

Deep Structured Implicit Functions

Kyle Genova^{1,2} Forrester Cole² Avneesh Sud² Aaron Sarna² Thomas Funkhouser^{1,2}

¹Princeton University ²Google Research

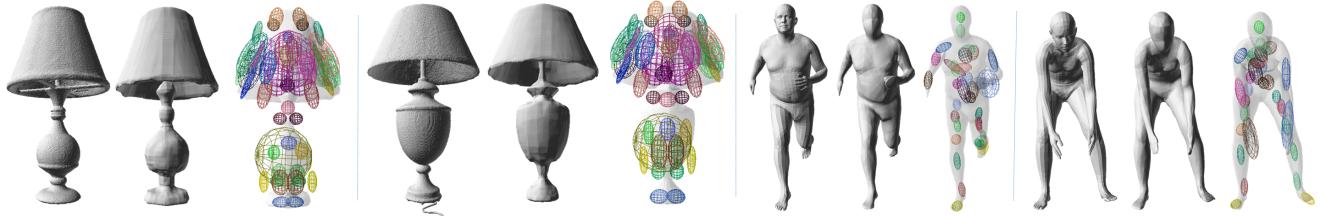


Figure 1. This paper introduces Deep Structured Implicit Functions, a 3D shape representation that decomposes an input shape (mesh on left in every triplet) into a structured set of shape elements (colored ellipses on right) whose contributions to an implicit surface reconstruction (middle) are represented by latent vectors decoded by a deep network.

Abstract

The goal of this project is to learn a 3D shape representation that enables accurate surface reconstruction, compact storage, efficient computation, consistency for similar shapes, generalization across diverse shape categories, and inference from depth camera observations. Towards this end, we introduce Deep Structural Implicit Functions (DSIF), a 3D shape representation that decomposes space into a structured set of local deep implicit functions. We provide networks that infer the space decomposition and local deep implicit functions from a 3D mesh or posed depth image. During experiments, we find that it provides 10.3 points higher surface reconstruction accuracy (F-Score) than the state-of-the-art (OccNet), while requiring fewer than 1% of the network parameters. Experiments on posed depth image completion and generalization to unseen classes show 15.8 and 17.8 point improvements over the state-of-the-art, while producing a structured 3D representation for each input with consistency across diverse shape collections.

1. Introduction

Representing 3D shape is a fundamental problem with many applications, including surface reconstruction, analysis, compression, matching, interpolation, manipulation, and visualization. For most vision applications, a 3D rep-

resentation should support: (a) reconstruction with accurate surface details, (b) scalability to complex shapes, (c) support for arbitrary topologies, (d) generalizability to unseen shape classes, (e) independence from any particular application domain, (f) encoding of shape priors, (g) compact storage, and (h) computational efficiency.

No current representation has all of these desirable properties. Traditional explicit 3D representations (voxels, meshes, point clouds, etc.) provide properties (a-e) above. They can represent arbitrary shapes and any desired detail, but they require storage and computation proportional to the shape complexity and reconstruction accuracy, and they don't encode shape priors helpful for 3D completion and reconstruction tasks. In contrast, learned representations (latent vectors and deep network decoders) excel at representing shapes compactly with low-dimensional latent vectors and encoding shape priors in network weights, but they struggle to reconstruct details for complex shapes or generalize to shape classes outside the training distribution.

Most recently, deep implicit functions (DIF) have been shown to be highly effective for reconstruction of individual objects [21, 8, 24, 42]. They represent an input observation as a latent vector \mathbf{z} and train a neural network to estimate the inside/outside or signed-distance function $f(\mathbf{x}, \mathbf{z})$ given a query location \mathbf{x} in 3D space. This approach achieves state of the art results for several 3D shape reconstruction tasks. However, they use a single, fixed-length latent feature vector to represent the entirety of all shapes and they evaluate a

complex deep network to evaluate the implicit function for every position \mathbf{x} . As a result, they support limited shape complexity, generality, and computational efficiency.

Meanwhile, new methods are emerging for learning to infer structured decomposition of shapes [36, 12]. For example, [12] recently proposed a network to encode shapes into Structured Implicit Functions (SIF), which represents an implicit function as a mixture of local Gaussian functions. They showed that simple networks can be trained to decompose diverse collections of shapes consistently into SIFs, where the local shape elements inferred for one shape (e.g., the leg of a chair) correspond to semantically similar elements for others (e.g., the leg of a table). However, they did not use these structured decompositions for accurate shape reconstruction due to the limited shape expressivity of their local implicit functions (Gaussians).

The key idea of this paper is to develop a pipeline that can learn to infer *Deep Structured Implicit Functions* (DSIF), a representation of 3D shape as a structured set of local deep implicit functions (Figure 1). This DSIF representation is similar to SIF in that it decomposes a shape into a set of overlapping local regions represented by Gaussians; however it also associates a latent vector with each local region that can be decoded with a DIF to produce finer geometric detail. Alternately, DSIF is similar to DIF in that it encodes a shape as a latent vector that can be evaluated with a neural network to estimate the inside/outside function $f(\mathbf{x}, \mathbf{z})$ for any location x ; however, the DSIF latent vector is decomposed into parts associated with local regions of space (SIF Gaussians), which makes it more scalable, generalizable, and computationally efficient.

In this paper, we not only propose the DSIF representation, but we also provide a common system design that works effectively for 3D autoencoding, depth image completion, and partial surface completion. First, we propose to use DIF to predict local functions that are *residuals* with respect to the Gaussian functions predicted by SIF – this choice simplifies the task of the DIF, as it must predict only fine details rather than the overall shape within each shape element. Second, we propose to use the SIF decomposition of space to focus the DIF encoder on local regions by gathering input 3D points within each predicted shape element and encoding them with PointNet [27]. Finally, we investigate several significant improvements to SIF (rotational degrees of freedom, symmetry constraints, etc.) and simplifications to DIF (fewer layers, smaller latent codes, etc.) to improve the DSIF representation. Results of ablation studies show that each of these design choices provides significant performance improvements over alternatives. In all, DSIF achieves 10-15 points better F-Score performance on shape reconstruction benchmarks than the state-of-the-art [21], with fewer than 1% of the network parameters.

