

Learning to Estimate Multi-view Pose from Object Silhouettes

Yoni Kasten^{1*}, True Price², David Geraghty², and Jan-Michael Frahm²

¹ NVIDIA Research

² Meta

ykasten@nvidia.com, {jtprice,dger,jmfrahm}@meta.com

Abstract. While Structure-from-Motion pipelines certainly have their success cases in the task of 3D object reconstruction from multiple images, they still fail on many common objects that lack distinctive texture or have complex appearance qualities. The central problem lies in 6DOF camera pose estimation for the source images: without the ability to obtain a good estimate of the epipolar geometries, all state-of-the-art methods will fail. Although alternative solutions exist for specific objects, general solutions have proved elusive. In this work, we revisit the notion that silhouette cues can provide reasonable constraints on multi-view pose configurations when texture and priors are unavailable. Specifically, we train a neural network to holistically predict camera poses and pose confidences for a given set of input silhouette images, with the hypothesis that the network will be able to learn cues for multi-view relationships in a data-driven way. We show that our network generalizes to unseen synthetic and real object instances under reasonable assumptions about the input pose distribution of the images, and that the estimates are suitable to initialize state-of-the-art 3D reconstruction methods.

1 Introduction

Three-dimensional object reconstruction, the process of converting imagery of an object into a representation of its geometry, is an increasingly mainstream component of augmented- and virtual-reality (AR/VR) research and applications, with much of this growth due to the increasing facility and scalability of capture technologies. In AR/VR entertainment, for example, commodity 3D scanning technology can efficiently generate photorealistic models for use in virtual worlds, reducing manual effort required by 3D artists. Likewise, many research applications now utilize realistic 3D models to drive synthetic data generation.

Historically, high-quality object capture methodologies have required a certain level of controlled capture, such as a fixed camera rig, or specific imaging equipment, such as a depth sensor [29, 38]. Modernized pipelines driven by structure-from-motion (SfM) followed by multi-view stereo (MVS) and depthmap fusion [48, 47] have increasingly democratized the process in recent years, enabling less-experienced users to run photogrammetry from a handheld camera,

* This work was completed while Yoni was an intern at Meta.

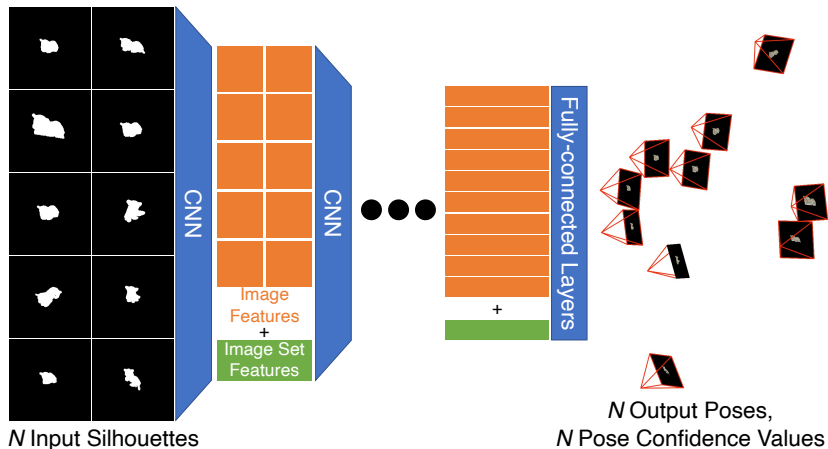


Fig. 1. Our deep neural network takes as input a set of silhouette masks of an object observed from different viewpoints. After applying several permutation-equivariant layers that combine image-specific and image-set-generic features, the network outputs a 6DOF pose and a pose confidence for each input image.

either with known temporal sequencing (*i.e.*, video capture) or capturing images as an unordered collection. These general-purpose pipelines also enable distributed collection, where photos from multiple users in different environments are leveraged to create a 3D model.

However, while casual 3D reconstruction is increasingly feasible, output reconstruction quality in these pipelines varies widely depending on the input imagery and target object, and there exist several categorical limitations, particularly for low-texture objects and objects having non-Lambertian surface reflectance properties. A number of approaches, for example the recent works of Yariv *et al.* [66] and Schmitt *et al.* [46], have pushed the envelope of dense surface estimation pipelines by jointly modeling object geometry, view-dependent lighting/reflectance effects, and – importantly – allowing for camera pose parameters to be refined as part of the optimization process, which is not generally possible with traditional MVS. These approaches have shown a remarkable improvement in completeness of the reconstructed object surface, as well as impressive quality for low-texture objects with complex appearance.

