

SDFit: 3D Object Pose and Shape by Fitting a Morphable SDF to a Single Image

Dimitrije Antić¹ Georgios Paschalidis¹ Shashank Tripathi²
 Theo Gevers¹ Sai Kumar Dwivedi² Dimitrios Tzionas¹

¹University of Amsterdam, The Netherlands ²Max Planck Institute for Intelligent Systems, Tübingen, Germany

{d.antic,g.paschalidis,th.gevers,d.tzionas}@uva.nl {sdwivedi,stripathi}@tue.mpg.de



Figure 1. We present SDFit, a novel framework that recovers an object’s 3D pose and shape from a single image. To this end, SDFit uses a learned, category-level, morphable SDF (mSDF) shape model, namely DIT [77], and fits this to images in a render-and-compare (a.k.a. analysis-by-synthesis) fashion. SDFit is robust to occlusions and uncommon poses, and requires no retraining for in-the-wild images. For visualizing the fitted mSDF overlaid on the input image, we show the mSDF’s normals with color coding. **Q Zoom in** for details.

Abstract

Recovering 3D object pose and shape from a single image is a challenging and ill-posed problem. This is due to strong (self-)occlusions, depth ambiguities, the vast intra- and inter-class shape variance, and the lack of 3D ground truth for natural images. Existing deep-network methods are trained on synthetic datasets to predict 3D shapes, so they often struggle generalizing to real-world images. Moreover, they lack an explicit feedback loop for refining noisy estimates, and primarily focus on geometry without directly considering pixel alignment. To tackle these limitations, we develop a novel render-and-compare optimization framework, called SDFit. This has three key innovations: First, it uses a learned category-specific and morphable signed-distance-function (mSDF) model, and fits this to an image by iteratively refining both 3D pose and shape. The mSDF robustifies inference by constraining the search on the manifold of valid shapes, while allowing for arbitrary shape topologies. Second, SDFit retrieves an initial 3D shape that likely matches the image, by exploiting foundational models for efficient look-up into 3D shape databases. Third, SDFit initializes pose by establishing rich 2D-3D

correspondences between the image and the mSDF through foundational features. We evaluate SDFit on three image datasets, i.e., Pix3D, Pascal3D+, and COMIC. SDFit performs on par with SotA feed-forward networks for unoccluded images and common poses, but is uniquely robust to occlusions and uncommon poses. Moreover, it requires no retraining for unseen images. Thus, SDFit contributes new insights for generalizing in the wild. Code is available at <https://anticdimi.github.io/sdfit>.

1. Introduction

Recovering 3D object pose and shape (OPS) from single images is key for building intelligent systems and mixed realities. However, the task is highly ill-posed due to strong challenges such as depth ambiguities, (self-)occlusions, and the huge variance in shape, appearance, and viewpoint. Yet, humans routinely solve this task by building and exploiting rich prior models through experience. Despite progress, computers still lack reliable methods and priors for reconstructing 3D objects from natural images. Our goal is to recover 3D object shape and pose from a natural image.

To this end, we draw inspiration from the “analogous” task of human pose and shape (HPS) estimation. Morphable generative body models [2, 31, 52, 71] such as SMPL [43] make HPS relatively reliable. Such models are data-driven and capture shape variance across a database of body scans. When fitting such models to single images [52, 68, 71], e.g., through the SMPLify [6] method, they act as a *strong shape prior*. That is, full-body shape can be reliably inferred even when bodies are *partially occluded*. Such occlusions are also common for object images taken in the wild.

However, perhaps counter-intuitively, there exists no SMPL-like model or SMPLify-like method for objects. But we cannot trivially adapt HPS methods for solving OPS as, despite commonalities, these tasks differ in three key ways: (1) *Shape variance* is much bigger for objects (which is both intra- and inter-class) than for bodies (which is only intra-class). For example, an armchair looks different from an airplane, but also from an office chair or a folding chair. (2) Objects have a wildly *varying topology* (e.g., chairs with a varying number of legs) while bodies have the same one. (3) To guide HPS fitting, OpenPose-like methods [7, 44] robustly detect in images 2D joints that directly correspond to 3D SMPL joints. In contrast, for general objects, detecting *correspondences* between 2D images and a textureless 3D model (let alone a *morphable* 3D model) is an open problem. Thus, OPS and HPS methods have evolved separately.

The current OPS paradigm is rendering synthetic images from 3D databases [15, 16, 67] for training deep networks to regress 3D shape from an image [1, 26–28, 62], or to generate it via image-conditioned diffusion [12, 17, 32, 46–48]. Such methods work well for in-distribution, unoccluded images, and common poses, but have three limitations: (1) They struggle generalizing to natural-looking, out-of-distribution images with occlusions and uncommon poses. (2) They mostly perform only feed-forward inference, and lack an explicit feedback loop for refining noisy estimates. (3) They mostly focus on geometry alone, largely ignoring object or camera pose, and by extension, pixel alignment.

Tackling the above limitations requires a strong shape prior for constraining the search to plausible shapes, i.e., for generating and refining plausible shape hypotheses. To this end, we exploit a category-level morphable signed-distance function (mSDF) model that generates 3D shape hypotheses through sampling its latent space (similar to SMPL [43]); here we use DIT [77]. This encodes the manifold of valid shapes, while allowing arbitrary topologies [51, 60, 77] and establishing dense correspondences across morphed shapes.

We exploit this to develop SDFit, a novel framework that fits the mSDF to an image (like SMPLify does for SMPL) by searching for a latent shape code and pose that best “matches” image cues; for an overview see Fig. 2. This has been done for 3D point clouds [36] but not for 2D images, which is much more challenging. We fill this gap here.

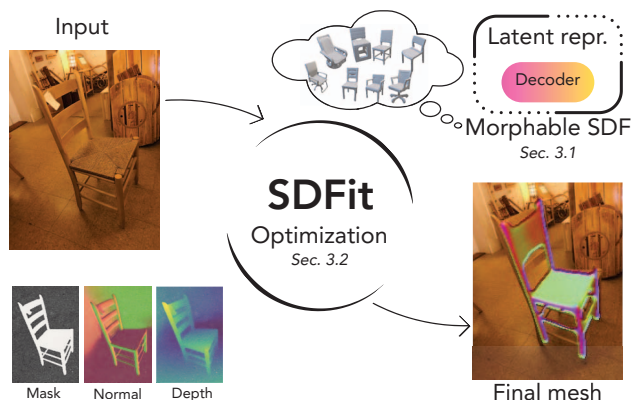


Figure 2. High-level overview of our SDFit framework. To recover both 3D object pose and shape, we fit a morphable signed-distance function (mSDF) model to observed image features (i.e., extracted normal, depth and binary masks) in a render-and-compare fashion.