2. Related Work

Traditional Shape Representations: There are many existing approaches for representing shape. In computer graphics, some of the foundational representations are meshes [2], point clouds [11], voxel grids [11], and implicit surfaces [28, 3, 4, 22, 23, 41]. These representations are popular for their simplicity and ability to operate efficiently with specialized hardware. However, they lack two important properties: they do not leverage a shape prior, and they can be inefficient in their expressiveness. Thus, traditional surface reconstruction pipelines based on them, such as Poisson Surface Reconstruction [17], require a substantial amount of memory and computation and are not good for completing unobserved regions.

Learned Shape Representations: To leverage shape priors, shape reconstruction methods began representing shape as a learned feature vector, with a trained decoder to a mesh [33, 13, 38, 14, 16], point cloud [10, 19, 43], voxel grid [9, 40, 6, 39], or octree [34, 30, 29]. Most recently, representing shape as a vector with an implicit surface function decoder has become popular, with methods such as Oc-Net [21], ImNet [8], DeepSDF [24], and DISN [42]. These methods have substantially improved the state of the art in shape reconstruction and completion. However, they do not scale or generalize very well because the fundamental representation is a single fixed-length feature vector representing a shape globally.

Structured Shape Representations: To improve scalability and efficiency, researchers have introduced structured representations that encode the repeated and hierarchical nature of shapes. Traditional structured representations include scene graphs [11], CSG trees [11], and partition of unity implicits [23], all of which represent complex shapes as the composition of simpler ones. Learned structured representations include CSGNet [32], GRASS [18], Volumetric Primitives [36], Superquadrics [25], and SIF [12]. These methods can decompose shapes into simpler ones, usually with high consistency across shapes in a collection. However, they have been used primarily for shape analysis (e.g. part decomposition, part-aware correspondence), not for accurate surface reconstruction or completion.

3. Deep Structured Implicit Functions

In this paper, we propose a new 3D shape representation based on Structured Deep Implicit Functions (DSIF). The DSIF is a function that can be used to classify whether a query point \mathbf{x} is inside or outside a shape. It is represented by a set of N shape elements, each parameterized by 10 analytic shape variables θ_i and M latent shape variables \mathbf{z}_i :

$$\text{DSIF}(\mathbf{x}, \Theta, \mathbf{Z}) = \sum_{i \in [N]} g(\mathbf{x}, \theta_i)(1 + f(\mathbf{x}, \mathbf{z}_i)) \quad (1)$$

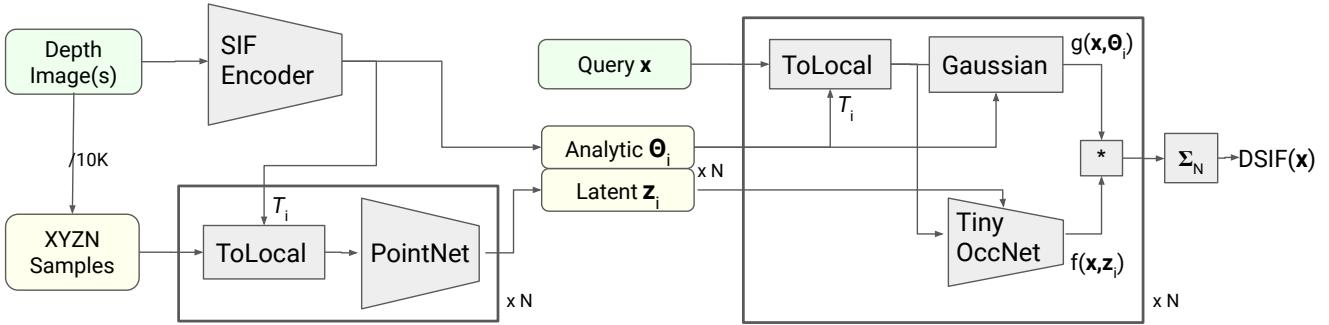


Figure 2. Network architecture. Our system takes in one or more posed depth images and outputs a DSIF function that can be used to classify inside/outside for any query point \mathbf{x} . It starts with a SIF encoder to extract a set of overlapping shape elements, each defined by a local Gaussian region of support parameterized by θ_i . It then extracts sample points/normals from the depth images and passes them through a PointNet encoder for each shape element to produce a latent vector \mathbf{z}_i . A TinyOccNet is used to decode each \mathbf{z}_i to produce an implicit function $f_i(\mathbf{x}, \mathbf{z}_i)$, which is combined with the local Gaussian function $g(\mathbf{x}, \theta_i)$ and summed with other shape elements to produce the output function $DSIF(\mathbf{x})$.

where $g(\mathbf{x}, \theta_i)$ is a local analytic implicit function and $f(\mathbf{x}, \mathbf{z}_i)$ is a deep implicit function. Intuitively, g provides a density function that defines a coarse shape and region of influence for each shape element, and f provides the shape details that cannot be represented by g .

Like a typical deep implicit function (DIF), our DSIF represents a 3D shape as an isocontour of an implicit function decoded with a deep network conditioned on predicted latent variables. However, our DSIF replaces the (possibly long) single latent code of a typical DIF with the concatenation of N pairs of analytic parameters θ_i and short latent codes \mathbf{z}_i – i.e., the global implicit function is decomposed into the sum of N local implicit functions. This key difference helps it to be more accurate, efficient, consistent, scalable, and generalizable (see Section 6).

Analytic shape function. The analytic shape function g defines a coarse density function and region of influence for each shape element. Any simple analytic implicit function with local support would do. We use an oriented, anisotropic, 3D Gaussian:

$$g(\mathbf{x}, \theta_i) = c_i e^{-\frac{\|\mathbf{T}_i \mathbf{x}\|^2}{2}} \quad (2)$$

where the parameter vector θ_i consists of ten variables: one for a scale constant c_i , three for a center point \mathbf{p}_i , three radii \mathbf{r}_i , and three Euler angles \mathbf{e}_i (this is the same parameterization as [12], except with 3 additional DoFs for rotation). The last 9 variables imply an affine transformation matrix \mathbf{T}_i that takes a point \mathbf{x} from object space coordinates to the local isotropic, oriented, centered coordinate frame of the shape element.