In this paper, we address a key remaining gap for state-of-the-art reconstruction methods: *camera pose initialization*, particularly when photometric methods fail. To this end, we introduce a neural-network-based alternative to SfM tackling the classical computer vision problem of multi-view pose from unknown object silhouettes [8, 7, 17, 31]. Given a set of binary object masks obtained from multiple images of an object, the goal is to produce a camera pose for each image, relative to an arbitrary, unspecified object coordinate frame. The driving concept here is that, when constraints like point correspondences cannot be utilized, the object contour in the image still provides signal on the space of possible relative camera poses. For example, each foreground pixel in one image has a correspond-

ing location in every other image, and thus the epipolar lines for those pixels must intersect the object silhouette in the other image. While previous work has wielded such principles using handcrafted features and/or controlled scenarios, our hypothesis is that a neural network should be able to naturally learn the joint space of camera viewpoints.

For the current work, we assume we are given a set of pre-extracted silhouette images with known camera intrinsic calibration, generally upright orientation, and medium-baseline camera motion. This image set is fed to an order-equivariant neural network (Fig. 1) that regresses a six-degrees-of-freedom (6DOF; *i.e.* rotation and translation) pose for each image simultaneously, as well as a confidence estimate that helps to identify images with higher levels of pose ambiguity. The network’s task during training is to learn how to map the object contours into a common latent representation while also taking into account the global state of all contours together, and then to form a mapping from this representation into a final 6DOF pose.

For training, we render randomly posed silhouettes of CAD models and directly optimize the network’s output to match the poses used for rendering. At test time, the network uses only the input silhouette images, without any knowledge about the 3D geometry of the observed object. Our method provides the following overall contributions:

- A deep-learning approach for silhouette-based multi-view 6DOF pose estimation for unknown objects. Previous works in this space have been very tailored to controlled settings or known objects [64], or have required carefully handcrafted features in a robust framework while only estimating 3DOF camera rotations [31].
- A neural network architecture leveraging DSS and DeepSets layers [33, 68] to achieve unordered multi-view pose estimation. Such architectures have not been previously used for this task. In our case, the selection of the output global coordinate system is arbitrary, and there is not just one “correct” solution, in contrast to previous applications of permutation-equivariant layers. We thus introduce a new loss function that is agnostic to the output global coordinate system (Eqs. (2-7)). This formulation is crucial for making the training problem possible.
- A loss formulation that incorporates the von Mises-Fisher distribution to allow for pose confidence regression. We demonstrate that our network’s confidence predictions reliably correspond to per-view pose accuracy results.
- Generalizability: While we train on only 15 object classes of CAD models, we show that the network generalizes to unseen object classes on a number of synthetic and real datasets, including datasets with imperfect masks.
- Putting silhouettes into practice: Considering the case of uncontrolled, unknown object capture, we demonstrate that silhouette-based reasoning offers a workable solution for low-texture objects where color-based reconstruction methods have inherent limitations. Examples are shown for a new “Glass Figurines” dataset, where our method succeeds in several challenging cases where a state-of-the-art SfM pipeline [48] fails.

2 Related Work

Camera pose estimation for object reconstruction has a long history in the field of computer vision. For unknown objects and unordered images, possibly the most well-established approaches are photogrammetric methods like Structure-from-Motion (SfM) [48]. These methods are driven by 2D feature correspondence search, where distinct 2D image keypoints are detected, described, and matched between the input images. Assuming that such 2D image-to-image correspondences can be reliably found, additional geometric reasoning is used to begin recovering 6DOF image poses. In contrast to incremental SfM methods that build the final 3D reconstruction one image at a time, the method we propose is more in line with global SfM approaches [12, 50] and recent holistic deep-learned approaches [37], where pose properties for all images are determined simultaneously. Typical global SfM methods rely on two-view pose estimates to initially solve for absolute image rotations, followed by a second stage to solve for absolute image positions. Recent work by Kasten *et al.* [25] has also suggested a one-step global approach by averaging essential matrices. While our neural network architecture does not leverage two-view relationships directly, it does employ a global representation of all images when deriving latent representations at different stages of the network.

It is also worth noting that many active-capture applications, for example object reconstruction pipelines that run on a smartphone [42, 53], augment the camera pose estimates with inertial measurement unit readings available on the device, which provide a strong prior for the differential motion of the camera. In our work, we assume a different capture scenario, where the object of interest may be moved between different frames, or even where the collection of object images is derived from different locations at different times. Moreover, all SfM-type methods heavily depend on the reconstructibility of the object of interest. For objects with low texture or complex appearance, these methods often fail because the photometric assumptions underlying keypoint detection and description are violated.