However, fitting an mSDF to an image is challenging not only due to depth ambiguities, but also due to requiring a good 3D shape and pose initialization, which is still unsolved. To initialize shape, we exploit OpenShape’s [39] multimodal latent space to retrieve an mSDF shape that matches the image; this is fast and scales to large databases [16]. To initialize pose, we decorate the initial shape with foundational features [49, 58, 76], and match these to features extracted from the image. This produces 2D-3D correspondences, used to recover pose. The above has been done only for fixed shapes (3D meshes) [14, 19, 50], so they are novel for morphable shapes (mSDFs). Eventually, our framework refines both 3D pose and shape via optimization with a feedback loop, i.e., it iteratively refines the mSDF hypothesis to minimize the discrepancy between respective mSDF-rendered and image-extracted feature maps, until convergence; see example reconstructions in Fig. 1.

We evaluate on three datasets [37, 61, 69] for 3D shape estimation (with and without occlusions), and for image alignment that involves both shape and pose estimation. Evaluation shows that our SDFit fitting framework performs on par with strong feed-forward regression- [28] and diffusion-based [12, 62] baselines for unoccluded images. However, SDFit excels under occlusions, while requiring no re-training for out-of-distribution images. Note that SDFit uniquely treats both pose and shape as first-class citizens.

In summary, the main contributions of our work are: (1) A novel framework (SDFit) that uses a 3D morphable SDF (mSDF) model as a strong 3D shape prior, and fits this to a single image, while being uniquely robust to occlusions. (2) A novel mSDF shape initialization, casted as a retrieval problem in a joint latent space of 2D images and 3D shapes. (3) A novel mSDF pose initialization, using foundational models to establish rich image-to-mSDF correspondences.

Code for SDFit is available for research purposes at <https://antidimi.github.io/sdfit>.

2. Related Work

Object Shape Estimation: Recent work on 3D shape inference from images represents shape in two main ways: (1) via explicit representations like voxel grids [11, 13], point clouds [20, 65], polygonal meshes [1, 22, 27, 64] and (2) via implicit representations like Neural Radiance Fields (NeRF) [29, 54] or Signed-Distance Fields (SDF) [51, 77]. The former is easier to model but struggles with complex structures, while the latter provides more compact and flexible alternatives by encoding shapes as continuous fields. We follow the latter, and specifically SDFs.

Approaches for 3D shape estimation follow three main paradigms, i.e., they are based on regression [1, 26–28, 62, 72], generation [12, 32, 41, 46, 48] or retrieval [39, 78].

Regression methods have significantly advanced 3D shape reconstruction from single images. This includes methods like SS3D [1], which is pretrained on ShapeNet [8] and fine-tuned on real-world images, leveraging category-level models for better performance. ShapeClipper [27] enhances this with CLIP-based shape consistency. Similarly, LRM [26] and TripoSR [62] predict NeRF using a transformer, achieving detailed 3D reconstruction. Recently, ZeroShape [28] infers camera intrinsics and depth as proxy states to improve reconstruction. However, these models often struggle generalizing to unseen categories and capturing the full diversity of complex or real-world shapes.

Generative methods, such as Zero123 [41], leverage foundational models for 3D shape estimation utilizing diffusion models to generate novel views from a single image, which are then used in multiview-to-3D methods such as One-2-3-45 [40, 55]. However, appearance quality (which is usually the priority) trades off against geometry quality. SDFusion [12] learns an image-conditioned diffusion process on the latent representation of the object SDF.

Retrieval methods, such as OpenShape [39], align multimodal data, such as images and point clouds. Then, given an image, they retrieve the closest-looking 3D object from a database. However, 3D databases have finite sizes, thus, retrieved shapes might not accurately match input images. Yet, this approach is fast and scales well to large databases, so we exploit this in our work for shape initialization.

Object Pose & Shape Estimation: Recent methods on single-image object pose estimation perform either direct pose parameter estimation [22, 70] or alignment of a 3D template model with an input modality (e.g., image, features, keypoints) [10, 23, 42, 63]. The former methods directly regress rotation, translation, and scale. The latter ones predict either sparse [23, 42] or dense 3D-3D [63] correspondences, or dense 2D-3D correspondences [10, 50], and exploit these to solve for pose via the PnP [9] algorithm. While effective, this depends on accurate camera or depth data, while also requiring an a-priori known shape. We take this approach to initialize the pose of our initial shape.

More recently, ROCA [24] jointly estimates object pose and shape. To this end, it improves the pose estimate via differentiable Procrustes optimization on a retrieved CAD model. However, the fixed shape of CAD models compromises reconstruction. Similarly, Pavllo et al. [54] also estimate pose and shape using NeRFs, without any refinement. In contrast, SDFit optimizes both pose and shape using 3D-aware feature “decoration” through foundation models.

3D-aware Foundational Models: Large foundational models have catalyzed many 2D vision tasks [3]. Banani et al. [4] find that DINOv2 [49] and StableDiffusion [58] features also facilitate 3D tasks. We use features from these models to establish dense image-to-3D correspondences.

3. Method

We recover 3D object pose and shape from a single image via a novel render-and-compare framework, called SDFit; for an overview see Fig. 3. At the core of this lies a 3D morphable signed-distance function (mSDF) model (Sec. 3.1), and exploiting recent foundational models [39, 49, 58, 76].

Our SDFit framework fits the mSDF to image cues (Sec. 3.2) by jointly optimizing over its shape and pose. However, optimization-based methods are prone to local minima, so they need a good initialization. To this end, SDFit first initializes the mSDF shape through a state-of-the-art (SotA) retrieval-based technique (Sec. 3.3). Then, it initializes pose by aligning the initial shape to rich, SotA foundational features extracted from the image (Sec. 3.4).

3.1. Shape Representation

We represent 3D object shape via a learned, category-level, morphable signed-distance function (mSDF) model.

mSDF: Here we use the DIT model [77]. Each shape is encoded by a unique latent code, $z \in \mathbb{R}^{256}$, in a compact space learned by auto-decoding a 3D dataset [8]. Mapping any 3D point, x , to a signed distance is parameterized by a network $f_{\theta}^{sdf} : \mathbb{R}^3 \times \mathbb{R}^{256} \rightarrow \mathbb{R}$ (with weights θ) conditioned on latent z . Each 3D shape, S , is encoded as the mSDF’s 0-level set, $S = \{x \in \mathbb{R}^3 \mid f_{\theta}^{sdf}(x; z) = 0\}$.

DIT decodes a latent z into signed distances through a warping function, $W(x; z)$, that “warps” any 3D point, x , to a canonical space defined by a learned SDF template, T . This models the inter-category shape variance w.r.t. the template, and defines dense correspondences to it. Note that training DIT comes with a useful byproduct, that is, it yields a collection of latent codes, \mathcal{Z} , for all training shapes z . We use these later to initialize the shape hypothesis (Sec. 3.3).

