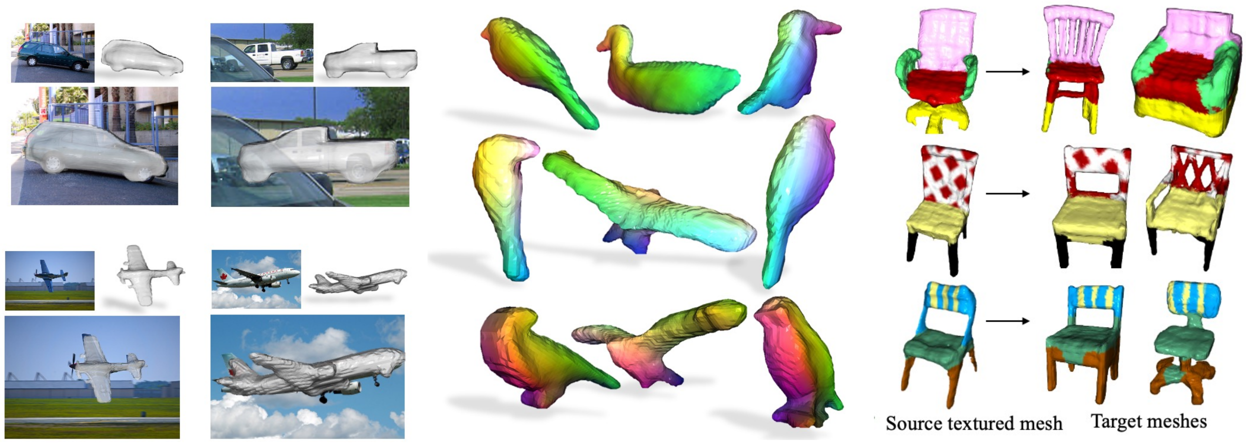
Topologically-Aware Deformation Fields for Single-View 3D Reconstruction

Shivam Duggal   Deepak Pathak
Carnegie Mellon University



Single-view 3D Reconstruction   ***Unsupervised*** Correspondence Estimation   ***Unsupervised*** Texture Transfer

Figure 1. Given an unpaired image collection (with known camera poses) of an object category at training time, our approach learns to: (a) *reconstruct the underlying 3D* given only a single image at test time, and (b) *model dense 3D correspondences* across category instances. The learned correspondence field is articulation-aware, topologically-aware and inherently captures structural properties of the category, enabling the task of *unsupervised texture transfer*. Videos and code at https://shivamduggal4.github.io/tars-3D/

Abstract

We present a new framework for learning 3D object shapes and dense cross-object 3D correspondences from just an unaligned category-specific image collection. The 3D shapes are generated implicitly as deformations to a category-specific signed distance field and are learned in an unsupervised manner solely from unaligned image collections without any 3D supervision. Generally, image collections on the internet contain several intra-category geometric and topological variations, for example, different chairs can have different topologies, which makes the task of joint shape and correspondence estimation much more challenging. Because of this, prior works either focus on learning each 3D object shape individually without modeling cross-instance correspondences or perform joint shape and correspondence estimation on categories with minimal intra-category topological variations. We overcome these restrictions by learning a topologically-aware implicit deformation field that maps a 3D point in the object space to a higher dimensional point in the category-specific canonical space. At inference time, given a single image, we reconstruct the underlying 3D shape by first implicitly deforming each 3D point in the object space to the learned category-specific canonical space using the topologically-aware deformation field and then reconstructing the 3D shape as a canonical signed distance field. Both canonical shape and deformation field are learned end-to-end in an inverse-graphics fashion using a learned recurrent ray marcher (SRN) as a differentiable rendering module. Our approach, dubbed TARS, achieves state-of-the-art reconstruction fidelity on several datasets: ShapeNet, Pascal3D+, CUB, and Pix3D chairs.

1. Introduction

Learning to understand the 3D geometric world underlying our 2D observation snapshots has been a longstanding problem in computer vision, yet the generalization is nowhere close to that in learning to recognize 2D concepts [15, 22, 23]. The reason is rather unsurprising: the lack of scalable ways to obtain 3D supervision in the wild, be it multiple views of the same object or GT shape. Unlike the current visual systems, humans can infer 3D structure just from a single image (even under large occlusions). If our (deep) learning models have to develop such a capability, we must first figure out how to understand the 3D structures from just an unaligned and diverse 2D image collection – the kind of data available *in abundance* on the web. However, any such approach must answer a fundamental question first – how should one represent the 3D structure?

Looking at the research in recent years, there is an overwhelming evidence in support of implicit representations credited to the advancement in neural implicit modeling [9, 36, 43, 44, 47, 53, 56]. While these implicit representations have attained the gold standards of high-fidelity

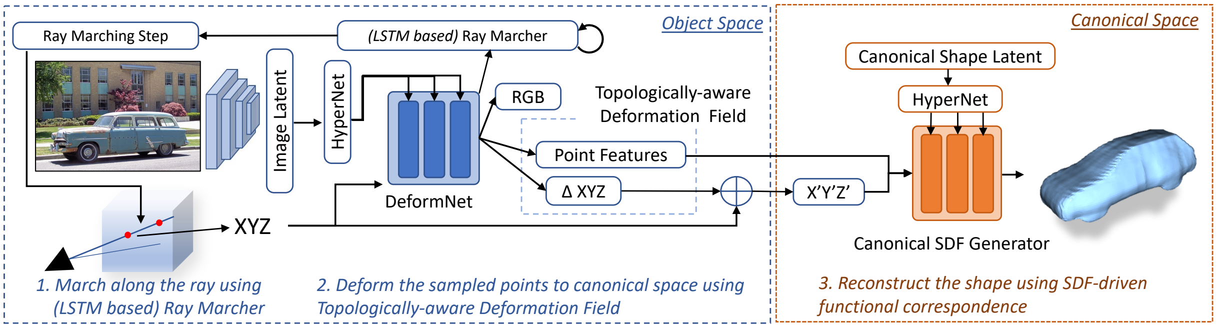


Figure 2. Overview of TARS: Given a single image, we first map a 3D point in object space to a higher-dimensional canonical space using our learned topologically-aware deformation field. The canonical point is then mapped to its SDF value using the Canonical Shape Generator module. We leverage an LSTM-based differentiable renderer to guide the learning of deformation and signed distance fields.