Learning-based methods for 3D object reasoning. A litany of methods have been proposed for camera pose detection for known objects, especially in single-view contexts. Early methods leveraged deformable parts models for discrete viewpoint prediction [13, 32, 40]. Related work [16] achieved 6DOF pose estimation via view synthesis with brute-force evaluation of geometry priors. With the advent of deep learning, numerous approaches have advanced single-view object detection and pose estimation, including for discrete prediction, 3D bounding box estimation, direct pose regression, and direct 2D-to-3D point correspondence regression [41, 62, 27, 43, 63, 54, 5, 2]. One recent extension to these works is HybridPose [51], which combines object pose regression with learned feature extraction and subsequent pose refinement. Also relevant to our work is SilhoNet [1], an object pose regressor that is trained to predict occlusion-aware and occlusion-agnostic object silhouette masks as an intermediate output. The silhouette is used as the primary cue for rotation, which allows the rotation regression submodule to train entirely on synthetic data.

Reconstruction-focused approaches have emphasized learning shape priors for object classes, especially for single-view geometry prediction. Choy *et al.* [4] trained a recurrent neural network for volumetric 3D reconstruction of multiple object classes. Beyond single-view shape estimation, this network is able to iteratively aggregate multiple images to refine the output, resulting in a coarse 3D model for instances of the trained-for classes. While this and related methods [10, 6, 61, 35] penalize errors in 3D geometry, prior work leveraging deformable shapes [3, 24] and subsequent works in differentiable projection and rendering have used object masks directly. Several methods [44, 65, 14, 56, 58, 26] train single-view volumetric or mesh reconstruction models by reprojecting voxels into other views and optimizing the predicted voxel occupancy against the ground-truth object mask. Some such methods have also reported results on 2 to 5 input views [44, 14, 58].

Several learning-based reconstruction methods exist that estimate camera pose for canonical object frames [70] or between image pairs [55, 20] with a silhouette-based loss. In the latter cases, a network is shown pair of images and jointly predicts (1) the relative pose between them and (2) a 3D geometry (a voxelization or a point cloud) for the object. The models are trained by reprojecting the predicted geometry into the first image and penalizing disagreements with the associated object mask. Each input image is independently processed, allowing for single-view applications of geometry and pose estimation at inference time.

A number of recent works learn a neural radiance field (NeRF) [36] while optimizing camera parameters [57, 30, 22, 67]. These methods either require camera initialization, or can only handle roughly forward-facing scenes. Very recently, [34] used a generative adversarial training strategy without input camera poses in a general camera setup with a known camera distribution. For each scene, they train from scratch a (NeRF, discriminator) network pair by sampling camera poses according to the distribution and training the NeRF to fool the discriminator for whether a patch is fake (rendered) or real. This training process is heavy, on the order of hours, and must be done separately for each scene without any generalizability to new scenes. While the majority of the work focuses on color image processing, the authors do present a single proof-of-concept result taking in a large collection of silhouette images as input. In contrast to NeRF methods, our approach trains a neural network that holistically processes image sets in a single pass and generalizes to unseen objects and object classes.

Pose from silhouettes of an unknown object. Methods for camera pose estimation from silhouettes date back more than two decades. Classical approaches utilize *epipolar mapping constraints*, where all epipolar silhouette lines mapped from one view must intersect the silhouette in another. For two views, epipolar constraints yield corresponding 2D object contour points with tangent epipolar lines. For multiple views, the constraints amount to finding a consistent visual hull for all images. Many early approaches optimized pose by identifying corresponding silhouette *frontier points*, either as single-take methods under controlled capture (*e.g.* a turntable or using mirrors) [59, 60, 7, 18, 19]

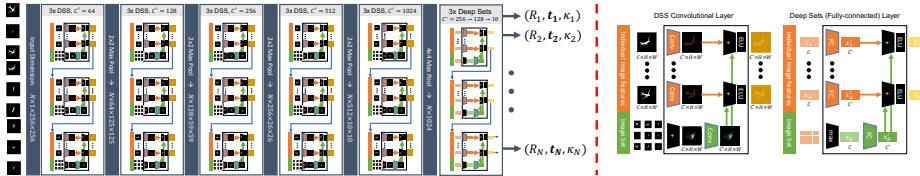


Fig. 2. Our permutation-equivariant network architecture. Starting from the N input silhouettes, five groups of three convolutional DSS layers are sequentially applied, interspersed by max-pooling operations. Then, three deep sets (fully-connected) layers are applied to finally obtain 10 output values per image, representing the image’s rotation, translation, and pose confidence. Each convolution is followed by a batch normalization layer (not shown), and we use ELU activations throughout the network.

or by optimizing a camera rig configuration under repeated observations [23, 8, 49]. Visual hull optimization [17] has also been proposed for controlled capture scenarios.

