IoU which instead provides a coarser estimate. This distinction is especially important when comparing shape in a canonical articulation where different instances of the same category have a significant overlap. It can be seen that StrobeNet consistently outperforms the baseline OccNet-C on both metrics by a large margin across the different categories. Especially note the large improvement in CD.

In Figure 5 we visualize the reconstruction quality of 2 instances from each category by OccNet-C and our StrobeNet. As can be seen, StrobeNet manages to capture detailed geometric properties of the input instance. In contrast, OccNet-C seems to neglect small-scale details as especially emphasized on the laptop keyboard and oven panel. In the glasses category, while OccNet-C generates plausible outputs, we noticed that they often do not match the true object geometry. This further highlights the previously discussed difference between CD and IoU. A possible explanation for the “regression to the mean” effect seen on OccNet-C is that supervision is done only at a canonical pose. StrobeNet on the other hand benefits from its structured prediction that enforces attention to details for the explicit canonicalization to succeed.

In Table 2 we evaluate StrobeNet on reconstructing in the input pose. We emphasize that existing methods (e.g. [51]) can reconstruct 3D shape from single images, they cannot aggregate multiple views to complete the geometry and instead hallucinate it. StrobeNet, in contrast, is specially designed to do so. This is demonstrated by the markedly lower reconstruction errors compared with OccNet-F. Albeit strictly superior wrt the baseline, interestingly StrobeNet seems to perform worse when measured on the input pose than on the canonical pose. We attribute this to the error in reposing. Note, however, that this error does not reflect a loss of details (see Figure 7) but rather that a small but consistent mismatch in pose (e.g. laptop lid angle) would cause large reconstruction errors.

5.4. Ablation study

**Design choices:** In Section 4 we discuss two main ingredients of our network architecture: (1) consistency loss: used to regularize the canonical reconstructions from different views; and (2) local features: used to propagate image cues through the lifted points. We evaluate their contribution to the overall performance of StrobeNet in Table 3 by comparing the reconstruction quality of our method in its complete form (top row) with its reduced variants. As can be seen, both ingredients play an important role in providing an overall improved performance.

**Adding views at inference:** A property of our pipeline is that thanks to canonicalization, it can support an arbitrary number of input views at inference even if trained on a smaller set size. This is confirmed by Figure 6 where we ran inference on 1 to 6 input views using a StrobeNet trained on only 2 views. As can be seen the reconstruction performance keeps increasing with the addition of input views.

5.5. Reposing

A unique feature of StrobeNet is its animatable output. By keeping the intermediate representations of the reconstructed 3D shape as a union of lifted pointclouds and joint predictions, we allow user-provided pose input to repose the pointclouds before passed to the reconstruction module. Since the reconstruction module is trained to reconstruct at input poses, it can gracefully support this appli-
forms an OccNet\cite{maturana2015voxnet}-based baseline in both CD and IoU. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations. We believe that non-rigid objects (e.g., clothing) whose deformation cannot be efficiently modeled using articulations.

In Figure 7, we visualize the reposing of a reconstruction from 2 input views, into multiple output articulations. We implement two types of reposing. The first, which we demonstrate on the laptop and oven categories (bottom two rows) follows the network reposing explained in section 4, namely: reposing the pointcloud from which a reconstruction is performed. The second reposing technique, which we demonstrate using the eyeglasses, reposes the mesh directly by starting at the canonical mesh reconstruction and modifying each vertex according to its nearest neighbor canonical point cloud joint assignment. We invite the reader to check the supplementary for more animations.

6. Limitations & Future Work

Our approach has several limitations. First, due to limited training data, we focused on synthetic animatable datasets. Extending our work to real scenes and is an important step forward. In Figure 1 we show first encouraging results on real data. Second, we assume a fixed number of joints per category, but this is not always true (e.g., chairs that can swivel, fold, etc.). Lastly, there is a large class on non-rigid objects (e.g. clothing) whose deformation cannot be efficiently modeled using articulations. We believe that multi-view aggregation of deformable objects from sparse input view is an important research direction and hope to see more research done along this line.