Rendering: Rendering an mSDF is not straightforward, so we extract a 3D mesh as a proxy that we exploit for differentiable rendering. In each iteration we take three steps: (1) we predict SDF values via f_{θ}^{sdf} on a 3D grid, (2) we extract a mesh using FlexiCubes [59], and (3) we pose it by applying a 6-DoF rigid transformation $(R, t) \in \text{SE}(3)$.

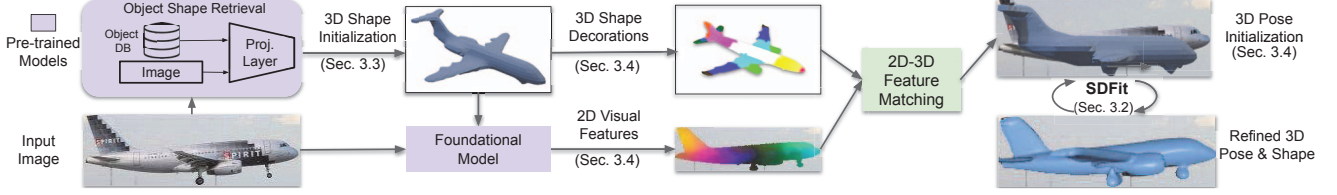


Figure 3. Our SDFit framework. We represent 3D shape via a learned morphable signed-distance function (mSDF) model [77] (Sec. 3.1). We first recover a likely initial shape from a database [8] via a SotA retrieval method [39] conditioned on the input image (Sec. 3.3). Next, we extract features from both the target image and the initial shape via foundational models [49, 76] to establish image-to-mSDF 2D-to-3D correspondences and initialize pose (Sec. 3.4). Last, we iteratively refine both shape and pose via render-and-compare (Sec. 3.2).

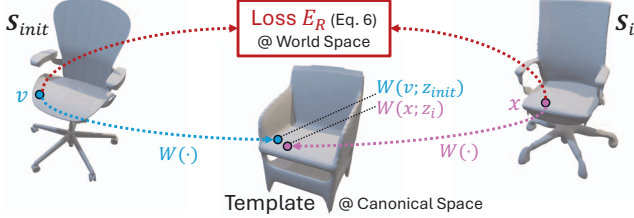


Figure 4. Loss E_R (Eq. (6)). Initial and hypothesis shapes, S_{init} and S_i , are warped to a canonical space via DIT’s warper $W(\cdot)$. For each warped vertex of S_i we find the closest warped vertex of S_{init} in canonical space, and compute MSE in world space.

3.2. Fitting Pose & Shape

To recover OPS from an image, SDFit optimizes over object shape, $z \in \mathbb{R}^{256}$, scale, $s \in \mathbb{R}^3$, and pose, $(R, t) \in \text{SE}(3)$, by minimizing via render-and-compare the energy function:

$$E = E_{\mathcal{M}} + \lambda_{\mathcal{N}} E_{\mathcal{N}} + \lambda_{\mathcal{D}} E_{\mathcal{D}} + \lambda_{DT} E_{DT} + \lambda_R E_R, \quad (1)$$

where \mathcal{M} is the mask, \mathcal{N} the normal map, \mathcal{D} the depth map, DT denotes a 2D distance transform, R denotes regularization, and λ are steering weights. The individual energy terms are:

$$E_{\mathcal{M}} = \text{MSE}(\widehat{\mathcal{M}}^i, \mathcal{M}) + \lambda_{IoU} \cdot \text{IoU}(\widehat{\mathcal{M}}^i, \mathcal{M}), \quad (2)$$

$$E_{\mathcal{D}} = \text{SSI-MAE}(\widehat{\mathcal{D}}^i, \mathcal{D}), \quad (3)$$

$$E_{\mathcal{N}} = \text{MSE}(\widehat{\mathcal{N}}^i, \mathcal{N}), \quad (4)$$

$$E_{DT} = \sum_{\hat{x} \in \widehat{\mathcal{C}}^i} \min_{x \in \mathcal{C}} \|\hat{x} - x\|_1. \quad (5)$$

where non-hat symbols are “ground-truth” observations, hat denotes maps rendered from the running mSDF hypothesis, i is the running iteration, \mathcal{C} the mask contour, MSE the mean squared error, IoU the intersection-over-union, while SSI-MAE is a scale- and shift-invariant depth loss [56].

To regularize fitting under self-occlusions, a regularization loss, E_R , encourages the running shape hypothesis, S_i , to be consistent with the initial estimate, S_{init} (Sec. 3.3). A simple way for this is to penalize deviation of the running z code from the code z_{init} of S_{init} , but, empirically, this causes local minima when S_{init} has a wrong topology (e.g., a chair that erroneously misses armrests).

Instead, SDFit geometrically regularizes to S_{init} so it can still refine the topology (e.g., chairs growing missing armrests). To this end, it uses the correspondences of S and S_{init} to the template, T , to map each vertex $x \in S_i$ to the closest vertex $u \in S_{init}$. Specifically, as shown in Fig. 4, (1) it warps S_i vertices on the template in *canonical space*, (2) it warps S_{init} vertices on the same template as well, and (3) for each warped vertex of S_i it finds the closest warped vertex of S_{init} , and eventually (4) computes the MSE for corresponding vertices in *world space*. In technical terms:

$$E_R = \text{MSE}(S_i, S_{init}), \quad (6)$$

$$S_{init} = \{v \mid \arg \min_{v \in S_{init}} \|W(v; z_{init}) - W(x; z_i)\|_2\}, \quad (7)$$

where $x \in S_i$ are vertices of S_i , $W(x; z_i)$ are these vertices mapped into the canonical space via the warper $W(\cdot)$, $v \in S_{init}$ are vertices of S_{init} , $W(v; z_{init})$ are these vertices mapped into the canonical space, and $S_i = f_{\theta}^{sdf}(x; z_i)$ is the i -th iteration shape hypothesis (in canonical space).

During optimization, in each iteration, SDFit evaluates the energy function E of Eq. (1), backpropagates gradients, and updates the hypothesis parameters z_i , R_i , and t_i .