One similar work to ours is that of Littwin *et al.* [31], which aims to estimate the rotation distance between cameras using a handcrafted measure of silhouette contour similarity. While this two-view measure is quite noisy, the authors show that, given a sufficiently large source image set, an inlier set of relative rotation measurements can be determined via a robust fitting procedure. In contrast, our network considers all available images jointly when making its predictions, and it additionally can reason about camera translations and pose confidence.

Finally, Xiao *et al.* [64] trained a neural network to regress 6DOF pose for an novel object in an image under the assumption that the object geometry is known at inference time. Their approach first computes separate shape and appearance encodings and then feeds these to a pose regression sub-network. While the authors did not analyze the network’s activations, it is quite possible that their network learns to encode object contours in the input image and compare these to possible projections of the 3D shape.

3 Method

We assume an input set of N images taken by N cameras capturing the same 3D object at different viewpoints. We further assume that the object silhouette masks are pre-extracted from the input images. In practice, this can be done either by classical or learning-based approaches [15] for 2D object segmentation. Our goal is to regress the camera poses solely from the silhouette masks.

To tackle this problem, we introduce a deep neural network architecture that learns to infer a set of 6DOF camera poses and pose confidences from silhouettes using a large training set of general 3D objects. To improve robustness against two-view ambiguities, our network considers all N input silhouettes jointly. While we directly optimize pose error during training, we observe that our network outputs poses that respect the silhouette constraints leveraged by earlier non-learning-based methods (see supplementary).

3.1 Network Architecture

Our network architecture (Fig. 2) is based on the recently introduced “deep sets of symmetric elements” (DSS) layers [33], which have shown to be effective across a variety of learning tasks involving inputs of unordered image sets. The input for each DSS layer is a set of N images with the same number of channels. Two learnable convolutional filters are then applied: a Siamese filter that is applied on each input image independently, and an aggregation module filter that is applied on the sum of all input images. The output of the second filter is then added to each output of the first filter, resulting in a new set of N images with a possibly different number of channels. Since summation is a permutation-invariant operation, it follows that a DSS layer is permutation-equivariant, meaning that applying a permutation to the N input images results in permuted outputs.

In our case, since the multi-view input silhouettes are unordered, we design our network to be permutation-equivariant. The inputs to the network are N , one-channel $\{0, 1\}$ silhouette binary masks, and the outputs are the corresponding N camera poses, each represented by 10 coordinates: 6 for the world-to-camera rotation using the $6D$ parameterization of [69], 3 for the world-to-camera translation, and another scalar to represent the confidence of the network in the estimated pose. We use a sequence of 5 DSS blocks, each consisting of 3 DSS layers with max-pooling operations between each block, followed by permutation-equivariant fully connected (“deep sets”) layers as in [68].

3.2 Confidence-based Loss Function

We use ground-truth camera poses (available at training time) for training the network. Let $(R_1, \mathbf{t}_1), \dots, (R_N, \mathbf{t}_N)$ denote the output camera poses from the deep network and $(\bar{R}_1, \bar{\mathbf{t}}_1), \dots, (\bar{R}_N, \bar{\mathbf{t}}_N)$ the respective ground-truth (GT) camera poses. For each input silhouette image, we predict a single pose confidence $\kappa \in \mathbb{R}$ corresponding to the scale parameter of a von Mises-Fisher (vMF) distribution [52].

First, it is worth taking a moment to discuss possible formulations for the network loss function and confidence prediction. On one hand, we could forego confidence estimation entirely and directly penalize the rotation and translation errors using, *e.g.*, an L2 penalty. We empirically found this approach to give similar accuracy to our formulation, with the caveat that the network no provides quality ratings for individual pose estimates. Alternatively, we could adopt a complete probability distribution like the Bingham distribution on rotation [11], which can properly model uncertainty directions in the tangent space of $\text{SO}(3)$. In practice, however, we found that introducing more confidence parameters for rotation made the network more difficult to train. This may be caused in part by the fact that the full space of $\text{SO}(3)$ is much larger than that of our assumed input viewpoints, which are generally upright and always object-centric. As such, we have chosen to model a single confidence parameter for rotation alone, and we show in our experiments that this approach is effective in separating good-quality pose estimates from those that are less certain.