reconstruction, they still rely on either 3D GT shape or dense multi-view supervision not only during training but sometimes also at inference [44], making them difficult to apply to the internet of images. Recent works [10, 26, 79, 82] have attempted to cut down the requirement of multi-view images from 100s to $2-10$. However, as long as any method needs more than a single image, it can not be used to 3D-fy trillions of images on internet – the setting considered in this paper. What kind of signals can we exploit from 2D image collection of a category at training time, that can help generate 3D for an unseen 2D image at test time? We turn to Plato.

Plato's philosophy of "Theory of Forms" relates every object in reality to a particular form or an idea (a platonic ideal). His famous example of "cupness" says that while there exists many cups, there is only one "idea" of cupness. We believe this is closely tied to human perception of objects. For instance, when we play the game of pictionary [1], given just a category-level description of an object, we can generally draw its high-level (category-level) representation. Only when we are provided with more observations or properties of an object (eg: a chair "with arms", "an SUV" car, an airplane "with wider wings"), we are able to draw that specific instance of the category. This philosophy has been classically adopted in deformable models [6] but require 3D supervision. More recently, with the advent of differentiable renderers [8, 29, 35], this has been adopted for estimating 3D from a single image [5, 19, 27, 32]. However, these methods are restricted to categories with minimal to no intra-category topological variations because of fixed mesh connectivity and absolute reconstructions are also of lower fidelity compared to implicit methods (see CMR results in Figure 6).

3D objects that belong to the same ("platonic") category generally inherit similar structural and semantic properties. In this work, we follow this ideology and propose a 3D reconstruction algorithm, which can: (a) learn from just an unaligned 2D image collection without any 3D or multi-view supervision at training and inference; (b) generalize to topologically diverse categories like chairs which mesh-based approaches can't; and (c) can learn dense 3D correspondence across different instance shapes for free by mapping the object instances to the category mean, allowing the model to exploit cross-image similarity. These intra-category correspondences are very beneficial for numerous vision and graphics tasks: geometry/shape understanding [3, 39, 70, 84], 3D manipulation [6, 39, 84], 2D image synthesis [10, 65, 74], 2.5D depth estimation [42, 57, 81], etc.

However, simply extending implicit models and learning implicit dense correspondences between topologically varying objects with just single view supervision is not straightforward. This is because of inherent continuous nature of MLPs used by implicit shape modeling techniques and the inherent discontinuities in correspondence field between any two topologically different objects. For any two instances with different topologies, correspondence field has to be dis-continuous in order to map one structure to the another. Please refer to supp. section B for more understanding. To overcome this issue of implicitly learned deformation fields, we propose *topologically-aware deformation fields*.

Given an object image, we first map a 3D point in the object space to the corresponding 3D-point in the category-level canonical space using our *DeformNet module*. Then, to address the above issue of implicit deformation fields and to learn correspondences between topologically varying shapes, we take inspiration from Level Set Method (LSM) [50, 51]. Level Set Methods support topological merging/breaking of shapes by representing any shape as a zero-level crossing of a higher-dimensional function. Inspired from them, *we transform our 3D canonical points to a higher-dimension by concatenating them with learned object-space point features.* We then estimate the underlying shape by mapping the higher-dimensional canonical points to the corresponding SDF values using the *Canonical Shape Generator* module. A high-level overview of our approach is shown in Figure 2.

We dub our approach **TARS** (**T**opologically-**A**ware **R**econstruction from **S**ingle-view), see overview in Figure 2. We utilize a differentiable renderer in our pipeline to learn both the deformation and the shape reconstruction modules using image collections containing single-view RGB obser-

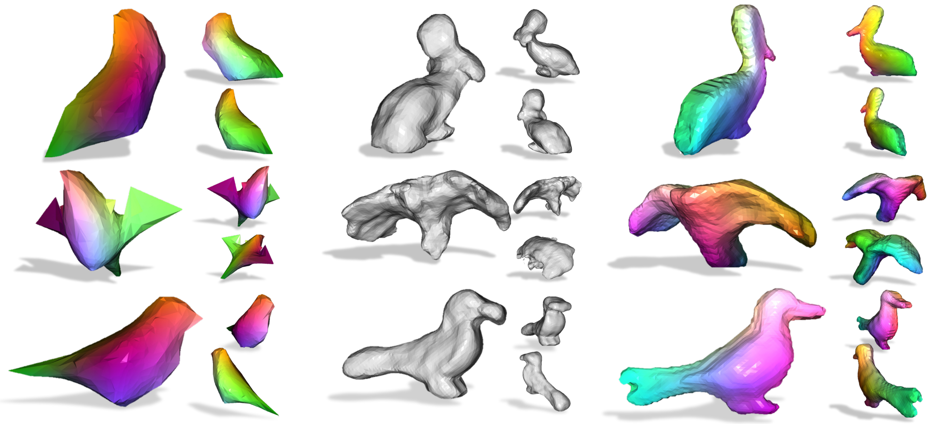


Figure 3. 3D Reconstruction on CUB-200-2011. Compared to prior works, not only our reconstructions are of higher fidelity, but the learned (color-coded) deformation fields are also articulation aware (eg: rotated heads, open wings). Unlike CMR, we do not hard-code symmetry.

vation, corresponding GT camera pose and object silhouette. Our differentiable render (inspired from SRN [60]) is a form of a neural render [64] which takes in features of a 3D point in the object space (visible from the input viewpoint) and predicts its corresponding depth value as seen from the input view point. Thus, during training we learn the object shape in two ways: (a) 2.5D depth representation learned using object-level features (via differentiable renderer), (b) 3D SDF learned using canonical shape features (via canonical shape generator). By enforcing the consistency between the two shape representations, we are able to effectively learn the correspondence field. Since this shape consistency is the courtesy of the differentiable renderer, we term it as the *differentiable render consistency* in the following sections.