3.3. Shape Initialization

SDFit initializes the shape code, z , by exploiting the retrieval-based OpenShape [39] model, denoted as f^{db} . This encodes multiple modalities (images, 3D point clouds) into a joint latent space, and facilitates searching for the 3D object, S_{init} , that best resembles an input image, \mathcal{I} , by: (1) embedding the shapes \mathcal{S} of mSDF training data via f^{db} ; (2) embedding image \mathcal{I} into the same latent space via f^{db} ; (3) retrieving the shape whose embedding most closely lies to the image embedding. More formally, the initial-shape latent code, z_{init} , is the code z whose 3D shape embedding, $f^{db}(S_z)$ lies closest to the image embedding, $f^{db}(\mathcal{I})$, via the cosine-similarity metric:

$$z_{init} = \arg \max_{z \in \mathcal{Z}} \frac{f^{db}(\mathcal{I}) \cdot f^{db}(S_z)}{\|f^{db}(\mathcal{I})\|_2 \|f^{db}(S_z)\|_2}, \quad (8)$$

where S_z is 0-level set of $f_{\theta}^{sdf}(x; z)$, z is the shape latent code, and \mathcal{Z} is a database of auto-decoded latent codes, each corresponding to a shape instance in the mSDF training set.

3.4. Pose Initialization

To initialize 3D pose from a single view, SDFit: (1) establishes correspondences between 2D pixels and 3D points, (2) estimates camera intrinsics, (3) filters out noisy correspondences with RANSAC, and (4) applies the PnP method.

To find correspondences, SDFit computes image features from the input image and rendered mSDF images. To this end, inspired from image-to-image matching, it leverages features from foundational models such as StableDiffusion (SD) [58] (or ControlNet [76]) and DINOv2 [49]. Specifically, it computes hybrid features that combine SD v1.5 (ControlNet) and DINOv2 ones, as these encode geometry and semantic cues [18, 45, 74] that are crucial for 3D understanding. In detail, it establishes 2D-3D pixel-vertex correspondences as described in the following paragraphs.

Image Features: SDFit uses the pretrained ControlNet [76] and DINOv2 [49] models. It conditions ControlNet on the prompt ‘‘A <category>, photorealistic, real-world’’, as well as on normal and depth maps estimated from an image for inpainting [18], i.e., hallucinating the original image from condition signals. Crucially, this pushes ControlNet to semantically differentiate between nearby pixels [18], so features extracted from its layers capture *semantic* cues. Then, it applies DINOv2 [49] on an image to extract features capturing *geometric* cues [4]. Last, it forms hybrid features by concatenating per pixel the complementary ControlNet and DINOv2 features.

In technical terms, for an input image \mathcal{I} , estimated [33] normal and depth maps, \mathcal{N} and \mathcal{D} , and a text prompt, SDFit uses a pretrained ControlNet to generate (inpaint) a ‘‘textured’’ image, \mathcal{I}^{tex} . To get ControlNet features, at the last diffusion step SDFit extracts features \mathcal{F}_2^{diff} and \mathcal{F}_4^{diff} from its UNet-decoder layers 2 and 4, respectively, upsamples these to the resolution of \mathcal{I} , and concatenates these to obtain the feature $\mathcal{F}^{diff} = \{\mathcal{F}_2^{diff} || \mathcal{F}_4^{diff}\}$. Note that here features from early layers emphasize semantic and geometric cues over texture ones [4, 18], which is beneficial as our mSDF models geometry but is textureless. To get \mathcal{F}^{DINOv2} features, it applies the DINOv2 model on the textured image \mathcal{I}^{tex} (applicable also for the mSDF, see next paragraph), to extract per-pixel geometric cues. To form the final features, it concatenates [74] per pixel the normalized \mathcal{F}^{diff} features with \mathcal{F}^{DINOv2} ones, as $\mathcal{F} = \{\alpha \mathcal{F}^{diff}, (1 - \alpha) \mathcal{F}^{DINOv2}\}$, where α is a steering weight. A detection mask, \mathcal{M} , steers focus only on object pixels. Below, the flattened features, $\mathcal{F}_{M(\mathcal{I})}$, are denoted as $\mathcal{F}_{\mathcal{I}}$ for notational brevity.

Shape (mSDF) Features: Recently, Diff3F [18] decorates 3D meshes with features extracted via ControlNet [76] and DINOv2 [49]. SDFit follows this to obtain features for the textureless mSDF and establish 2D-3D correspondences with image features in a zero-shot fashion. This is a novel use of Diff3F for a long-standing problem. Note that this does not require a known object-part connectivity [6, 66].



Figure 5. Features for mSDF. Rather than decorating each mSDF from scratch, we can query precomputed features via DIT’s correspondences and warper, $W(\cdot)$. We either pre-decorate the category template ($feat@T$) or the initial shape per image ($feat@S_{init}$).

Note also that the DIT [77] mSDF model establishes dense correspondences across all morphed shapes within a class.

SDFit can perform the above in two different ways:

- (1) ‘‘SDFit $feat@T$ ’’ (Fig. 5-left): It decorates a mesh extracted from the mSDF template, T , only once per category, offline. Then, for every morphed mSDF shape, it queries decoration features from the already decorated T .
- (2) ‘‘SDFit $feat@S_{init}$ ’’ (Fig. 5-right): It decorates a mesh extracted from the initial mSDF shape, S_{init} (Sec. 3.3). It does so once per image, as S_{init} differs across images. The above options trade efficiency for accuracy; the former is computationally cheaper, but the latter is more accurate.

In any case, for decoration, SDFit first extracts a mesh [59] from either T or from S_{init} . Then, it samples J views on a unit sphere around it, and for each view $j \in J$, it renders normal maps, \hat{N}^j , and depth maps, \hat{D}^j . Then, it extracts per-pixel feature maps, \mathcal{F}_S^j , in the same way as for image features, discussed above. Since the P^j camera parameters are known, each view-specific feature map, \mathcal{F}_S^j , gets unprojected onto 3D mesh vertices. Last, for each vertex, the unprojected features across views are aggregated to form the final feature, $\mathcal{F}_S \in \mathbb{R}^{|S| \times 2368}$, where $|S|$ is the number of vertices and 2368 is the feature dimension.

Object-to-Image Alignment: Using the extracted image features, $\mathcal{F}_{\mathcal{I}}$, and shape feature maps, \mathcal{F}_S , SDFit establishes 2D-3D pixel-vertex correspondences, \mathcal{C} , by finding in feature space the most similar vertex for each pixel:

$$\mathcal{C} = \{\{i, s\} = \arg \max_{s \in S_{init}} \mathcal{A}_{i,s}, \text{ for all pixels } i\}. \quad (9)$$

$$\mathcal{A}_{i,s} = \frac{\mathcal{F}_{\mathcal{I}}^i \cdot \mathcal{F}_S^s}{\|\mathcal{F}_{\mathcal{I}}^i\|_2 \|\mathcal{F}_S^s\|_2}, \quad (10)$$

where \mathcal{A} is a cosine-similarity matrix, and $\mathcal{F}_{\mathcal{I}}^i$ and \mathcal{F}_S^s are i -th pixel and s -th vertex features. By exploiting the correspondences, \mathcal{C} , SDFit implicitly finds the visible mSDF points, as only these can be matched to pixels (see Sec. S.1).