To model 3DOF camera rotation and its confidence, we adopt a maximum-likelihood formulation where we predict a 2DOF probability distribution mean and scale for each axis of the local camera frame. We define three vMF distributions that share the same scale parameter κ , each with a probability density function of

$$f_i(\mathbf{x}; \mathbf{r}_i, \kappa) = C_3(\kappa) e^{\kappa \mathbf{r}_i^T \mathbf{x}}, \quad (1)$$

where $\mathbf{r}_i \in \mathbb{S}^2$ for $i \in \{1, 2, 3\}$ are the (unit) row vectors of the predicted rotation matrix for the image, and $C_3(\kappa) = \frac{\kappa}{2\pi(e^\kappa - e^{-\kappa})}$ forms a normalization factor.

We aim to predict distributions that explain the GT rotation axes with as high of a probability as possible. Denoting the GT row as $\bar{\mathbf{r}}_i \in \mathbb{S}^2$, the log-likelihood for this vector to be sampled from the corresponding distribution is

$$l_i(R_0) = \log(C_3(\kappa)) + \kappa \mathbf{r}_i^T (R_0 \bar{\mathbf{r}}_i), \quad (2)$$

where R_0 is a global rotation ambiguity of our solution relative to the GT cameras.

For log-likelihood l_i^j of camera j , we can compute R_0 as

$$R_0^* = \operatorname{argmax}_{R_0} \sum_{j=1}^N \sum_{i=1}^3 l_i^j(R_0). \quad (3)$$

In the supplementary material, we show that this can simply done by weighted relative rotation averaging:

$$\tilde{\mathbf{q}}_0^* = \frac{1}{N} \sum_{j=1}^N \kappa_j \mathbf{q}_j^{-1} \bar{\mathbf{q}}_j, \quad \mathbf{q}_0^* = \frac{\tilde{\mathbf{q}}_0^*}{\|\tilde{\mathbf{q}}_0^*\|}, \quad (4)$$

where \mathbf{q}_i , $\bar{\mathbf{q}}_i$, and \mathbf{q}_0^* are the quaternions corresponding to R_i , \bar{R}_i , and R_0^* , respectively. The final loss function for the rotation and confidence outputs is

$$L_{R,\kappa} = \frac{1}{3N} \sum_{j=1}^N \sum_{i=1}^3 -l_i^j(R_0^*). \quad (5)$$

For our predicted translation vectors, a camera-center loss is applied by considering global translation and scaling ambiguities. Denoting $\mathbf{c}_i = -R_0^{*T} R_i^T \mathbf{t}_i$ and $\bar{\mathbf{c}}_i = -\bar{R}_i^T \bar{\mathbf{t}}_i$ as predicted and GT camera centers, respectively, the camera-center loss is defined by

$$L_c = \frac{1}{N} \sum_{i=0}^N \left\| \frac{\mathbf{c}_i - \mathbf{c}}{s} - \frac{\bar{\mathbf{c}}_i - \bar{\mathbf{c}}}{\bar{s}} \right\|, \quad (6)$$

where the mean vectors $\mathbf{c} = \frac{1}{N} \sum_{j=1}^N \mathbf{c}_j$, $\bar{\mathbf{c}} = \frac{1}{N} \sum_{j=1}^N \bar{\mathbf{c}}_j$ account for the global translation ambiguity, and we divide by the mean distance between each camera center and the average center: $s = \frac{1}{N} \sum_{j=1}^N \|\mathbf{c} - \mathbf{c}_j\|$, $\bar{s} = \frac{1}{N} \sum_{j=1}^N \|\bar{\mathbf{c}} - \bar{\mathbf{c}}_j\|$. Our total loss function is:

$$L = \beta L_c + L_{R,\kappa} \quad (7)$$

with scalar weight β balancing the two loss parts. In our experiments, we use $\beta = 2$, which was chosen based on examining the validation set error. L_c values are in $[0, 1]$, and values of $L_{R,\kappa}$ are typically around -10 .

3.3 Training

For training our network, we use a collection of synthetic object models from multiple object categories. For each training iteration, we render N 256×256 px silhouettes with random camera poses around the given object. Each input set of N images is generated with azimuth and elevation sampled uniformly in the range $[-30^\circ, 30^\circ]$, while the camera roll from the scene vertical is sampled from a normal distribution with a standard deviation of 5° . This viewing range is selected to approximate a typical set of casually captured viewpoints of one side of an object; for example, the DTU dataset [21] used in our experiments has a similar range of viewing angles. Each camera is initially positioned to look at the object origin (defined as the median vertex), with a distance from the origin sampled uniformly within the range $[3.2r, 6r]$, where r is the object radius. The camera translations are then perturbed by an offset sampled from $\mathcal{N}(\mathbf{0}, (0.005r)I_{3 \times 3})$. The object itself is rotated randomly around its origin.