### Table 1: Articulation-canonicalized shape reconstruction from different number of input views. Our StrobeNet (Ours) consistently outperforms an OccNet\cite{maturana2015voxnet}-based baseline in both CD and IoU over all categories.

| Class | Metric | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ |
|-------|--------|------|-------|------|-------|------|-------|------|-------|
| Single view | OccNet-C | 5.38 ± 3.11 | 22.94 ± 11.53 | 1.31 ± 0.27 | 50.92 ± 10.04 | 2.31 ± 0.59 | 50.46 ± 13.16 | 2.96 | 41.44 |
| | StrobeNet (Ours) | 3.72 ± 2.88 | 24.45 ± 9.04 | 0.83 ± 0.21 | 60.77 ± 10.36 | 1.42 ± 0.53 | 63.75 ± 12.04 | 2.55 | 49.66 |
| 2 views | OccNet-C | 4.47 ± 2.63 | 23.75 ± 13.43 | 1.08 ± 0.24 | 58.24 ± 9.10 | 2.09 ± 0.53 | 55.01 ± 11.49 | 2.54 | 45.67 |
| | StrobeNet (Ours) | 1.59 ± 0.60 | 31.59 ± 8.33 | 0.77 ± 0.21 | 64.94 ± 11.73 | 1.40 ± 0.41 | 67.98 ± 11.46 | 1.25 | 54.84 |
| 4 views | OccNet-C | 5.51 ± 4.02 | 28.17 ± 13.88 | 1.03 ± 0.19 | 60.30 ± 6.62 | 1.95 ± 0.53 | 57.19 ± 12.48 | 2.83 | 48.55 |
| | StrobeNet (Ours) | 1.78 ± 0.78 | 32.58 ± 7.74 | 0.73 ± 0.14 | 65.21 ± 5.93 | 1.34 ± 0.53 | 70.91 ± 12.32 | 1.28 | 56.23 |

### Table 2: Reconstruction in Input Shape shape reconstruction from different number of input views. Our StrobeNet (Ours) consistently outperforms an OccNet\cite{maturana2015voxnet}-based baseline in both CD and IoU.

| Class | Metric | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ |
|-------|--------|------|-------|------|-------|------|-------|------|-------|
| Single view | OccNet-F | 9.65 ± 2.69 | 19.97 ± 11.73 | 1.42 ± 0.27 | 52.31 ± 9.00 | 5.00 ± 1.93 | 51.72 ± 12.91 | 5.36 | 43.33 |
| | StrobeNet | 4.19 ± 2.13 | 13.54 ± 8.02 | 1.08 ± 0.31 | 56.17 ± 11.06 | 2.12 ± 1.42 | 61.32 ± 9.17 | 2.46 | 43.67 |
| 2 views | OccNet-F | 8.49 ± 2.37 | 27.77 ± 8.59 | 1.75 ± 0.47 | 45.77 ± 9.58 | 4.82 ± 1.78 | 50.25 ± 10.36 | 5.02 | 41.26 |
| | StrobeNet | 2.16 ± 0.81 | 21.26 ± 8.83 | 1.23 ± 0.52 | 50.71 ± 11.96 | 1.95 ± 0.89 | 60.86 ± 8.87 | 1.78 | 44.28 |
| 4 views | OccNet-F | 9.10 ± 2.69 | 28.20 ± 10.09 | 1.36 ± 0.30 | 56.98 ± 7.88 | 3.13 ± 1.21 | 52.95 ± 10.41 | 4.53 | 46.04 |
| | StrobeNet | 2.00 ± 0.74 | 31.13 ± 8.50 | 1.06 ± 0.34 | 57.85 ± 10.34 | 1.95 ± 0.67 | 63.31 ± 8.15 | 1.67 | 50.76 |

### Table 3: Our ablation on two pipeline ingredient: consistency loss and local features shows their effectiveness in improving overall reconstruction accuracy.

| Class | Metric | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ | CD ↓ | IoU ↑ |
|-------|--------|------|-------|------|-------|------|-------|------|-------|
| Ours | 1.59 ± 0.60 | 31.59 ± 8.33 | 0.77 ± 0.21 | 64.94 ± 11.73 | 1.40 ± 0.41 | 67.98 ± 11.46 | 1.25 | 54.84 |
| w/o consistency | 2.09 ± 0.81 | 29.54 ± 9.73 | 0.83 ± 0.20 | 60.03 ± 9.34 | 1.42 ± 0.36 | 63.83 ± 7.26 | 1.44 | 51.13 |
| w/o local feature | 2.53 ± 1.71 | 25.94 ± 10.28 | 0.81 ± 0.22 | 61.66 ± 11.33 | 1.38 ± 0.49 | 65.81 ± 10.47 | 1.98 | 51.14 |
References

[1] Antonio Agudo, Melcior Pijoan, and Francesc Moreno-Noguer. Image collection pop-up: 3d reconstruction and clustering of rigid and non-rigid categories. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2607–2615, 2018. 2