Moreover, SDFit estimates intrinsic camera parameters, K , via the off-the-shelf PerspectiveFields [30] model applied on image \mathcal{I} . Last, it uses the estimated correspondences, \mathcal{C} , and intrinsics, K , to apply the RANSAC [21] and PnP [9] algorithms for estimating the object pose, R_{init}, t_{init} . This pose, along with the initial object shape, z_{init} (Sec. 3.3), initializes our fitting framework (Sec. 3.2).

4. Experiments

4.1. Implementation Details

mSDF: We use the DIT [77] model trained on ShapeNet [8]. But our approach is agnostic to the chosen mSDF, that is, as richer mSDFs get developed, SDFit also gets better.

Pose Initialization (Sec. 3.4): We establish image-to-shape 2D-3D correspondences by matching deep features. However, these might be imperfect as this is still an open problem. Therefore, we compute pose as follows. First, we apply RANSAC+PnP on the established correspondences, and generate two hypotheses by mirroring pose around the vertical axis. Then, we refine each hypothesis over 200 iterations and select the one with the lower E_D from Eq. (2).

Normal, Depth & Mask Maps: For the objective function of Eq. (1) we need an observed “ground truth” segmentation mask, \mathcal{M} , normal map, \mathcal{N} , and depth map, \mathcal{D} , and respective maps rendered from the mSDF, $\widehat{\mathcal{M}}$, $\widehat{\mathcal{D}}$, and $\widehat{\mathcal{N}}$.

For $\widehat{\mathcal{M}}$, $\widehat{\mathcal{D}}$, $\widehat{\mathcal{N}}$, we extract a mesh via FlexiCubes [59] with a grid size of $N = 32$ and render with Nvdiffrast [35].

We estimate \mathcal{N} and \mathcal{D} by applying the OmniData [33] model on the input image. The masks \mathcal{M} can be provided by datasets (e.g., in Pix3D [61]), while in the opposite case (e.g., for Pascal3D+ [69]) we segment objects by applying the rembg [57] method (as in ZeroShape [28]).

Fitting (Sec. 3.2): We optimize with Adam [34]. For the first 300 iterations, we refine the initial pose, (R_{init}, t_{init}) , and scale, s_{init} , keeping shape S_{init} fixed. For the next 1000 iterations, we jointly optimize shape, scale and pose.

4.2. Metrics

We use four complementary numeric metrics as follows.

Chamfer Distance (CD): CD quantifies the similarity of two 3D point clouds X and Y as the average (bidirectional) distance from each point in a cloud to the nearest point in the other one. Then, with $|\cdot|$ denoting cardinality:

$$CD = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} \|x - y\|_2 + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} \|x - y\|_2. \quad (11)$$

F-Score: Given a rejection threshold, d , the F-Score at distance d ($F@d$) is the harmonic mean of *precision@ d* and *recall@ d* , reflecting the proportion of the surface accurately reconstructed within the correctness threshold, d .

Intersection-over-Union (IoU): IoU encodes the alignment of an estimated 3D shape with image pixels, by quantifying the alignment between a target mask (detected in image) and estimated mask (projected 3D shape onto 2D) as:

$$IoU = (TP) / (TP + FP + FN) \times 100, \quad \text{where:} \quad (12)$$

TP is true positives, FP false positives, FN false negatives.

CLIP Similarity: To assess how plausible 3D shapes look like for a given class (e.g., a “chair”), we first compute the CLIP embedding of the class name and of a rendered 3D-shape image, and then their CLIP Similarity [25].

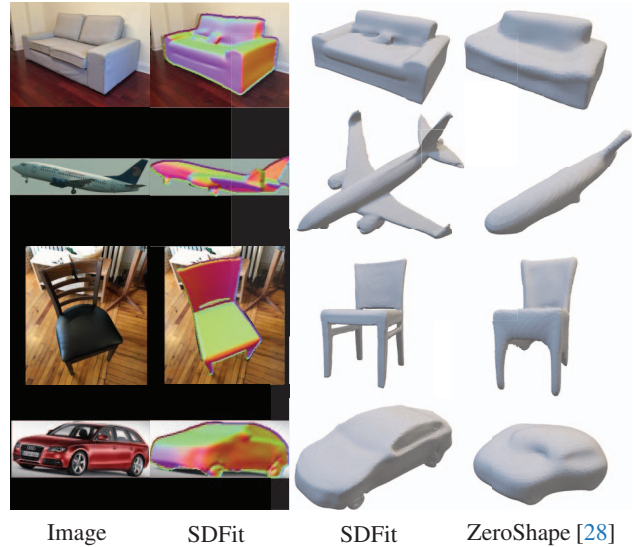


Figure 6. Shape recovery for SDFit ($\text{feat}@S_{init}$) and ZeroShape [28]. SDFit jointly fits pose and shape to the image, helping pixel alignment. It also excels at recovering occluded parts via the mSDF’s learned shape prior. Overlays show the mSDF’s normals.

4.3. Evaluation

We evaluate on *shape reconstruction*, capturing only geometry, and *image alignment*, capturing both shape and pose.

Shape Reconstruction: We evaluate on the standard Pix3D dataset [61], which pairs real-world images with ground-truth CAD models, using ZeroShape’s [28] test set. Specifically, we compare our fitting-based SDFit method against the regression-based ZeroShape [28] and TripoSR [62], and the diffusion-based SDFusion [12] model. For a more direct comparison, we also train a class-specific ZeroShape [28], denoted as ZeroShape-CLS.

The results, presented in Tab. 3, show that SDFit performs on par with ZeroShape and TripoSR in terms of the CD metric. However, we notice that ZeroShape often defaults to “blobby” shapes (see Fig. 6). We hypothesize that this might be because it is “only” feed-forward, so it cannot correct potential mistakes. Moreover, its search space might be insufficiently constrained, so it often struggles to produce shapes resembling the depicted object class.

However, the CD metric cannot fully capture such artifacts, as it measures only geometric proximity to ground-truth shapes, and ignores semantics. To capture semantics, we assess how well a recovered 3D shape aligns with the target class via CLIP similarity [25, 79]. To this end, we first compute the CLIP embedding for the class name. Then, we render synthetic images of the fitted mSDF from five canonical viewpoints, and compute their CLIP embedding. Last, we compute the cosine-similarity score between the class embedding and the five mSDF embeddings, taking the max score to account for poor views. As shown in Tab. 3, SDFit yields shapes that better reflect the target class.