In all experiments, we use $N = 10$ input views for training. This number was reported by [31] to be a large-enough support set in the multi-view setting. We trained the network by minimizing Eq. (7) on a training split of 15 object categories. We used the ADAM optimizer [28] with a learning rate of 0.001. To improve initial training acceleration, we began with a batch size of 5 (*i.e.*, 5 groups of 10 random poses) with all images coming from the same object. To better maintain training acceleration in later epochs, we switched to a batch size of 1 after ~ 60 epochs, and we trained overall for ~ 250 epochs.

4 Experiments

4.1 Datasets

We evaluated our trained network on 3D objects from a validation split of object models from 3D Warehouse. We tested unseen objects and camera configurations from our 15 training classes, plus 5 unseen object classes.

In addition to manually created models, we further evaluated the network on a new dataset, RealScan, that consists of 30 high-resolution scans of a variety of real 3D objects ranging from stuffed animals to office supplies. We projected these scans with the same sampling described in Sec. 3.3 for 100 sets of 20 random views, with each set of views coming from either the front, back, top, sides, or bottom of the object. These high-polygon meshes, as well as their projected contours, are very different from the ones that are used for training the network.

We further applied our method to real images from (1) the DTU MVS dataset [21] and (2) a new ‘‘Glass Figurines’’ dataset containing objects that are difficult to reconstruct using traditional SfM methods. For DTU, we evaluated the 15 back-row cameras (available for scans with id number > 80) whose cameras are far enough from the object such that most of the object is visible in the image. For the 8 scans, we used the input masks that were extracted manually by [66, 39]. The Glass Figurines dataset consists of 11 objects with 10 images each, plus manually extracted objects masks and ground-truth camera poses computed using ArUco Tags [45, 9]. We plan to publicly release the dataset.

Table 1. Camera pose accuracy and reprojection IOU for the 3D Warehouse dataset.

Metrics		R [°]	t_s [ratio]	t_d [°]	R [IOU]	t [IOU]	t_s [IOU]	t_d [IOU]	$R + t$ [IOU]
Average per-class validation statistic over all 15 training classes									
valid.	Mean (Med.)	6.40 (4.59)	0.03 (0.02)	2.36 (1.50)	0.84 (0.88)	0.70 (0.73)	0.92 (0.94)	0.70 (0.74)	0.46 (0.47)
	↑ 5 (Oracle)	5.17 (3.88)	0.03 (0.03)	2.17 (2.23)	0.86 (0.87)	0.71 (0.71)	0.93 (0.93)	0.72 (0.71)	0.49 (0.48)
	↓ 5 (Oracle)	7.63 (8.92)	0.04 (0.04)	2.55 (2.48)	0.82 (0.81)	0.68 (0.69)	0.91 (0.91)	0.69 (0.69)	0.42 (0.44)
Unseen test classes									
bathrub	Mean (Med.)	6.61 (4.69)	0.03 (0.02)	2.07 (1.33)	0.92 (0.95)	0.86 (0.88)	0.95 (0.96)	0.87 (0.89)	0.67 (0.74)
	↑ 5 (Oracle)	5.33 (4.13)	0.03 (0.03)	1.90 (1.92)	0.94 (0.94)	0.87 (0.87)	0.95 (0.96)	0.88 (0.88)	0.70 (0.69)
	↓ 5 (Oracle)	7.89 (9.09)	0.03 (0.04)	2.23 (2.22)	0.91 (0.90)	0.85 (0.85)	0.94 (0.94)	0.87 (0.87)	0.64 (0.65)
car	Mean (Med.)	6.12 (4.53)	0.03 (0.02)	2.27 (1.45)	0.92 (0.94)	0.84 (0.86)	0.94 (0.96)	0.85 (0.87)	0.62 (0.69)
	↑ 5 (Oracle)	4.92 (3.81)	0.03 (0.03)	1.99 (2.13)	0.93 (0.94)	0.85 (0.85)	0.95 (0.95)	0.86 (0.86)	0.66 (0.64)
	↓ 5 (Oracle)	7.32 (8.43)	0.03 (0.03)	2.55 (2.40)	0.90 (0.90)	0.83 (0.83)	0.94 (0.94)	0.84 (0.84)	0.58 (0.60)
chair	Mean (Med.)	6.79 (4.96)	0.03 (0.02)	2.67 (1.70)	0.83 (0.87)	0.74 (0.78)	0.91 (0.94)	0.75 (0.79)	0.49 (0.51)
	↑ 5 (Oracle)	5.38 (4.14)	0.03 (0.03)	2.38 (2.52)	0.85 (0.87)	0.76 (0.75)	0.92 (0.92)	0.76 (0.76)	0.53 (0.51)
	↓ 5 (Oracle)	8.19 (9.43)	0.04 (0.04)	2.97 (2.83)	0.80 (0.78)	0.73 (0.73)	0.90 (0.91)	0.74 (0.74)	0.45 (0.47)
lamp	Mean (Med.)	10.60 (7.26)	0.04 (0.03)	3.57 (2.24)	0.77 (0.83)	0.63 (0.68)	0.89 (0.93)	0.64 (0.69)	0.32 (0.27)
	↑ 5 (Oracle)	9.19 (6.57)	0.04 (0.04)	3.48 (3.51)	0.79 (0.80)	0.65 (0.65)	0.90 (0.90)	0.66 (0.66)	0.34 (0.33)
	↓ 5 (Oracle)	12.00 (14.62)	0.05 (0.05)	3.66 (3.63)	0.75 (0.73)	0.61 (0.62)	0.88 (0.88)	0.62 (0.63)	0.30 (0.31)
mailbox	Mean (Med.)	11.15 (5.13)	0.06 (0.03)	3.98 (2.12)	0.82 (0.88)	0.73 (0.78)	0.93 (0.95)	0.74 (0.78)	0.36 (0.30)
	↑ 5 (Oracle)	8.98 (7.56)	0.05 (0.05)	3.49 (3.42)	0.84 (0.86)	0.74 (0.75)	0.94 (0.93)	0.75 (0.76)	0.38 (0.38)
	↓ 5 (Oracle)	13.32 (14.73)	0.07 (0.07)	4.47 (4.54)	0.81 (0.79)	0.72 (0.71)	0.93 (0.93)	0.73 (0.71)	0.33 (0.33)