Deep shape function. The deep implicit function f defines local shape details within a shape element by modulating g (one f function is shared by all shape elements). To compute f , we use a network architecture based on Occupancy

Networks [21] (we call ours TinyOccNet). As in the original OccNet, ours is organized as a fully-connected network conditioned on the latent code \mathbf{z}_i and trained using conditional batch normalization. However, one critical difference is that we transform the point \mathbf{x} by \mathbf{T}_i before feeding it to the network:

$$f(\mathbf{x}, \mathbf{z}_i) = \text{TinyOccNet}(\mathbf{T}_i \mathbf{x}, \mathbf{z}_i) \quad (3)$$

Another critical difference is that f_i only modulates the local implicit function g_i rather than predicting an entire, global function. As a result, TinyOccNet has fewer network layers (9 vs. 33), shorter latent codes (32 vs. 256), and many fewer network parameters (8.6K vs 2M) than the original OccNet, and still achieves higher overall accuracy (see Section 6).

Symmetry constraints. For shape collections with man-made objects, we constrain a subset of the shape elements (half) to be symmetric with respect to a selected set of transformations (reflection across a right/left bisecting plane). These “symmetric” shape elements are evaluated twice for every point query, once for \mathbf{x} and once for $S\mathbf{x}$, where S is the symmetry transformation. In doing so, we effectively increase the number of shape elements without having to compute/store extra parameters for them. Adding partial symmetry encourages the shape decomposition to match global shape properties common in many shape collections and gives a boost to accuracy (Table 4).

4. Processing Pipeline

The processing pipeline for computing DSIF is shown in Figure 2. All steps of the pipeline are differentiable and trained end-to-end. At inference time, the input to the system is a 3D surface or depth image, and the output is a set of shape element parameters Θ and latent codes Z for each of

N overlapping local regions, which can be decoded to predict inside/outside for any query location \mathbf{x} . Complete surfaces can be reconstructed for visualization by evaluating DSIF(\mathbf{X}) at points on a regular grid and running Marching Cubes [20].

The exact configuration of the encoder architecture varies with input data type. We encode a **3D mesh** by first rendering a stack of 20 depth images at 224 x 224 resolution from a fixed set of equally spaced views surrounding the object. We then give the depth images to an early-fusion ResNet50 [15] to regress the shape element parameters Θ . Meanwhile, we generate a set of 10K points with normals covering the whole shape by estimating normals from the depth image(s) and unprojecting randomly selected pixels to points in object space using the known camera parameters. Then, for each shape element, we select a sampling of 1K points with normals within the region of influence defined by the predicted analytic shape function, and pass them to a PointNet [27] to generate the latent code \mathbf{z}_i . Alternatively, we could have encoded 3D input surfaces with CNNs based on mesh, point, or voxel convolutions, but found this processing pipeline to provide a good balance between detail, attention, efficiency, and memory. In particular, since the local geometry of every shape element is encoded independently with a PointNet, it is difficult for the network to “memorize” global shapes and it therefore generalizes better.

We encode a **depth image** with known camera parameters by first converting it into a 3 channel stack of 224 x 224 images representing the XYZ position of every pixel in object coordinates. We then feed those channels into a ResNet50 to regress the shape element parameters Θ , and we regress the latent codes \mathbf{Z} for each shape element using the same process as for 3D meshes.

4.1. Training Losses

The pipeline is trained with the following loss L :

$$L(\Theta, \mathbf{Z}) = w_P L_P(\Theta, \mathbf{Z}) + w_C L_C(\Theta) \quad (4)$$

Point Sample Loss L_P . The first loss L_P measures how accurately the DSIF(\mathbf{x}) predicts inside/outside of the ground-truth shape. To compute it, we sample 1024 points near the ground truth surface (set \mathcal{S}) and 1024 points uniformly at random in the bounding box of the shape (set \mathcal{U}). We combine them with weights $w_i \in \{w_{\mathcal{S}}, w_{\mathcal{U}}\}$ to form set $C = \mathcal{U} \cup \mathcal{S}$. The near-surface points are computed using the sampling algorithm of [12]. We scale by a hyperparameter α , apply a sigmoid to the decoded value DSIF(\mathbf{x}), and then compute an L_2 loss to the ground truth indicator function $I(\mathbf{x})$ (see [12] for details):

$$L_P(\Theta, \mathbf{Z}) = \frac{1}{|C|} \sum_{\mathbf{x}_i \in C} w_i \|\text{sig}(\alpha \text{DSIF}(\mathbf{x}_i, \Theta, \mathbf{Z})) - I(\mathbf{x}_i)\|$$

Shape Element Center Loss L_C . The second loss L_C encourages the center of every shape element to reside within the target shape. To compute it, we estimate a signed distance function on a low-res 32x32x32 grid \mathbf{G} for each training shape. The following loss is applied based on the grid value $G(\mathbf{p}_i)$ at the center \mathbf{p}_i of each shape element:

$$L_C(\Theta) = \begin{cases} \sum_{\theta_i \in \Theta} G(\mathbf{p}_i)^2 & G(\mathbf{p}_i) > \beta \\ 0 & G(\mathbf{p}_i) \leq \beta \end{cases}$$

Here, β is a threshold chosen to account for the fact that \mathbf{G} is coarse. It is set to half the width of a voxel cell in \mathbf{G} . This setting makes it a conservative loss: it says that when \mathbf{p}_i is definitely outside the ground truth shape, \mathbf{p}_i should be moved inside. L_C never penalizes a center that is within the ground truth shape.

It is also possible for the predicted center to lie outside the bounding box of \mathbf{G} . In this case, there is no gradient for L_C , so we instead apply the inside-bounding-box loss from [12] using the object-space bounds of \mathbf{G} .

5. Experimental Setup

We execute a series of experiments to evaluate the proposed DSIF shape representation, compare it to alternatives, study the effects of its novel components, and test it in applications. Unless otherwise noted, we use $N = 32$ shape elements and $M = 32$ dimensional latent vectors during all experiments.

Datasets. When not otherwise specified, experiments are run on the ShapeNet dataset [7]. We use the train and test splits from 3D-R2N2 [9]. We additionally subdivide the train split to create an 85%, 5%, 10% train, validation, and test distribution. We pre-process the shapes to make them watertight using the depth fusion pipeline from Occupancy Networks [21]. We train models multi-class (all 13 classes together) and show examples only from the test split.