The closest approach to ours is SDF-SRN [33], a neural implicit shape modeling approach for single-view 3D reconstruction. Unlike us, they directly map a point in the object space to the corresponding SDF value and hence do not output dense correspondences across object instances.

We validate the effectiveness of our learned shapes on multiple datasets: ShapeNet [7], Pascal3D+ [77], CUB-200-2011 [73] and Pix3D chairs [61]. Our method, TARS, outperforms priors works in term of 3D reconstruction fidelity and generates shapes with better global structure and finer instance-specific details. Unlike prior deformable single-view reconstruction works [19, 27, 32], which are restricted to categories like cars/ cubs, we take the first major step in modeling topologically-challenging categories (chairs). The learned topologically-aware deformation field captures structural properties of the category (without any supervision), thus enabling unsupervised texture-transfer (Figure 1).

2. Related Work

Reconstructing 3D from 2D observations has been an actively studied problem [9, 30, 43, 45, 53, 57]. Until recently, the majority of the high-fidelity reconstruction results were credited to the availability of some form of 3D data [25, 45], and because of this, the majority of the success had been restricted to synthetic datasets [7]. Reconstruction of real-world 3D shapes was either done by transferring the learned synthetic models to real-world objects [4, 16, 75, 76] or required special 3D sensors [16, 45, 80]. However, collecting dense 3D data is cumbersome, challenging, and even not possible for certain categories (eg: birds). With advancements in inverse graphics and differentiable rendering [8, 29, 35, 37, 38, 68], the requirement of 3D supervision has been significantly reduced. More recently, significant progress has been made in this direction and the reconstruction quality has reached its golden standard, particularly thanks to the combination of neural implicit representations and differentiable rendering [44, 48, 60, 78]. However, the majority of these works still require dense supervision in form of paired multi-view images. Such a setting may not be possible for the internet of images. Our work, TARS, focuses on further mitigating the dependency on dense supervision by operating only on single-view data.

3D Reconstruction with Single-view Supervision: The task of single-view 3D reconstruction has been comparatively less-explored. Kar *et al*. [28], Kanazawa *et al*. [27] took a major step in this direction by learning 3D structures from a large collection of unpaired images. They learned to reconstruct the underlying shape by learning the deformations on top of a (learned) category-specific mean mesh. Further research along this direction focused on reducing supervision [19, 32], enhancing geometry [8, 67] and texture fidelity [5]. However, these works are restricted to the reconstruction of object categories with topologically similar instances (eg: birds). Like CMR [27], we leverage the structural knowledge embedded in image collection in form of learned deformations to a learned category-specific mean shape but overcome their topological restrictions. Recently, Lin *et al*. [33] directed the success of neural implicit mod-

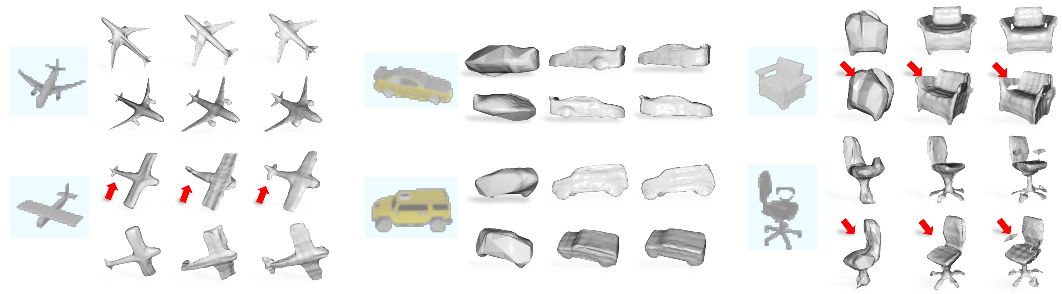


Figure 4. 3D Reconstruction on Shapenet. Compared to mesh-based SoftRas, both neural implicit approaches yield higher 3D fidelity. Our approach additionally provides correspondences (even across topologically-varying structures), while matching SDF-SRN's shape fidelity.

eling [60] to the task of single-view 3D reconstruction and achieved state of the art results in terms of geometric fidelity. Our work further boosts the fidelity standards by jointly learning category-specific deformations and SDF fields.

Neural Rendering: Recent works [36,44,48,49,60,71,78] for rendering implicit surfaces have majorly leveraged some form of ray-tracing (ray marching, volumetric or surface rendering). The recent survey on neural rendering [63] classifies the rendering as: (a) *image-based rendering approaches*, which generate 2D content without explicitly modeling 3D (by transforming/ warping the input images) (b) *explicit 3D based approaches*. In our work, we utilize SRN [60] as our neural renderer. SRN [60] performs LSTM-based ray marching to implicitly generate a 2.5D depth map corresponding to the input image. Therefore, SRN lies at the intersection of image-based and explicit rendering approaches. By using SRN [60] as image-based renderer, we learn shapes in two ways: image to depth map translation learned using object-space features, and image to SDF learned using canonical-space features. Consistency between the two shape representations is the key contributor to our performance.