Metric: CD ↓		Pix3D					COMIC				
Method	Inpaint	Chair	Sofa	Mean	Box	Bottle	Camera	Knife	Mug	Spray	Mean
ZeroShape [28]	✗	7.08	5.71	6.40	9.32	5.37	9.25	3.24	7.31	4.67	6.53
ZeroShape-CLS [28]	✗	5.94	5.14	5.54	7.14	4.90	6.40	1.98	4.44	4.90	4.96
TripoSR [62]	✗	6.73	7.45	7.09	10.7	6.48	9.66	3.28	6.23	4.81	6.86
ZeroShape [28]	✓	5.64	4.81	5.23	7.91	4.38	8.42	2.77	6.52	3.83	5.64
ZeroShape-CLS [28]	✓	5.71	4.83	5.27	6.96	4.38	6.05	1.55	3.92	6.54	4.90
TripoSR [62]	✓	6.03	6.95	6.49	6.39	4.25	7.06	2.19	4.91	4.60	4.90
SDFit-feat@ S_{init}	✗	4.08	3.40	3.74	1.21	4.04	6.52	1.11	2.94	3.66	3.25

Table 1. **Shape reconstruction evaluation under occlusion** on Pix3D [61] (synthetic-patch occlusions) and COMIC [37] (hand-object grasp occlusions). For Pix3D we occlude 40% of the object bounding box (see Sec. S.2 in Sup. Mat.). We report the Chamfer Distance (CD).

	Type	CD ↓	F@1 ↑	F@2 ↑	CLIP-10 ² ↑
SDFusion [12]	Diff.	3.95	0.16	0.38	29.87
ZeroShape [28]	Reg.	3.44	0.23	0.47	29.33
ZeroShape-CLS [28]	Reg.	4.55	0.21	0.46	29.40
TripoSR [62]	Reg.	3.41	0.13	0.32	29.87
SDFit-feat@ S_{init}	Opt.	3.53	0.25	0.46	30.53

Table 3. **Shape reconstruction evaluation** on Pix3D [61]. We report the mean Chamfer Distance (CD), F-Score at two thresholds (F@1 and F@2), and CLIP similarity across the *Chair* and *Sofa* categories; each value is the average over these classes.

Shape Reconstruction under Occlusion: We quantitatively evaluate robustness to occlusion by (1) rendering synthetic occluding patches on Pix3D images covering 40% of the object’s bounding box (Fig. 7), and (2) using the COMIC [37] dataset of hand-object grasps.

Results are shown in Tab. 1, and a sensitivity analysis in Fig. 8, and Sec. S.2. We see that SDFit clearly outperforms ZeroShape. Note that SDFit has stable performance for increasingly stronger occlusions, while ZeroShape heavily degrades. We think that this is because SDFit relies on geometric cues only from unoccluded regions, which remain intact, while ZeroShape relies on “global” appearance cues that are strongly influenced by occlusions. Moreover, SDFit uses an explicit shape prior (mSDF) and a feedback loop, while ZeroShape uses an implicit shape prior (baked into network weights) and does only feed-forward inference.

To help baselines handle occlusions, we give them the privilege of inpainting [73]. That is, we remove occluders, infill the missing object pixels [73], and apply the baseline on the new unoccluded image. The privileged baselines have an improved performance, but SDFit clearly outperforms these. This is because SDFit relies only on features from unoccluded regions, and leverages the mSDF shape manifold of valid shapes to recover the occluded parts.

We compare the best performers of Tab. 3, i.e., SDFit and ZeroShape [28], in Fig. 6. ZeroShape struggles recovering self-occluded parts. Instead, SDFit recovers these via the regularizer E_R in Eq. (7). This aligns with the previous paragraph, i.e., SDFit is more robust to self-occlusions or occlusions by third parties. This is because SDFit exploits correspondences (see Fig. 5) between S_{init} and the running mSDF hypothesis for supervising self-occluded regions.

# Views	Diff. steps	Runtime (sec)	CD ↓
100	100	600	3.53
16	10	27	3.67
8	30	30	3.69
8	10	10	3.58

Table 2. Runtime analysis for shape decoration. We assess the impact of the number of views, diffusion steps, and runtime for *Chair* and *Sofa* in Pix3D [61].



Figure 7. Examples of synthetic occluders of varying size. The labels denote the percentage of object (bounding-box) occlusion.

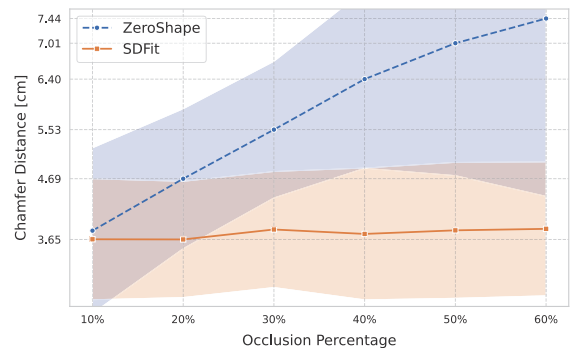


Figure 8. Occlusion sensitivity analysis. We evaluate shape reconstruction (Y-axis) on the Pix3D [61] test set with a varying degree of occlusion (X-axis). SDFit outperforms ZeroShape in both mean and standard deviation (lower is better) and remains stable under increasing occlusion, while ZeroShape heavily degrades.

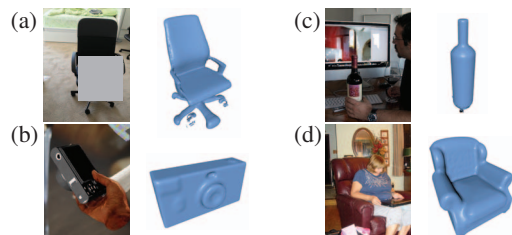


Figure 9. Reconstructions of SDFit-feat@ S_{init} on images of three datasets: (a) Pix3D [61] with synthetic occluding patches, used in Fig. 8 and Tab. 1, (b) COMIC [37], and (c, d) COCO [38].

Image Alignment: We evaluate joint shape-and-pose estimation for inferring pixel-aligned 3D objects. We use the Pascal3D+ [69] dataset and specifically the test split of Pavllo et al. [53] for the *car* and *airplane* classes.