4.2 Results

Camera pose accuracy results are presented for the 3D Warehouse dataset in Table 1 and for the RealScan dataset in Table 2. For both, we evaluate on 10 random views of the object per test instance. Due to space limitations, we only show a representative subset of the RealScan results, and for 3D Warehouse, we show the average per-class validation result across all 15 training classes, and for our 5 unseen testing classes. See our supplementary material for complete results. In each row, we show mean and median errors, plus a confidence-ordered breakdown of the mean error, for a variety of metrics. All “Top 5” and “Bottom 5” metrics are taken using our confidence ranking from highest to lowest; we also show “Oracle” rankings for these that consider the ordering of lowest rotation error to higher rotation error. (The oracle ordering is the same for all columns.) The oracle provides a lower bound on the Top-5 error and thus can be used to assess the effectiveness of our confidence predictions.

When our confidence output is near to or better than the oracle, this indicates that our network has learned a reasonable confidence for pose. We also report intersection-over-union (IOU), computed by re-rendering the test object using our predicted poses after global alignment to the GT poses. An example IOU result is shown in Fig. 3, along with a visualization of the visual hull for our estimated poses.

In Tables 1 and 2, we report our mean rotation (R), translation-scale (t_s), and translation-direction (t_d) error. t_s is the absolute value of: one minus the magnitude ratio of the predicted and GT translation vectors. t_d is the angle between the predicted and GT translation vectors. Also in of Table 1, we isolate the different network outputs: the sixth column shows our rotation combined with GT translation, the next our translation with GT rotation, and so on.

Table 2. Pose accuracy for a representative subset of RealScan.