Metrics. We evaluate shape reconstruction results with mean intersection-over-union (IoU) [21], mean Chamfer distance [21], and mean F-Score [35] at $\tau = 0.01$. As suggested in [35], we find that IoU is difficult to interpret for low values, and Chamfer distance is outlier sensitive, and so we focus our discussions mainly on F-Scores.

Baselines. We compare most of our results to the two most related prior works: Occupancy Networks [21] (OccNet), the state-of-the-art in deep implicit functions, and Structured Implicit Functions [12] (SIF), the state-of-the-art in structural decomposition.

6. Experimental Evaluation

In this section, we report results of experiments that compare DSIF and baselines with respect to how well they satisfy desirable properties of a 3D shape representation.

| Category | IoU | | | Chamfer | | | F-Score | | |
|------------|--------|------|-------------|--------------|-------|--------------|---------|------|-------------|
| | OccNet | SIF | Ours | OccNet | SIF | Ours | OccNet | SIF | Ours |
| airplane | 77.0 | 66.2 | 91.2 | 0.016 | 0.044 | 0.010 | 87.8 | 71.4 | 96.9 |
| bench | 71.3 | 53.3 | 85.6 | 0.024 | 0.082 | 0.017 | 87.5 | 58.4 | 94.8 |
| cabinet | 86.2 | 78.3 | 93.2 | 0.041 | 0.110 | 0.033 | 86.0 | 59.3 | 92.0 |
| car | 83.9 | 77.2 | 90.2 | 0.061 | 0.108 | 0.028 | 77.5 | 56.6 | 87.2 |
| chair | 73.9 | 57.2 | 87.5 | 0.044 | 0.154 | 0.034 | 77.2 | 42.4 | 90.9 |
| display | 81.8 | 69.3 | 94.2 | 0.034 | 0.097 | 0.028 | 82.1 | 56.3 | 94.8 |
| lamp | 56.5 | 41.7 | 77.9 | 0.167 | 0.342 | 0.180 | 62.7 | 35.0 | 83.5 |
| rifle | 69.5 | 60.4 | 89.9 | 0.019 | 0.042 | 0.009 | 86.2 | 70.0 | 97.3 |
| sofa | 87.2 | 76.0 | 94.1 | 0.030 | 0.080 | 0.035 | 85.9 | 55.2 | 92.8 |
| speaker | 82.4 | 74.2 | 90.3 | 0.101 | 0.199 | 0.068 | 74.7 | 47.4 | 84.3 |
| table | 75.6 | 57.2 | 88.2 | 0.044 | 0.157 | 0.056 | 84.9 | 55.7 | 92.4 |
| telephone | 90.9 | 83.1 | 97.6 | 0.013 | 0.039 | 0.008 | 94.8 | 81.8 | 98.1 |
| watercraft | 74.7 | 64.3 | 90.1 | 0.041 | 0.078 | 0.020 | 77.3 | 54.2 | 93.2 |
| mean | 77.8 | 66.0 | 90.0 | 0.049 | 0.118 | 0.040 | 81.9 | 59.0 | 92.2 |

Table 1. **Autoencoder results.** Comparison of 3D reconstruction errors for different autoencoders on the test set of 3D-R²N².

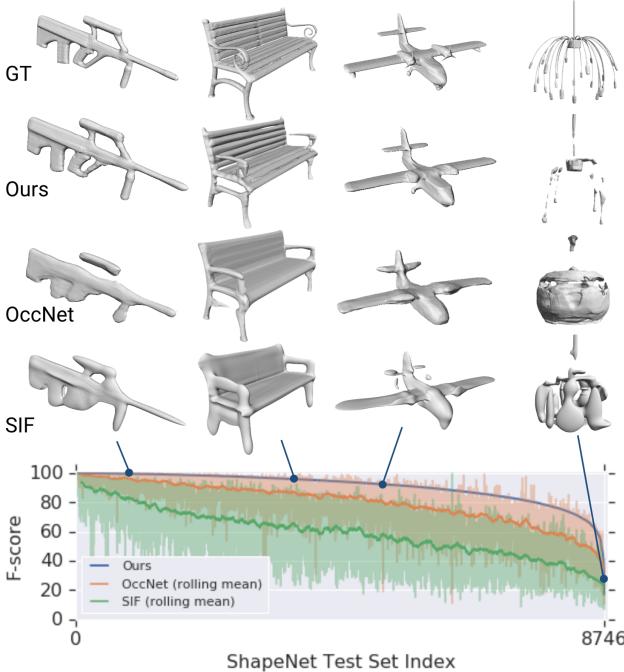


Figure 3. **Autoencoder examples.** F-scores for the test set (8746 shapes) are shown ordered by the DSIF F-score, with examples marked with their position on the curve. Our reconstructions (blue curve) are most accurate for 93% of shapes (exact scores shown faded). The scores of OccNet and SIF follow roughly the same curve as DSIF (rolling means shown bold), indicating shapes are similarly difficult for all methods. Solid shapes such as the rifle are relatively easy to represent, while shapes with irregular, thin structures such as the lamp are more difficult.

Accuracy. Our first experiment compares 3D shape representations in terms of how accurately they can encode/decode shapes. For each representation, we compare a 3D→3D autoencoder trained on the multiclass training data, use it to reconstruct shapes in the test set, and then evaluate how well the reconstructions match the originals (Table 1). DSIF’s mean F-Score is 92.2, 10.3 points higher