Since SotA methods [12, 28] focus mostly on shape recovery, ignoring pose, we establish our own baselines, by extending these methods with our pose initialization and RnC fitting as follows: We first infer shape through SotA

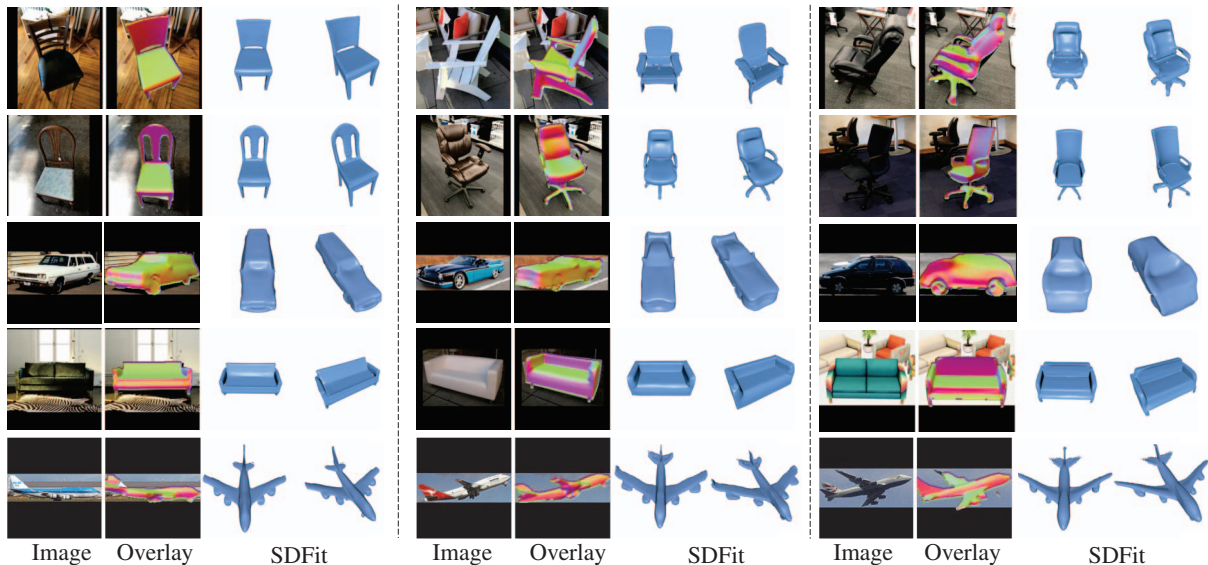


Figure 10. Qualitative results for SDFit ($\text{feat}@S_{init}$) on images of the Pix3D [61] and Pascal3D+ [69] datasets. We show the estimated 3D shape (left to right) in camera- (as normal map), front- and side-view. **Q Zoom in** to see details.

regression [28] or diffusion [12] methods. We then keep the estimated shape fixed, and optimize over pose and scale with our render-and-compare module (RnC) (Sec. 3.2). For fairness, we initialize the object pose and camera intrinsics of all baselines using SDFit’s pose initialization (Sec. 3.4). For ZeroShape, we initialize only the translation since it assumes that the world and camera frames are aligned.

Table 4 reports the 2D Intersection-over-Union (%). “SDFit- $\text{feat}@S_{init}$ ” outperforms all baselines, while “SDFit- $\text{feat}@T$ ” is on par with “ZeroShape+RnC” but outperforms others. This shows that SDFit (whose RnC module is used to extend baselines) and ZeroShape can be complementary. Note that baselines refine only the pose through RnC. Instead, SDFit uniquely refines both pose and shape by morphing the mSDF – this is a key advantage.

Note that SDFit- $\text{feat}@T$ trades speed for accuracy; it is faster but less accurate than SDFit- $\text{feat}@S_{init}$, as the topology of S_{init} matches the image better than T . For further ablations of our modules, see Sec. S.3 in Sup. Mat.

Qualitative Results: We show extensive results of our SDFit- $\text{feat}@S_{init}$ for in-the-wild Pascal3D+ [69] and Pix3D [61] images in Fig. 10. Despite the diverse shapes, appearances, and challenging imaging conditions (e.g., poor lighting, uncommon poses) in real-world images, SDFit recovers plausible, pixel-aligned 3D shapes, showing promising generalization. Unlike purely data-driven methods, SDFit does not need retraining for unseen images.

Moreover, in Fig. 9 we show reconstructions of SDFit *under occlusion* on Pix3D [61] and COMIC [37], as well as on COCO [38] images that are taken in the wild. SDFit’s reconstructions look robust to strong occlusions, reflecting the findings of Tab. 1 and Fig. 8, and Sup. Mat. Sec. S.2.

Metric: 2D IoU (%) \uparrow	Pascal3D+		Pix3D		Mean IoU
	Plane	Car	Chair	Sofa	
SDFusion [12] + RnC	N/A	N/A	59.5	N/A	N/A
OpenShape [39] + RnC	48.5	79.5	54.3	85.2	66.8
ZeroShape [28] + RnC	76.4	89.1	61.3	85.9	78.1
SDFit $\text{feat}@T$	77.6	86.9	61.1	89.7	78.8
SDFit $\text{feat}@S_{init}$	81.5	91.6	75.9	92.7	85.4

Table 4. Image-alignment performance on the Pascal3D+ [69] and Pix3D [61] datasets. The shape predictions of competing methods are aligned to the image in a render-and-compare (RnC) fashion similarly to our SDFit. We report the per-category IoU metric, as well as the Mean IoU across all categories.

Runtime: Our SDFit method fully converges in ~ 3 min, often obtaining a satisfactory result within the first 45–60 sec, using an Nvidia 4080 GPU, with an additional 20 sec for image feature extraction [49, 76], and 10 sec for shape decoration. For the latter, see Tab. 2, where the top-most row corresponds to Diff3F [18]; heavily reducing the number of views and diffusion steps does not harm accuracy, while reducing runtime 60x w.r.t. Diff3F. We hypothesize that Diff3F’s many views introduce redundant information that causes noise accumulation during feature aggregation.

5. Conclusion

We develop SDFit, a novel method for fitting an explicit morphable 3D shape prior to single images. This uniquely refines both shape and pose using an explicit feedback loop. This achieves much better pixel alignment than SotA methods, and is exceptionally robust to occlusions. We believe that this is interesting for the broader 3D community and will inspire work that combines the best of our work and learning-based models. To this end, our code is available.

6. Acknowledgments

Acknowledgements: We thank Božidar Antić, Yuliang Xiu and Muhammed Kocabas for useful insights. We acknowledge EuroHPC JU for awarding the project ID EHPC-AI-2024A06-077 access to Leonardo BOOSTER. This work also used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-7589. This work is partly supported by the ERC Starting Grant (project STRIPES, 101165317, PI: D. Tzionas). D. Tzionas has received a research gift from Google, and from the NVIDIA Academic Grant Program.