[2] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. Advances in neural information processing systems, 21:41–48, 2008. 1, 2

[3] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Trajectory space: A dual representation for nonrigid structure from motion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(7):1442–1456, 2010. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In 2018 International Conference on 3D Vision (3DV), pages 98–109. IEEE, 2018. 3

[5] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8387–8397, 2018. 3

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In ACM SIGGRAPH 2005 Papers, pages 408–416. 2005. 3

[7] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In Consumer Depth Cameras for Computer Vision, pages 71–98. Springer, 2013. 3

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017. 5

[9] Luca Ballan, Aparna Taneya, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In European Conference on Computer Vision, pages 640–653. Springer, 2012. 3

[10] Adrien Bartoli, Vincent Gay-Bellile, Umberto Castellani, Julien Peyras, Sören Olsen, and Patrick Sady. Coarse-to-fine low-rank structure-from-motion. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. 2

[11] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662), volume 2, pages 690–696. IEEE, 2000. 1, 2

[12] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In Advances in Neural Information Processing Systems, pages 9609–9619, 2019. 2

[13] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun 2020. 2, 3, 5

[14] Christopher Bongsoo Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. CoRR, abs/1604.00449, 2016. 2

[15] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision, 107(2):101–122, 2014. 2

[16] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In ACM SIGGRAPH 2008 papers, pages 1–10. 2008. 3

[17] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 605–613, 2017. 1, 2, 3

[18] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence, 32(8):1362–1376, 2009. 2

[19] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 755–762. IEEE, 2010. 3

[20] Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3d reconstruction. Advances in Neural Information Processing Systems, 33, 2020. 2

[21] R. Girdhar, D.F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. 2016. 2

[22] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), volume 2, pages 2402–2409. IEEE, 2006. 1, 2

[23] Vladislav Golyanik, Torben Fetzer, and Didier Stricker. Accurate 3d reconstruction of dynamic scenes from monocular image sequences with severe occlusions. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 282–291. IEEE, 2017. 2

[24] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. CoRR, abs/1802.05384, 2018. 1

[25] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 2

[26] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In Proceedings
[53] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In Proceedings of the IEEE International Conference on Computer Vision, pages 5379–5389, 2019.

[54] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpn: Canonical 3d pose networks for non-rigid structure from motion. In Proceedings of the IEEE International Conference on Computer Vision, pages 7688–7697, 2019.

[55] David Novotny, Roman Shapovalov, and Andrea Vedaldi. Canonical 3d deformers maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction. Advances in Neural Information Processing Systems, 33, 2020.

[56] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsf: Learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103, 2019.

[57] Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. arXiv preprint arXiv:1711.10669, 2017.

[58] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[59] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 84–93, 2020.

[60] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. International Journal of Computer Vision, 35(2):151–173, 1999.

[61] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Niessner, Gordon Wetzstein, and Michael Zollhoefer. Deepvoxels: Learning persistent 3d feature embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[62] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In Advances in Neural Information Processing Systems, pages 1121–1132, 2019.

[63] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1386–1395, 2017.

[64] Srinath Sridhar, Davis Rempe, Julien Valentin, Sofien Bouaziz, and Leonidas J. Guibas. Multiview aggregation for learning category-specific shape reconstruction. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

[65] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. IEEE computer graphics and applications, 27(3):21–31, 2007.

[66] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. ACM Transactions on Graphics (ToG), 35(6):1–11, 2016.

[67] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE transactions on pattern analysis and machine intelligence, 30(5):878–892, 2008.

[68] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. arXiv preprint arXiv:2007.08504, 2020.

[69] Shubham Tulsiani, Or Litany, Charles R. Qi, He Wang, and Leonidas J. Guibas. Object-centric multi-view aggregation, 2020.

[70] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2626–2634, 2017.

[71] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shurong Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[72] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), pages 52–67, 2018.

[73] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qingping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8876–8884, 2019.

[74] Olivia Wiles and Andrew Zisserman. Silnet: Single-and multi-view reconstruction by learning from silhouettes. arXiv preprint arXiv:1711.07888, 2017.

[75] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum. MarrNet: 3D Shape Reconstruction via 2.5D Sketches. 2017.

[76] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11097–11107, 2020.

[77] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In Advances in Neural Information Processing Systems, pages 492–502, 2019.

[78] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12224–12233, 2020.
[79] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In *Advances in Neural Information Processing Systems*, pages 2263–2274, 2018.

[80] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):901–914, 2018.

[81] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1542–1549, 2014.

[82] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):1–12, 2014.