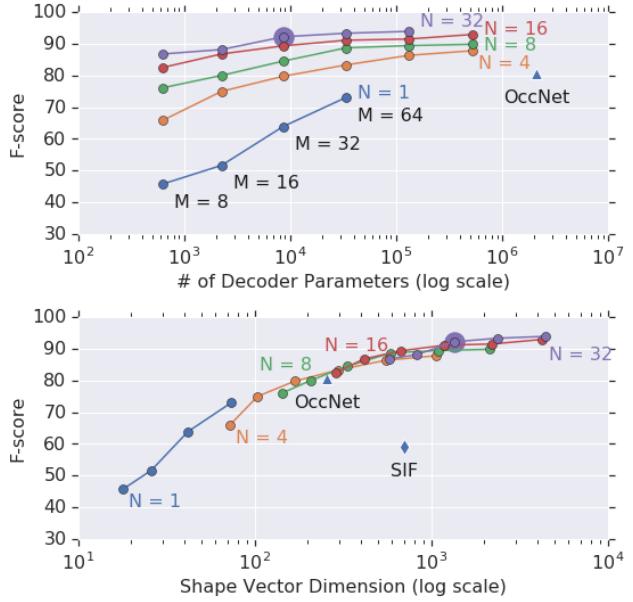


Figure 4. **Representation efficiency.** F-score vs. model complexity. Curves show varying M for constant N . Other methods marked as points. **Top:** F-score vs. count of decoder parameters. The $N = 32, M = 32$ configuration (large dot) reaches >90% F-score with <1% of the parameters of OccNet, and is used as the benchmark configuration in this paper. **Bottom:** F-score vs. shape vector dimension ($|\Theta| + |\mathbf{Z}|$ for DSIF). DSIF achieves similar reconstruction accuracy to OccNet at the same dimensionality, and can use additional dimensions to further improve accuracy.

than OccNet, and 33.2 points higher than SIF. A more detailed breakdown of the results appears in Figure 3, which shows the F-scores for all models in the test set – DSIF improves on OccNet’s score for 93% of test shapes. The increase in accuracy translates into a large qualitative improvement in results (shown above in Figure 3). For example, DSIF often reproduces better geometric details (e.g., back of the bench) and handles unusual part placements more robustly (e.g., handles on the rifle).

Efficiency. Our second experiment compares the efficiency of 3D shape representations in terms of accuracy vs. storage/computation costs. Since DSIF can be trained with different numbers of shape elements (N) and latent feature sizes (M), a family of DSIF representations is possible, each with a different trade-off between storage/computation and accuracy. Figure 4 investigates these trade-offs for several combinations of N and M and compares the accuracy of their autoencoders to baselines. Looking at the plot on the top, we see that DSIF performs comparably to OccNet and outperforms SIF at the same number of bytes, and is capable of scaling to larger embeddings. Similarly, the plot on the bottom shows that DSIF provides more accurate reconstructions than baselines at every decoder size – our decoder with $N = 32$ and $M = 32$ is 0.004× the size of OccNet

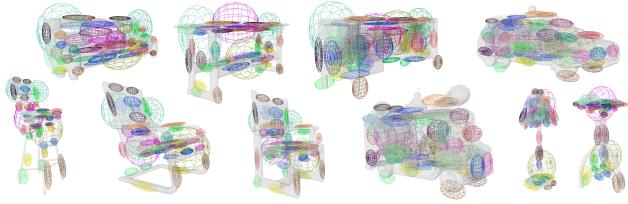


Figure 5. Representation consistency. Example shape decompositions produced by our model trained multi-class on 3D-R²N². Shape elements are depicted by their support ellipsoids and colored consistently by index. Note that the shape element shown in brown is used to represent the right-front leg of the chairs, tables, desks, and sofas, as well as the front-right wheel of the cars.

and provides 1.13× better F-Score.

Consistency. Our third experiment investigates the ability of DSIF to decompose shapes consistently into shape elements. This property was explored at length in [12] and shown to be useful for structure-aware correspondences, interpolations, and segmentations. While not the focus of this paper, we find qualitatively that the consistency of the DSIF representation is slightly superior to SIF, because the shape element symmetries and rotations introduced in this paper provide the DoFs needed to decompose shapes with fewer elements. On the other hand, the local DIFs are able to compensate for imperfect decompositions during reconstruction, which puts less pressure on consistency. Figure 5 shows qualitative results of the decompositions computed for DSIF. Please note the consistency of the colors (indicating the index of the shape element) across a broad range of shapes.

Generalizability. Our third experiment studies how well trained autoencoders generalize to handle unseen shape classes. To test this, we used the auto-encoders trained on 3D-R²N² classes and tested them without fine-tuning on a random sampling of meshes from 10 ShapeNet classes that were not seen during training. Table 2 shows that the mean F-Score for DSIF on these novel classes is 84.4, which is 17.8 points higher than OccNet and 41.4 points higher than SIF. Looking at the F-Score for every example in the bottom of Figure 6, we see that DSIF is better on 91% of examples. We conjecture this is because DSIF learns to produce consistent decompositions for a broad range of input shapes when trained multiclass, and because the local TinyOccNets learn to predict shape details only for local regions with their limited capacity. This two-level factoring of structure and detail seems to help DSIF generalize.

Domain-independence. Our fifth experiment investigates whether DSIF can be used in application domains beyond the man-made shapes found in ShapeNet. As one example, we trained DSIF without any changes to autoencode meshes of human bodies in a wide variety of poses sampled from [37]. Specifically, we generated 5M meshes by

| Category | Chamfer | | | F-Score | | |
|-----------|---------|--------|--------------|---------|--------|-------------|
| | SIF | OccNet | Ours | SIF | OccNet | Ours |
| bed | 0.224 | 0.130 | 0.068 | 32.0 | 59.3 | 81.4 |
| birdhouse | 0.192 | 0.125 | 0.075 | 33.8 | 54.2 | 76.2 |
| bookshelf | 0.121 | 0.083 | 0.036 | 43.5 | 66.5 | 86.1 |
| camera | 0.191 | 0.117 | 0.083 | 37.4 | 57.3 | 77.7 |
| file | 0.071 | 0.041 | 0.029 | 65.8 | 86.0 | 93.0 |
| mailbox | 0.146 | 0.060 | 0.040 | 38.1 | 67.8 | 87.6 |
| piano | 0.181 | 0.107 | 0.078 | 39.8 | 61.4 | 82.2 |
| printer | 0.144 | 0.085 | 0.043 | 40.1 | 66.2 | 84.6 |
| stove | 0.104 | 0.049 | 0.030 | 52.9 | 77.3 | 89.2 |
| tower | 0.105 | 0.050 | 0.047 | 45.9 | 70.2 | 85.7 |
| mean | 0.148 | 0.085 | 0.053 | 43.0 | 66.6 | 84.4 |

Table 2. Generalization to unseen classes. Comparison of 3D reconstruction accuracy when 3D autoencoders are tested directly on ShapeNet classes not seen during training. Note that our method (DSIF) has a higher F-Score by 17.8 points.