3D Reconstruction with Dense Correspondences: Learning category-specific deformable shapes have been found to be prominently useful for 3D reconstruction [6,17,27,28,39]. These approaches generally learn instance shapes as deformations to the initial shape bases. Prior works along this line (reconstruction via deformation) have reconstructed 3D shapes either in volumetric grid representation [17, 72] or mesh representation [27, 28, 40]. We learn both the deformation field and the 3D shape (signed distance field) implicitly. Unlike deformations to mesh, learning deformations to an implicit field is much more challenging, because of the loss of explicit structure (mesh connectivity). Recently, [13, 83] learned category specific deformation and signed distance fields implicitly. However, unlike our approach (TARS), they require dense 3D supervision during training.

3. Method

Given a single image of an object, our goal is to reconstruct the underlying 3D shape. Rather than directly reconstructing the shape, we learn to reconstruct the object's 3D shape by implicitly mapping it to a (learned) category-specific canonical shape. In order to do so, we leverage a category-specific collection of unpaired object images (along with camera poses and object silhouettes) as our training corpus. This allows us to incorporate category-specific knowledge into our shape reconstruction pipeline. Our pipeline (as shown in Figure 2) consists of three core components: (a) *Deformnet*, for prediction of topologically-aware deformation fields, (b) *Canonical Shape Generator*, for reconstruction of object's 3D shape (as SDF) and, (c) *Differentiable Renderer module*, to render the learned SDF and hence guide the learning of Deformnet and Canonical Shape Generator during the training phase. In the following sections, we first discuss these modules and then stick them together to define our inference and training regimes.

3.1. Topologically-Aware Deformation Fields

Learning Implicit Deformation Fields: The goal of DeformNet (g) is to learn dense 3D point deformations from object-space to canonical-space. More formally, given an image I, and a 3D point (x_{object}) in object space, the deformation estimation task is defined as:

$$x_{\text{object}} + g(x_{\text{object}}, I) = x_{\text{canonical:3D}} \tag{1}$$

where $x_{\text{canonical:3D}}$ is the corresponding point in the canonical space. The mapping between the two points (x_{object} and $x_{\text{canonical:3D}}$) is learned by leveraging signed distance function (SDF) as the functional map [52] between the two spaces i.e. SDF of x_{object} w.r.t object's surface should be same as SDF of $x_{\text{canonical:3D}}$ w.r.t canonical shape surface.

We implement DeformNet module as an MLP. To learn the deformation field, we condition the DeformNet module on the input image through a hyper-network. The input image is first passed through ImageNet pre-trained ResNet encoder [23] to generate a latent-code. Inspired from [33, 58–60], the computed latent code is then used by the hyper-network to predict the weights for the DeformNet MLP. We observe that using the hyper-network rather than directly learning the weights of the MLP leads to smoother shapes.

Figure 5. 3D Reconstruction on Pascal3D+ planes. Compared to prior works, our approach performs well even with the challenging real-world observations, generating 3D shapes which are less noisy and better represent the overall structure of the ground-truth shapes.

Point-features for Learning Topologically-Aware Deformation Fields: Unlike prior works [5, 19, 27, 32], our goal is to reconstruct 3D shapes even for object categories with large intra-category topological variations (eg: see chairs in Figure 1, Figure 6 and Figure 7). In order to do so, we need to ensure that our deformation field can map any input object with an arbitrary topology to the canonical shape with a fixed topology. However, learning such a deformation field using an MLP is a challenging task. This is because of the continuous nature of the MLP. While the continuous nature of MLP assists in learning the 3D shapes implicitly, such a property hurts the learning of cross-object deformations. This is because the deformation field between objects of different topologies could be discontinuous (supp. Figure 9). To overcome this issue and effectively learn both the deformation and the shape fields, we take inspiration from the level-set methods (LSM) theory. LSM [50] allow topological merging and breaking of structures by modeling any surface as a zero-level crossing of a higher-dimensional function. We take inspiration from these works [14, 24, 50, 51], and learn a higher-dimension deformation field (7D in our implementation) instead of previously learned 3D deformation fields. Concurrently, Park *et al*. [55] proposed similar insights for learning deformations between multiple views of the same object instance. To learn the higher-dimensional deformation field, we also learn object-space point features, $h(x_{\text{object}})$ using the intermediate-level features of DeformNet, alongside learning the above-defined 3D deformation field (Eq. 1). Thus, we deform a point ($x_{\text{object}} \in \mathbb{R}^3$) in object space to a higher-dimensional canonical point ($x_{\text{canonical:HD}} \in \mathbb{R}^{3+k}$) ($k$ equals the dimension of the learned point features), where $x_{\text{canonical:HD}}$ is simply the concatenation of 3D canonical point ($x_{\text{canonical:3D}}$) and learned point features, $h(x_{\text{object}})$. We notice that learning these point features leads to reconstructions with sharper details and better preservation of topology of GT shape (see Figure 8).

We also predict view-independent RGB value of the input 3D point using intermediate-level features of DeformNet.

3.2. Canonical Shape Reconstruction

Now that we have deformed the 3D points in object space to the corresponding points in canonical space, our next task is to learn the 3D shape in form of SDF field. To estimate the SDF value of the 3D object point (x_{object}), we pass the corresponding higher-dimensional canonical point ($x_{\text{canonical:HD}}$) through the Canonical Shape Generator module (f). We learn the weights of shape generator using a hyper-network. The hyper-network is conditioned on a canonical shape latent-code (L), which is jointly learned during training. The canonical reconstruction task is defined:

$$f(x_{\text{canonical:HD}}, L) = s$$

, where s is the signed-distance value of $x_{\text{canonical:3D}}$ w.r.t the canonical shape surface (also equals signed distance value of x_{object} w.r.t input object's surface by the property of the established functional map).

3.3. Differentiable Renderer Consistency

In this section, we define the differentiable renderer and our proposed differentiable renderer consistency term which are used in our training pipeline (Figure 2). The differentiable renderer is used to generate 2D renderings of the learned 3D shape during training, which are then compared against input object's GT 2D observations (RGB map and silhouette). Following [33], we utilize SRN [60] as our LSTM-based differentiable renderer. The renderer works by performing the ray marching procedure, where every marching step is learned in form of a depth estimate from the current 3D point along the current camera ray direction. Please refer to SRN paper [60] for more details.