References

- [1] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pretrain, self-train, distill: A simple recipe for super-sizing 3D reconstruction. In *CVPR*, pages 3763–3772, 2022. 2, 3
- [2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *TOG*, 24:408–416, 2005. 2
- [3] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. arXiv:2307.13721, 2023. 3
- [4] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *CVPR*, pages 21795–21806, 2024. 3, 5, 1
- [5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025. 1
- [6] 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 *ECCV*, pages 561–578, 2016. 2, 5
- [7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *TPAMI*, 43(1):172–186, 2021. 2
- [8] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiang Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012, 2015. 3, 4, 6
- [9] Bo Chen, Álvaro Parra, Jiwei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating PnP optimization. In *CVPR*, pages 8097–8106, 2019. 3, 5
- [10] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. MonoRUN: Monocular 3D object detection by reconstruction and uncertainty propagation. In *CVPR*, pages 10379–10388, 2021. 3
- [11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 3
- [12] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tuyakov, Alex Schwing, and Liangyan Gui. SDFusion: Multimodal 3D shape completion, reconstruction, and generation. In *CVPR*, pages 4456–4465, 2023. 2, 3, 6, 7, 8, 1
- [13] 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. *ECCV*, 9912:628–644, 2016. 3
- [14] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshminpathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *CVPR*, 2025. 2
- [15] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *NeurIPS*, 2023. 2
- [16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, pages 13142–13153, 2023. 2
- [17] Congyue Deng, Chiyu Max Jiang, Charles R. Qi, Xichen Yan, Yin Zhou, Leonidas J. Guibas, and Dragomir Anguelov. NeRD: Single-view NeRF synthesis with Language-Guided diffusion as general image priors. In *CVPR*, pages 20637–20647, 2023. 2
- [18] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J. Mitra. Diffusion 3D features (Diff3F): Decorating untextured shapes with distilled semantic features. In *CVPR*, pages 4494–4504, 2024. 5, 8
- [19] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *CVPR*, 2025. 2
- [20] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3D object reconstruction from a single image. *CVPR*, pages 2463–2471, 2017. 3
- [21] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5
- [22] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. *ICCV*, pages 9784–9794, 2019. 3
- [23] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *ECCV*, pages 516–532, 2022. 3
- [24] Can Gümeli, Angela Dai, and Matthias Nießner. ROCA: Robust CAD model retrieval and alignment from a single image. In *CVPR*, pages 4012–4021, 2022. 3
- [25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Confer-*

- ence on Empirical Methods in Natural Language Processing, *EMNLP 2021*, pages 7514–7528, 2021. 6
- [26] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *ICLR*, 2024. 2, 3
- [27] Zixuan Huang, Varun Jampani, Anh Thai, Yuanzhen Li, Stefan Stojanov, and James M. Rehg. ShapeClipper: Scalable 3D shape learning from single-view images via geometric and CLIP-based consistency. In *CVPR*, pages 12912–12922, 2023. 3
- [28] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M. Rehg. ZeroShape: Regression-based zero-shot shape reconstruction. In *CVPR*, pages 10061–10071, 2024. 2, 3, 6, 7, 8, 1
- [29] Won Jun Jang and Lourdes de Agapito. CodeNeRF: Disentangled neural radiance fields for object categories. *ICCV*, pages 12929–12938, 2021. 3
- [30] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *CVPR*, pages 17307–17316, 2023. 5
- [31] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 2
- [32] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3D implicit functions. arXiv:2305.02463, 2023. 2, 3
- [33] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3D common corruptions and data augmentation. In *CVPR*, pages 18963–18974, 2022. 5, 6
- [34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [35] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *TOG*, 39(6), 2020. 6
- [36] Haoang Li, Jinhu Dong, Binghui Wen, Ming Gao, Tianyu Huang, Yun-Hui Liu, and Daniel Cremers. DDIT: Semantic scene completion via deformable deep implicit templates. In *ICCV*, pages 21837–21847, 2023. 2
- [37] Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. CHORD: Category-level hand-held object reconstruction via shape deformation. In *ICCV*, pages 9410–9420, 2023. 2, 7, 8
- [38] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 7, 8
- [39] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yin hao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *NeurIPS*, 2023. 2, 3, 4, 8
- [40] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization. *NeurIPS*, 2024. 3
- [41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, P. Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. *ICCV*, pages 9264–9275, 2023. 3
- [42] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. AutoShape: Real-time shape-aware monocular 3D object detection. In *ICCV*, pages 15621–15630, 2021. 3
- [43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):248:1–248:16, 2015. 2
- [44] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. In *CVPRW*, 2019. 2
- [45] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023. 5
- [46] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion 360° reconstruction of any object from a single image. In *CVPR*, pages 8446–8455, 2023. 2, 3
- [47] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. PC²: Projection-conditioned point cloud diffusion for single-image 3D reconstruction. In *CVPR*, pages 12923–12932, 2023.
- [48] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-E: A system for generating 3D point clouds from complex prompts. arXiv:2212.08751, 2022. 2, 3
- [49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 2, 3, 4, 5, 8, 1
- [50] Evin Pinar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. FoundPose: Unseen object pose estimation with foundation features. In *ECCV*, 2024. 2, 3
- [51] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2, 3
- [52] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2
- [53] Dario Pavlo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3D meshes from real-world images. In *ICCV*, pages 13859–13869, 2021. 7

- [54] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *CVPR*, pages 4391–4401, 2023. 3
- [55] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *ICLR*, 2024. 3
- [56] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 44(3):1623–1637, 2022. 4
- [57] Rembg: A tool to remove images background. <https://github.com/danielgatis/rembg>, 2022. 6
- [58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 2, 3, 5
- [59] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *TOG*, 42(4):37:1–37:16, 2023. 3, 5, 6
- [60] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 2
- [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*, pages 2974–2983, 2018. 2, 6, 7, 8, 1
- [62] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv:2403.02151*, 2024. 2, 3, 6, 7
- [63] He Wang, Srinath Sridhar, Jingwei Huang, Julien P. C. Valentin, Shuran Song, and L. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *CVPR*, pages 2642–2651, 2019. 3
- [64] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, W. Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *ECCV*, pages 55–71, 2018. 3
- [65] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. PQ-NET: A generative part Seq2Seq network for 3D shapes. *CVPR*, pages 826–835, 2020. 3
- [66] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3D animals in the wild. In *CVPR*, pages 8792–8802, 2023. 5
- [67] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-Vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *CVPR*, pages 803–814, 2023. 2
- [68] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10957–10966, 2019. 2
- [69] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, pages 75–82, 2014. 2, 6, 7, 8
- [70] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018. 3
- [71] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, pages 6183–6192, 2020. 2
- [72] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, pages 8843–8852, 2021. 3
- [73] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv:2304.06790*, 2023. 7
- [74] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *NeurIPS*, 2023. 5, 1
- [75] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *CVPR*, pages 3076–3085, 2024. 1
- [76] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to Text-to-Image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 4, 5, 8, 1
- [77] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3D shape representation. In *CVPR*, pages 1429–1439, 2021. 1, 2, 3, 4, 5, 6
- [78] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *ICLR*, 2024. 3
- [79] Thomas Hanwen Zhu, Ruining Li, and Tomas Jakab. DreamHOI: Subject-driven generation of 3D human-object interactions with diffusion priors. *arXiv:2409.08278*, 2024. 6