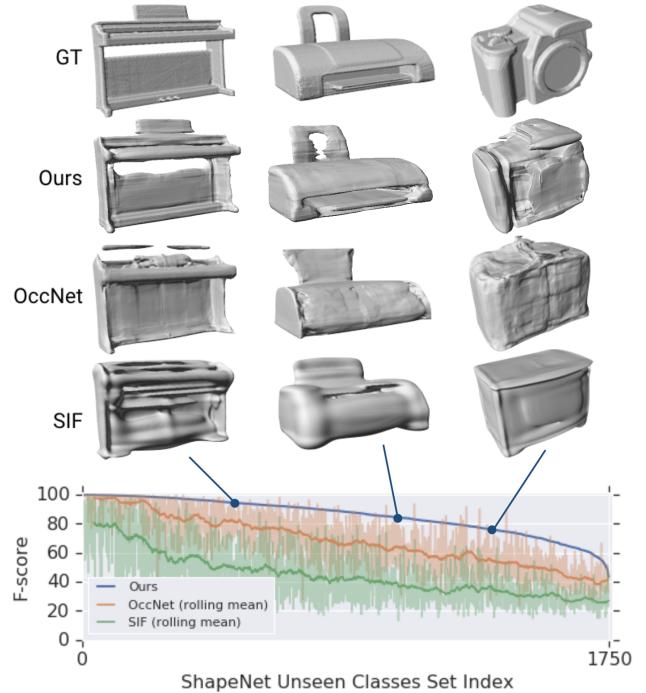


Figure 6. Generalization examples. Example shape reconstructions for piano, printer, and camera classes, which did not appear in the training data. F-score is plotted below ordered by DSIF score, similar to Figure 3. Our method (blue curve) achieves the best accuracy on 91% of the novel shapes.

randomly sampling SMPL parameters (CAESAR fits for shape, mocap sequence fits for pose). We use a 80/5/15 train/val/test split similar to [37], and measured the error of the learned autoencoder on the held-out test set. The challenge for this dataset is quite different than for ShapeNet – the autoencoder must be able to represent large-scale, non-

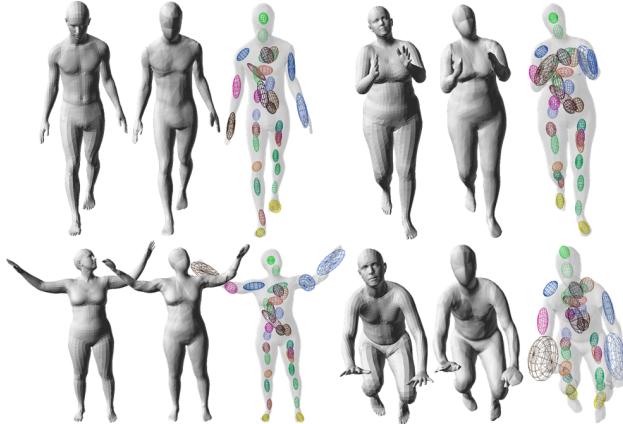


Figure 7. Human body modeling. Surface reconstructions and decompositions for 4 random SMPL [5] human meshes from the SURREAL [37] dataset. For each triple, from left to right: SMPL mesh, our reconstruction, our shape decomposition. These results demonstrate unsupervised correspondence between people in different poses as well as accurate reconstructions of organic shapes.

rigid deformations in addition to shape variations. Our reconstructions achieve 93% mIOU compared to 85% mIOU for SIF. The results of DSIF reconstructions and the underlying SIF templates are shown in Figure 7. Despite a lack of supervision on pose or subject alignment, our approach reconstructs a surface close to the original and establishes coarse correspondences via the structure decomposition.

7. Applications

In this section, we investigate how the proposed DSIF representation can be used in applications. Although SIF (and similarly DSIF) has previously been shown useful for 3D shape analysis applications like structure-aware shape interpolation, surface correspondence, and image segmentation [12], we focus our study here on 3D surface reconstruction from partial observations.

7.1. 3D Completion from a Single Depth Image

Task. Reconstructing a complete 3D surface from a single depth image is an important vision task with applications in AR, robotics, etc. To investigate how DSIF performs on this task, we modified our network to take a single depth image as input (rather than a stack of 20) and trained it from scratch on depth images generated synthetically from random views of the 3D-R2N2 split of shapes. The depth images were 512x512 to approximate the resolution of real depth sensors (though all CNN inputs remain 224x224 due to memory restrictions). The depth images were rendered from view points sampled from all view directions and at variable distances to mimic the variety of scan poses. Each depth image was then converted to a three channel XYZ

| Category | IoU | | Chamfer | | F-Score | |
|------------|---------|------|---------|--------------|---------|-------------|
| | OccNet* | Ours | OccNet* | Ours | OccNet* | Ours |
| airplane | - | 80.2 | 0.047 | 0.017 | 70.1 | 89.2 |
| bench | - | 70.9 | 0.070 | 0.039 | 64.9 | 81.9 |
| cabinet | - | 82.8 | 0.113 | 0.077 | 70.1 | 77.9 |
| car | - | 81.4 | 0.099 | 0.051 | 61.6 | 72.4 |
| chair | - | 70.6 | 0.234 | 0.102 | 50.2 | 69.6 |
| display | - | 82.4 | 0.095 | 0.062 | 62.8 | 80.0 |
| lamp | - | 62.1 | 0.991 | 0.215 | 44.1 | 66.4 |
| rifle | - | 81.5 | 0.049 | 0.014 | 66.4 | 92.3 |
| sofa | - | 81.4 | 0.108 | 0.083 | 61.2 | 71.7 |
| speaker | - | 80.2 | 0.350 | 0.148 | 52.4 | 67.3 |
| table | - | 73.5 | 0.249 | 0.114 | 66.7 | 78.0 |
| telephone | - | 92.3 | 0.035 | 0.019 | 86.1 | 92.0 |
| watercraft | - | 76.0 | 0.115 | 0.050 | 54.5 | 77.5 |
| mean | - | 78.1 | 0.197 | 0.076 | 62.4 | 78.2 |

Table 3. Depth completion accuracy. Our method (DSIF) provides better 3D surface completions than an OccNet* trained on our XYZ image inputs.