Prior works on deformation-driven inverse graphics [19, 27, 54, 66] rendered the learned 3D shape (be it mesh, density field or signed distance field) to compute the loss terms for training. Unlike them, instead of rendering the signed-distance field (like [36, 71]) learned by the Canonical Shape Generator, we utilize SRN as an image-based neural renderer. It takes as inputs the intermediate-level object features of the DeformNet module and predicts a 2.5D depth map of the input object (as viewed from input viewpoint). *This allows us to enforce consistency between the two shape representations learned in our training pipeline: (a) 2.5D depth map learned via object-space point features, (b) 3D signed distance value learned via canonical-space point features.* We enforce the signed distance value of the last and the second last 3D points along the (renderer's) marched rays (for rays hitting the object) to be $-$ve and $+$ve respectively. The consistency term has been adopted from SDF-SRN [33]. However, they established this consistency between the two shape representations learned in the same object space. Unlike them, our purpose to utilize such a consistency to allow efficient learning of the deformation field.

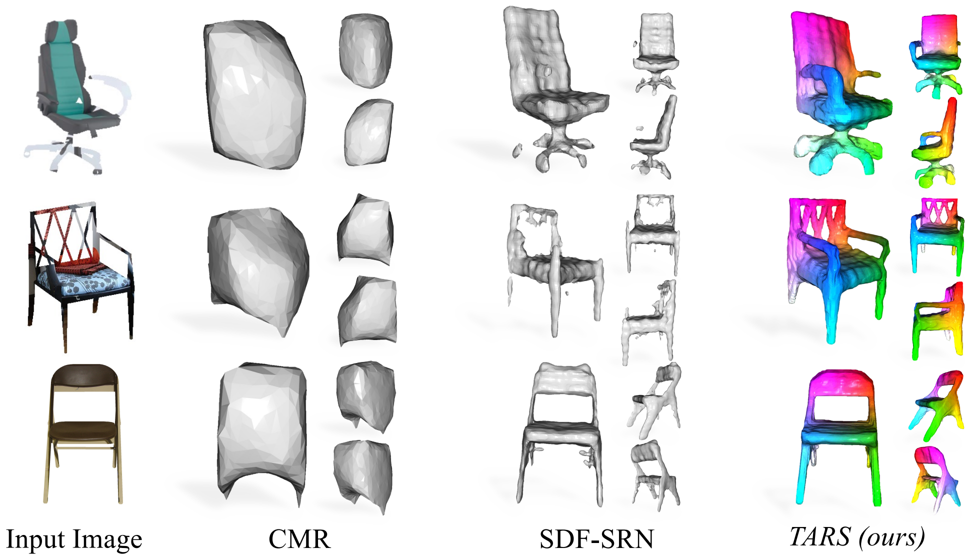


Figure 6. 3D Reconstruction on Pascal3D+ (default) Chairs. Compared to the implicit approaches, CMR completely fails to model the topologically-varying chairs category. (Color denotes mapping to canonical space)

3.4. Inference and Training regimes

Inference: In order to reconstruct the 3D shape underlying the input image, we first densely sample points within a unit-cube and map them from object space to canonical-space using the topologically-aware deformation field. The SDF values of the deformed object points are then estimated using Canonical Shape Generator. Finally, we utilize marching cubes [41] to generate a 3D mesh from the learned SDF field.

Training: Our training procedure is similar to the recent image-based implicit shape modeling and novel-view synthesis works [33, 44, 60]. We begin with shooting variable number of camera rays from the input camera viewpoint. We iteratively march along each camera ray and for each 3D point (x^i_{object}) along the ray, we predict: (a) corresponding canonical point ($x^i_{\text{canonical:HD}}$) using the DeformNet, (b) corresponding SDF value of the canonical point using the shape generator and, (c) the ray marching step (d^i) using the LSTM renderer. The next 3D point along the ray is then estimated as: $x^{i+1}_{\text{object}} = x^i_{\text{object}} + d^i \overrightarrow{r}$ ($\overrightarrow{r}$ is the unit ray direction). The above procedure is repeated n times ($i \in \mathrm{n}$) along all rays (where $\mathrm{n} =$ # of ray marching steps). Our training objective is similar to SDF-SRN's [33] and is defined as:

$$\ell_{\text{total}} = \ell_{\text{rgb}} + \ell_{\text{sdf}} + \ell_{\text{reg}}$$

RGB loss term (ℓ_{rgb}) is simply the mean-squared error between a 3D point's predicted RGB value and the GT pixel intensity of the corresponding rendered pixel.

SDF loss term (ℓ_{sdf}) enforces the proposed differentiable renderer consistency. For camera rays intersecting the 3D object (guided by the GT object silhouette), SDF loss term enforces all points other than the last ray point to have SDF > 0 (outside the object surface) and the last ray point to have SDF < 0 (inside the surface). SDF value is penalized to be greater than 0, for all points on non-intersecting rays. Following SDF-SRN [33], we also utilize the distance transform of the input object mask to penalize the lower-bound of the SDF values of points lying outside the surface. Please check SDF-SRN [33] for more details on the distance-transform loss term.

Regularization terms (ℓ_{reg}): We utilize two regularization terms: Eikonal loss (ℓ_{eik}) and Deformation smoothness (ℓ_{def}). We apply eikonal loss on canonical points ($x_{\text{canonical:3D}}$) and def. smoothness on object-space points.

$$\ell_{\text{eik}} = \sum_{x \in \Omega} ||\nabla f(x + g(x, I)) - 1||_2^2$$

$$\ell_{\text{def}} = \sum_{x \in \Omega} ||\nabla g_x(x) + \nabla g_y(x) + \nabla g_z(x)||_2^2$$

For both the regularization terms, we sample from the unit cube (Ω) bounding a normalized 3D object.