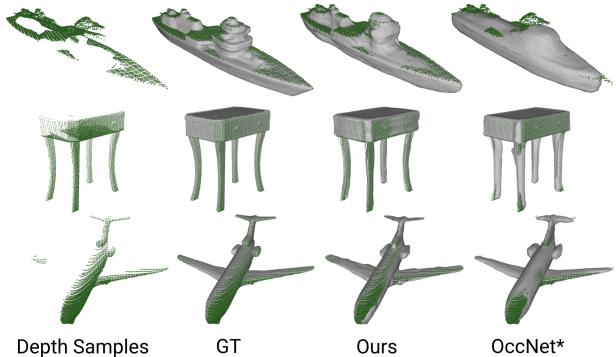


Figure 8. Depth completion examples. Visualizations of surfaces predicted from posed depth images (depicted by green points). Our method provides better details in both the observed and unobserved parts of the shape.

image using the known camera parameters.

Baseline. For comparison, we trained an OccNet network from scratch on the same data. Because the OccNet takes a point cloud rather than depth images, we train an XYZ image encoder network to regress the 256-D OccNet embedding. This OccNet* model provides an apples-to-apples baseline that isolates differences due only to the representation decoding part of the pipeline.

Results. Table 3 shows results of this 3D depth completion experiment. We find that the F-Score of DSIF is 15.8 points higher than OccNet* (78.2 vs. 62.4). Figure 8 highlights the difference in the methods qualitatively. As in the 3D case, we observe that DSIF’s local part encoders result in substantially better performance on hard examples.

Ablation study. To further understand the behavior of DSIF during depth completion, we ablate three components of our pipeline (Table 4). First, we verify that having local point-

| Method | IoU | Chamfer | F-Score |
|--------------|------|---------|---------|
| Full (D) | 77.2 | 0.078 | 77.6 |
| No PointNet | 69.1 | 0.098 | 66.2 |
| No Transform | 71.9 | 0.180 | 71.9 |
| No Symmetry | 76.7 | 0.076 | 76.6 |

Table 4. **Depth completion ablation study.** Local PointNet encoders, camera transformations, and partial symmetry all improve performance. Independently and locally encoding the \mathbf{z}_i with PointNet is particularly good for generalization (see Section 6).

nets to encode the local feature vectors is useful, rather than simply predicting them directly from the input image. Second, we show that providing an XYZ image as input to the network is much more robust than providing a depth image. Finally, we show that taking advantage of the explicit structure via partial symmetry improves results qualitatively and achieves the same quality with fewer degrees of freedom. The biggest of these differences is due to the PointNet encoding of local shape elements, which reduces the F-Score by 11.4 points if it is disabled.

7.2. Reconstruction of Partial Human Body Scans

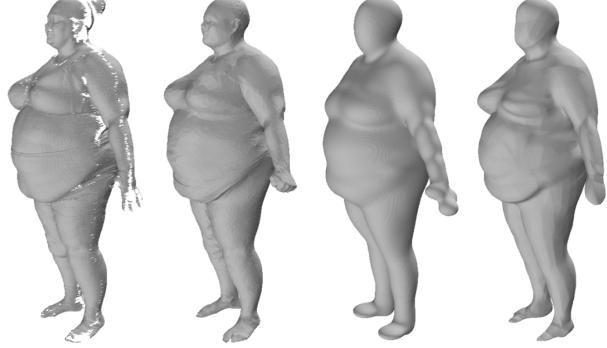
Task. Acquisition of complete 3D surface scans for a diverse collection of human body shapes has numerous applications [1]. Unfortunately, many real world body scans have holes (Figure 9(a)), due to noise and occlusions in the scanning process. We address the task of learning to complete and beautify the partial 3D surfaces without any supervision or even a domain-specific template.

Dataset and baselines. The dataset for this experiment is CAESAR [31]. We use our proposed 3D autoencoder to learn to reconstruct a DSIF for every scan in the CAESAR dataset, and then we extract watertight surface from the DSIFs (using the splits from [26]). For comparisons, we do the same for SIF (another unsupervised method) and a non-rigid deformation fit of the S-SCAPE template [26].

Results. Figure 9 shows representative results. Note that DSIF captures high-frequency details missing in SIF reconstructions. Although the approach based on S-SCAPE provides better results, it requires a template designed specifically for human bodies as well as manual supervision (landmarks and bootstrapping), whereas DSIF is domain-independent and unsupervised. These results suggest that DSIF could be used for 3D reconstruction of other scan datasets where templates are not available.

8. Conclusion

Summary of research contributions: In this paper, we propose Deep Structured Implicit Functions (DSIF), a new 3D representation that describes a shape implicitly as the



a) Scan [31] b) Template fit [26] c) SIF [12] d) Ours
Figure 9. Surface reconstruction from partial human scans.

sum of local 3D functions, each evaluated as the product of a Gaussian and a residual function predicted with a deep network. We describe a method for inferring a DSIF from a 3D surface or posed depth image by first predicting a structured decomposition into shape elements, encoding 3D points within each shape element using PointNet [27], and decoding them with a residual TinyOccNet [21]. This approach provides an end-to-end framework for encoding shapes in local regions arranged in a global structure.

We show that this DSIF representation improves both reconstruction accuracy and generalization behavior over previous work – its F-Score results are better than the state-of-the-art [21] by 10.3 points for 3D autoencoding of test models from trained classes and by 17.8 points for unseen classes. We show that it dramatically reduces network parameter count – its local decoder requires approximately 0.4% of the parameters used by [21]. We show that it can be used to complete posed depth images – its depth completion results are 15.8 percentage points higher than [21]. Finally, we show that it can be used without change to reconstruct complete 3D surfaces of human bodies from partial scans.