4. Experiments Details

Datasets We train and evaluate our proposed approach as well as the baselines on following datasets: Shapenet [7], Pascal3D+ [77], CUB-200-2011 [73] and Pix3D chairs [61]. Each training example consists of cropped RGB image (centered around the object), corresponding segmentation map and camera pose. At inference time, we only need the object image as input. Please check supp. for more details.

Baselines We compare against the state-of-the-art methods on the task of 3D reconstruction: (a) SoftRas [35]: rasterization-based differentiable mesh renderer. (b) SDF-SRN [33]: neural implicit modeling approach for single-view

Figure 7. 3D Reconstruction on Pix3D (trained on Shapenet).

reconstruction. It is closest to ours but does *not* learn any correspondences across instances. (c) CMR [27]: deformation-driven mesh reconstruction approach which uses NMR [29] as the differentiable renderer and also learns dense correspondences. We achieve state-of-the-art quantitative results (or are at par) for most categories on all datasets while jointly learning dense correspondences. The qualitative comparisons with baselines highlight the efficiency of our approach.

Evaluation Metrics: Correctly and efficiently evaluating the reconstruction quality has been a point of debate [2, 34, 46, 62]. In this work, we evaluate the reconstruction quality by comparing the reconstructed shape with GT using (a) Chamfer distance (b) Earth Mover's Distance (EMD) and (c) Precision, Recall, F-score at 0.1 threshold.

5. Experimental Results

5.1. Qualitative and Quantitative Comparisons

3D reconstruction on CUBS-200-2011: We compare against CMR [27] and SDF-SRN [33] on the CUBS dataset in Figure 3. While SDF-SRN independently reconstructs each 3D object given the input image, CMR reconstructs each instance shape by deforming the category-specific mean mesh. On the other hand, TARS learns to reconstruct 3D instances implicitly by deforming the object space points to the canonical space. Compared to both deformation based reconstruction approaches, SDF-SRN generates noisy shapes *(see noisy wings of the bird in row 2, Figure 3)*. Credited to the implicit nature of our approach, the reconstructed shapes better respect the articulations of the GT objects (see rotated heads in Figure 1, open wings in Figure 3 row 2).

3D reconstruction on Shapenet: Figure 4 and Table 1 showcase qualitative and quantitative comparison of the reconstructed shapes on the Shapenet dataset. As a mesh-based reconstruction algorithm, SoftRas [35] is able to reconstruct cars and planes, but fails on the chairs category, reason being the large intra-category topological variations. It fails to capture the details and only recovers the global shape underlying the input image. Being a neural implicit reconstruction

Cat.	Method	Chamfer ↓			EMD ↓	Precision ↑	Recall	F score ↑
		acc.	cov.	overall		(%)	(%)	(%)
Car	SoftRas [35]	0.372	0.302	0.337	0.723	93.04	96.62	94.80
	SDF-SRN [33]	**0.141**	0.144	0.142	0.452	**99.76**	99.84	**99.80**
	TARS (ours)	**0.141**	**0.140**	**0.140**	**0.446**	99.70	**99.81**	99.75
Chair	SoftRas [35]	0.572	0.475	0.523	1.017	82.56	89.18	85.74
	SDF-SRN [33]	**0.352**	0.315	0.333	0.854	**94.18**	95.21	**94.69**
	TARS (ours)	**0.353**	**0.312**	**0.332**	**0.817**	93.43	**95.39**	94.40
Airplane	SoftRas [35]	0.215	0.207	0.211	0.588	98.74	98.42	98.58
	SDF-SRN [33]	**0.193**	0.154	0.173	0.576	98.55	99.11	98.83
	TARS (ours)	**0.194**	**0.152**	**0.173**	**0.533**	**98.79**	**99.34**	**99.06**

Table 1. 3D reconstruction results on ShapeNet. Compared to mesh based SoftRas algorithm, both the implicit approaches: SDF-SRN and our approach perform significantly better on all metrics.

approach, SDF-SRN [33] captures both the global structure and the fine details. TARS matches the reconstructed shape fidelity of the SDF-SRN reconstructions, both quntitatively and qualitatively *(see the tail of the airplane, arms of the revolving chair in Figure 4)*, and also learns cross-instance structural correspondences for free. Thanks to the proposed higher-dimensional deformation field, our reconstructions respect the topology of the GT shape *(both the arms of the couch in Figure 4 have holes in them)*.

3D reconstruction on Pascal3D+: The default Pascal3D+ dataset provides 2D-3D paired data by associating PASCAL VOC [18] and Imagenet [12] images with the closest matching CAD model. Since, the same set of CAD models are used for both training and test set objects, generating object silhouettes (used both during training and inference) by rendering the 3D CAD models creates a bias between the train and the test sets. Thus, generalization results of prior reconstruction methods [11, 33] on the Pascal3D+ dataset should be taken with a grain of salt. Unlike prior works [33], we demonstrate qualitative comparison on both the default biased Pascal3D+ dataset and an unbiased version of the same dataset. The main purpose to showcase results on both the default and the unbiased datasets is to dis-entangle the inherent limitations of prior works from the lack of generalization. Please refer to supp. dataset section for more details. We compare against CMR [27] and SDF-SRN [33] on three categories of Pascal3D+ dataset (cars, planes, and chairs). As shown in Figure 6, CMR [27] suffers significantly on the chairs category and even fails to capture the global shape, the reason being their mesh representation which doesn't allow breaking the initial mesh connectivity (/ topology). Even on the planes category of both the default dataset (supp. Figure 21) and the unbiased dataset (Figure 5, supp. Figure 13, supp. Figure 14), CMR fails to capture the details and rather generates self-intersecting and similar-looking meshes for different plane instances. SDF-SRN [33] does capture the overall shape details well. However, because of the challenging nature of the real-world images, compared to its performance on Shapenet dataset [7], it under-performs and generate much noisier reconstructions on the real-world Pascal dataset *(see noise on the reconstructed planes in Figure 5, ripples on the reconstructed SDF-SRN cars in supp. Figure 20,* noisy *reconstructed sofa in Figure 6)*. Further, it

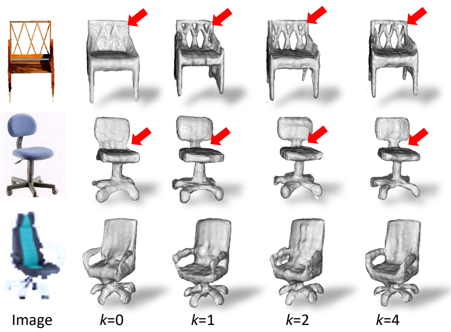


Figure 8. Ablation of dimensionality (k+3) of deformation field on Pascal3D+ (default) chairs. k equals the dimensionality of the additional point features.

fails to maintain the topological details of the GT 3D shapes *(eg: lack of details on the back of the chair in row 3, missing arms of chairs in row 2, 3 of Figure 6)*. In comparison, as can be seen from the results on both the default dataset (Figure 6, supp. Figure 20, 21, 22) and the unbiased dataset (supp. Figure 13, 14), our reconstructions are (a) much less noisier, (b) respect the topology of the underlying shapes, and (c) better captures both the global shape and the finer details.

3D reconstruction on Pix3D Chairs: To showcase the generalization capability of our proposed approach, we demonstrate qualitative comparisons on the Pix3D chairs dataset. Figure 7 compares shapenet-trained SDF-SRN [33] and our approach on the Pix3D test split. Despite the challenging nature of Pix3D dataset (diverse 3D shapes, variable texture, material and environment conditions), both the approaches generalizes well from the synthetic Shapenet dataset to the real-world Pix3D dataset. Further, thanks to the topologically-aware deformation field, our approach maintains the topological structure of the GT shapes (see reconstructed holes in chairs in row 2, 3 of Figure 7). Please refer to supp. for comparison of reconstruction approaches trained only on real-world chairs (Pix3D train + Pascal3D).

Learned deformation field: We visualize our learned deformation fields in Figure 1, 3 and supp. Figure 10. The color codes denote the corresponding canonical 3D points obtained by mapping the 3D object space points to the canonical space using the learned deformation fields. Our deformation field consistently learns to deform similar object parts to similar regions of the learned mean shape *without any form of part supervision* (eg: legs of all chairs in supp. Figure 10 are consistently painted similarly with yellow, green, blue and pinkish-white). *Similar deformation consistency is observed in the cubs category, despite the structural and non-rigid articulation dependent variations (see Figure 3).* This validates that deformation-based approaches can inherently learn category-specific structural relations *(without any supervision)* leveraging just single-view image collections.

Category	Method	Chamfer ↓ acc.	Chamfer ↓ cov.	Chamfer ↓ overall	EMD ↓	Precision ↑ (%)	Recall (%)	F score ↑ (%)
Car	Ours (w/o point features)	0.379	0.473	0.426	0.949	96.97	91.77	94.30
	Ours	**0.363**	**0.386**	**0.374**	**0.763**	**97.00**	**95.63**	**96.31**
Chair	Ours (w/o point features)	0.539	0.485	0.512	1.291	86.95	91.44	89.14
	Ours	**0.527**	**0.426**	**0.476**	**1.171**	**89.75**	**94.83**	**92.22**
Airplane	Ours (w/o point features)	0.576	0.561	0.568	1.341	87.77	93.16	90.38
	Ours	**0.547**	**0.530**	**0.538**	**1.302**	**88.50**	**93.76**	**91.05**

Table 2. Point features ablation on Pascal 3D+ chairs dataset.

Leveraging deformation fields for texture transfer: We showcase the utility of the learned deformation field for the task of texture transfer in Figure 1. We first manually paint a 3D mesh and then transfer the painted texture to other meshes using the learned deformation field of the two meshes. As can be seen in the figure, structurally similar parts of both the source and the target meshes are painted similarly. The checkered stripe patterns and the parallel stripe patterns of the source meshes of row 2 and 3 respectively, are maintained in the target meshes, highlighting the structural details captured by the learned deformation fields.

5.2. Ablation Study

We ablate the efficiency of the learned point features on Pascal3D+ categories in Table 2 and Figure 8. Despite the bias in Pascal3D+ default dataset, such an ablation is useful as it helps understand the inherent limitation of the implicit deformation approaches, by ruling out the lack of generalization as a potential factor for the lack of fidelity. As can be seen from the table, learning higher-dimensional deformation field leads to considerable improvement in chamfer coverage and EMD metric (while still improving chamfer accuracy metric). *This highlights that point features are contributing in the enhancement of details and structures present in GT shapes, and are thus crucial for reconstructing topologically varying categories.* Figure 8 qualitatively validates this fact. We didn't observe significant improvements for $k > 4$, where k equals point features dimensionality.

6. Conclusion

In this work, we presented an approach which can learn to reconstruct 3D shapes, given a (category-specific) collection of unpaired 2D images. The proposed approach, TARS, tackles the problem of single-view reconstruction by implicitly learning to deform different object instances to a learned category specific mean shape. By transforming the 3D deformation field to a higher-dimensional field, we corroborated that the learned-deformation field is topologically-aware. As a result, our reconstructed shapes capture the global structure of the underlying GT shape and also resembles the GT shapes much than than the prior works in terms of fine structural and topological details. Furthermore, the learned deformation field implicitly captures the structural properties of the category, without any explicit supervision. Overall, our results represent an encouraging step towards generalization of reconstruction systems to the internet of images.