Limitations and future work: Though the results are encouraging, there are limitations that require further investigation. First, we decompose space into a flat set of local regions – it would be better to consider a multiresolution hierarchy. Second, we leverage known camera poses when reconstructing shapes from depth images – it would be better to estimate them. Third, we estimate a constant number of local regions – it would be better to derive a variable number dynamically during inference (e.g., with an LSTM). Finally, we just scratch the surface of how structured and implicit representations can be combined – this is an interesting topic for future research.

Acknowledgements: We thank Daniel Vlasic for invaluable discussions and help generating CAESAR data. We also thank Boyang Deng for sharing OccNet* training code, Max Jiang for creating single-view depth renders, and Fangyin Wei and JP Lewis for feedback on the manuscript.

References

- [1] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22, pages 587–594. ACM, 2003. 8
- [2] Bruce G Baumgart. A polyhedron representation for computer vision. In *Proceedings of the May 19-22, 1975, national computer conference and exposition*, pages 589–596. ACM, 1975. 2
- [3] James F Blinn. A generalization of algebraic surface drawing. *ACM transactions on graphics (TOG)*, 1(3):235–256, 1982. 2
- [4] Jules Bloomenthal and Ken Shoemake. Convolution surfaces. *ACM SIGGRAPH Computer Graphics*, 25(4):251–256, 1991. 2
- [5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016, Lecture Notes in Computer Science*. Springer International Publishing, Oct. 2016. 7
- [6] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016. 2
- [7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4
- [8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *CoRR*, abs/1812.02822, 2018. 1, 2
- [9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2, 4
- [10] 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. 2
- [11] James D Foley, Foley Dan Van, Andries Van Dam, Steven K Feiner, John F Hughes, J Hughes, and Edward Angel. *Computer graphics: principles and practice*, volume 12110. Addison-Wesley Professional, 1996. 2
- [12] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. *arXiv preprint arXiv:1904.06447*, 2019. 2, 3, 4, 6, 7, 8
- [13] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. *arXiv preprint arXiv:1906.02739*, 2019. 2
- [14] 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*, pages 216–224, 2018. 2
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. 4, 11
- [16] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2
- [17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, pages 61–70. Eurographics Association, 2006. 2
- [18] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):52, 2017. 2
- [19] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2
- [20] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’87*, pages 163–169, New York, NY, USA, 1987. ACM. 4
- [21] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *CoRR*, abs/1812.03828, 2018. 1, 2, 3, 4, 8, 11
- [22] Shigeru Muraki. Volumetric shape description of range data using blobby model. *ACM SIGGRAPH computer graphics*, 25(4):227–235, 1991. 2
- [23] Yutaka Ohtake, Alexander Belyaev, Marc Alexa, Greg Turk, and Hans-Peter Seidel. *Multi-level partition of unity implicits*, volume 22. ACM, 2003. 2
- [24] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 1, 2
- [25] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2
- [26] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 2017. 8
- [27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 2, 4, 8, 11
- [28] Antonio Ricci. A constructive geometry for computer graphics. *The Computer Journal*, 16(2):157–160, 1973. 2

- [29] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017. 2
- [30] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *2017 International Conference on 3D Vision (3DV)*, pages 57–66. IEEE, 2017. 2
- [31] Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002. 8
- [32] Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. Csgnet: Neural shape parser for constructive solid geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5523, 2018. 2
- [33] Edward J Smith, Scott Fujimoto, Adriana Romero, and David Meger. Geometrics: Exploiting geometric structure for graph-encoded objects. *arXiv preprint arXiv:1901.11461*, 2019. 2
- [34] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2088–2096, 2017. 2
- [35] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 4
- [36] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2643, 2017. 2
- [37] Gü̈l Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6, 7
- [38] 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. 2
- [39] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pages 82–90, 2016. 2
- [40] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Liguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2
- [41] Geoff Wyvill, Craig McPheeers, and Brian Wyvill. Data structure for soft objects. In *Advanced Computer Graphics*, pages 113–128. Springer, 1986. 2
- [42] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *arXiv preprint arXiv:1905.10711*, 2019. 1, 2
- [43] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. *arXiv preprint arXiv:1906.12320*, 2019. 2

A. Hyperparameters

Table 5 contains all hyperparameter values used for training the model. Architecture details for individual networks are below.

ResNet50. We use a ResNet50 [15] V2 that is trained from scratch. The 20 depth images are concatenated channel-wise prior to encoding.

PointNet. We modify the original PointNet [27] architecture by removing the 64×64 orthogonal transformation to improve speed and reduce memory requirements.

OccNet. Our TinyOccNet variant follows the same structure as the original OccNet [21]. However, we reduce the number of residual blocks from 5 to 1. The latent layer feature widths are also decreased proportionally to the vector dimensionality.

Local Point Cloud Extraction. We sample a subset of points for encoding by the local PointNet as follows. We first transform all 10,000 points to the local frame. Then we choose a distance threshold $r = 4.0$ measured in local units. Since the local frame is scaled proportionally to the radius, this threshold is approximately four radii in the world frame. We randomly sample 1,000 points without replacement within r and return those as the set of points to be encoded. If 1,000 points do not exist, we expand r until 1,000 total points are found.

Global Point Cloud Creation. In order to create 10,000 points from one or more input depth images, we randomly sample valid points without replacement from the depth images. If 10,000 valid pixels do not exist, we repeat random points as necessary before moving to the local extraction phase.

Activation Functions. Since the generated network activations y are in the range $[-\infty, \infty]$, we apply activation functions to latents y to interpret them as the analytic parameters θ_i . The following functions are used. For constants c_i : $-|y_{c,i}|$. For ellipsoid radii r_i : $0.15 \times \text{sig}(y_{r,i})$. For ellipsoid euler-angles e_i : $\max(\min(\frac{\pi}{4}, y_{e,i}), -\frac{\pi}{4})$. For ellipsoid positions p_i : $\frac{y_{p,i}}{2}$.

| Name | Value |
|------------------|--------------------|
| α | 100.0 |
| w_S | 0.1 |
| w_U | 1.0 |
| w_C | 10.0 |
| w_P | 1.0 |
| Batch Size | 24 |
| Adam β_1 | 0.9 |
| Adam β_2 | 0.999 |
| Learning Rate | 5×10^{-5} |
| Surface Isolevel | -0.07 |

Table 5. Hyperparameters and optimization details for training the autoencoder network.