7. Acknowledgement

We would like to thank Shamit Lal, Jason Zhang, Alex Li, Ananye Agarwal for feedback on the paper. We also thank Chen-Hsuan Lin for providing details on the SoftRas baseline and the Pascal3D dataset, and thank Ankit Ramchandani for inpainting chair meshes for texture transfer experiment.

References

[1] https://quickdraw.withgoogle.com/. 2

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas J Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017. 7

[3] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *NeurIPS*, 2008. 2

[4] Jan Bechtold, Tatarchenko Maxim, Fischer Volker, and Brox Thomas. Fostering generalization in single-view 3d reconstruction by learning a hierarchy of local and global shape priors. In *CVPR*, 2021. 3

[5] Anand Bhattad, Aysegul Dundar, Guilin Liu, Andrew Tao, and Bryan Catanzaro. View generalization for single image textured 3d models, 2021. 2, 3, 5

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 1999. 2, 4

[7] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015. 3, 6, 7, 12

[8] Wenzheng Chen, Jun Gao, Huan Ling, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 2, 3

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 1, 3

[10] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *CVPR*. IEEE, jun 2021. 2

[11] 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 *ECCV*, 2016. 7, 12

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7, 16

[13] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *CVPR*, 2021. 4, 15, 16

[14] A. Dervieux and F. Thomasset. A finite element method for the simulation of a rayleigh-taylor instability. In Reimund Rautmann, editor, *Approximation Methods for Navier-Stokes Problems*, 1980. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[16] Shivam Duggal, Zihao Wang, Wei-Chiu Ma, Sivabalan Manivasagam, Justin Liang, Shenlong Wang, and Raquel Urtasun. Mending neural implicit modeling for 3d vehicle reconstruction in the wild. In *WACV*, 2022. 3

[17] Francis Engelmann, J. Stückler, and B. Leibe. Samp: Shape and motion priors for 4d vehicle reconstruction. *WACV*, 2017. 4

[18] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, jun 2010. 7

[19] Shubham Goel, Angjoo Kanazawa, , and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 2, 3, 5

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. 16

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. 16

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1, 4, 12

[24] Namdar Homayounfar, Yuwen Xiong, Justin Liang, Wei-Chiu Ma, and Raquel Urtasun. Levelset r-cnn: A deep variational method for instance segmentation, 2020. 5

[25] Qixing Huang, Hai Wang, and Vladlen Koltun. Single-view reconstruction via joint analysis of image and shape collections. *ACM Trans. Graph.*, 34(4), July 2015. 3

[26] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis, 2021. 2

[27] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2, 3, 4, 5, 7, 12, 13, 15

[28] Abhishek Kar, Shubham Tulsiani, João Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. *CVPR*, pages 1966–1974, 2015. 3, 4

[29] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 2, 3, 7

[30] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16:150–162, 03 1994. 3

[31] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation. *CoRR*, abs/1511.08498, 2015. 12

[32] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 2, 3, 5

[33] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *NeurIPS*, 2020. 3, 4, 5, 6, 7, 8, 12, 13, 14, 15

[34] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. *arXiv preprint arXiv:1912.00280*, 2019. 7

[35] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *ICCV*, Oct 2019. 2, 3, 6, 7, 13, 14

[36] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, 2020. 1, 4, 5

[37] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, 2020. 3

[38] Matthew Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In *ECCV*, 2014. 3

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 4

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 4

[41] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, 1987. 6

[42] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1, 2020. 2

[43] Lars M. Mescheder, Michael Oechsle, M. Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 1, 3

[44] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 4, 6

[45] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 3

[46] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3d point clouds. *arXiv preprint arXiv:2102.04014*, 2021. 7

[47] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 1

[48] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 3, 4

[49] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 4

[50] Stanley Osher and Ronald P. Fedkiw. Level set methods: An overview and some recent results. *J. Comput. Phys*, 169:463–502, 2001. 2, 5

[51] Stanley Osher and James A. Sethian. Fronts propagating with curvature dependent speed: algorithms based on hamilton–jacobi formulations. *Journal of Computational Physics*, pages 12–49, 1988. 2, 5

[52] Maks Ovsjanikov, Mirela Ben-chen, Justin Solomon, Adrian Butscher, Leonidas Guibas, and Lix École Polytechnique. Functional maps: A flexible representation of maps between shapes. 4

[53] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1, 3

[54] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 5

[55] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 5

[56] 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 *ICCV*, 2019. 1

[57] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 2, 3

[58] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. In *NeurIPS*, 2020. 4

[59] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *arXiv*, 2020. 4

[60] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 3, 4, 5, 6, 13

[61] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 3, 6, 12

[62] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do

single-view 3d reconstruction networks learn? *CVPR*, 2019. 7, 14

[63] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. State of the Art on Neural Rendering. *Computer Graphics Forum (EG STAR 2020)*, 2020. 4

[64] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason M. Saragih, Matthias Nießner, Rohit Pandey, Sean Ryan Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. State of the art on neural rendering. *CoRR*, abs/2004.03805, 2020. 3

[65] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. *CoRR*, abs/2007.14808, 2020. 2

[66] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*. IEEE, 2021. 5

[67] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *CoRR*, abs/2007.08504, 2020. 3

[68] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 3

[69] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. *CoRR*, abs/1704.06254, 2017. 12

[70] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In *3DV*, 2020. 2

[71] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4, 5

[72] Rui Wang, Nan Yang, J. Stückler, and Daniel Cremers. Directshape: Photometric alignment of shape priors for visual vehicle pose and shape estimation. *ArXiv*, abs/1904.10097, 2019. 4

[73] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 3, 6, 12

[74] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo, 2018. 2

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

[76] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *ECCV*, 2018. 3

[77] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 3, 6, 12

[78] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 33, 2020. 3, 4

[79] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2020. 2

[80] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors. In *CVPR*, June 2020. 3

[81] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016. 2

[82] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 2

[83] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation, 2020. 4, 15, 16

[84] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, July 2017. 2