	Metrics	R [$^{\circ}$]	t_s [ratio]	t_d [$^{\circ}$]
		mean median	mean median	mean median
Cheetah	Mean (Med.)	8.78 (6.43)	0.04 (0.03)	3.37 (2.57)
	\uparrow 5 (Oracle)	7.14 (5.50)	0.03 (0.03)	2.80 (3.13)
	\downarrow 5 (Oracle)	10.42 (12.06)	0.04 (0.04)	3.94 (3.62)
Chess Knight	Mean (Med.)	10.29 (8.01)	0.04 (0.03)	4.84 (3.61)
	\uparrow 5 (Oracle)	8.12 (5.99)	0.04 (0.04)	4.97 (4.64)
	\downarrow 5 (Oracle)	12.47 (14.60)	0.05 (0.04)	4.70 (5.03)
Glasses	Mean (Med.)	9.15 (7.57)	0.05 (0.03)	4.39 (3.00)
	\uparrow 5 (Oracle)	7.45 (5.95)	0.04 (0.04)	3.91 (4.23)
	\downarrow 5 (Oracle)	10.85 (12.35)	0.05 (0.06)	4.87 (4.56)
Plastic Cup	Mean (Med.)	15.33 (12.96)	0.05 (0.04)	6.45 (5.41)
	\uparrow 5 (Oracle)	12.84 (9.70)	0.05 (0.05)	6.50 (6.75)
	\downarrow 5 (Oracle)	17.82 (20.96)	0.05 (0.05)	6.41 (6.16)
Stapler	Mean (Med.)	8.88 (4.68)	0.03 (0.02)	2.77 (1.81)
	\uparrow 5 (Oracle)	7.44 (5.50)	0.03 (0.03)	2.43 (2.70)
	\downarrow 5 (Oracle)	10.31 (12.25)	0.03 (0.03)	3.10 (2.83)
Toy Bunny	Mean (Med.)	8.29 (5.92)	0.03 (0.02)	2.92 (1.95)
	\uparrow 5 (Oracle)	6.06 (4.51)	0.03 (0.03)	2.78 (2.80)
	\downarrow 5 (Oracle)	10.52 (12.08)	0.03 (0.03)	3.06 (3.04)
Wooden Spoon	Mean (Med.)	10.04 (8.18)	0.05 (0.03)	3.52 (2.57)
	\uparrow 5 (Oracle)	7.68 (5.76)	0.04 (0.04)	3.51 (3.65)
	\downarrow 5 (Oracle)	12.40 (14.32)	0.05 (0.05)	3.53 (3.40)

Table 3. Camera pose accuracy for the DTU dataset.

Id	R [$^{\circ}$]		t_s [ratio]		t_d [$^{\circ}$]	
	mean	median	mean	median	mean	median
83	4.69	4.11	0.02	0.02	3.63	3.73
97	16.44	16.73	0.11	0.08	9.54	8.72
105	4.35	4.16	0.01	0.01	2.57	2.37
106	6.20	5.08	0.02	0.02	1.78	1.84
110	4.59	3.43	0.02	0.01	0.68	0.68
114	3.13	3.12	0.02	0.02	0.78	0.66
118	7.17	6.34	0.03	0.02	5.55	4.89
122	9.76	8.74	0.03	0.03	6.23	5.43

Concerning the results themselves, we observe that we obtain consistent generalization from our training data to our unseen test classes and more realistic object scans. For many objects, rotation error is around 8° on average, with a substantially lower median error, and we observe a similar error distribution for validation and test instances (Fig. 4). We also observe generally low translation errors, and that our confidence ranking is consistently able to achieve rotation errors within a few degrees of the oracle. This ranking is also on par with the oracle in the IOU metrics. From the IOU metrics, we also see that our rotation estimates are generally high quality, achieving 80-90% IOU for nearly all test cases. Translation fares slightly worse, especially for the direction estimate, for which the IOU metric is very sensitive. We observe lower IOUs for both rotation and translation (rightmost column), which is expected since it reflects the full network output. See the supplementary for additional RealScan visualizations.

Qualitatively, our network understandably performs worse for objects with rotational symmetry, for example the 3D Warehouse lamps and the RealScan plastic cup. In the latter case, while the cup has a handle, this handle is not always visible in the input images, and so an unambiguous pose estimate cannot be determined. IOU is also a conservative metric for pose estimation, especially for thin structures like the RealScan eyeglasses and spoon, because even with a perfect rotation estimate, a small amount of translation error can cause the reprojection to shift considerably.

As for real-world datasets, results for the DTU and Glass Figurines datasets are presented in Tables 3 and 4, respectively. Different from the previous experiments, we provided our network with all 15 DTU images as input. We observe low rotation and translation errors for the majority of the objects in both datasets.

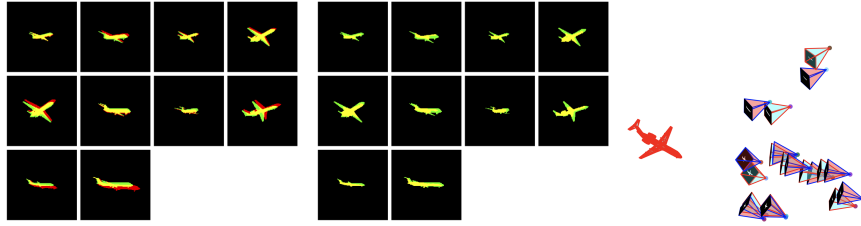


Fig. 3. Estimated poses for 10 silhouette masks of an airplane. Left: Object rejections by our cameras (red) versus the original input masks (green), ordered from from greatest confidence (top left) to least (bottom right). Middle: Visual hull projection for our method (yellow) versus the original input masks (green), with the same ordering. Right: Predicted poses (blue frusta) relative to GT poses (red frusta), with the target object mesh in red.

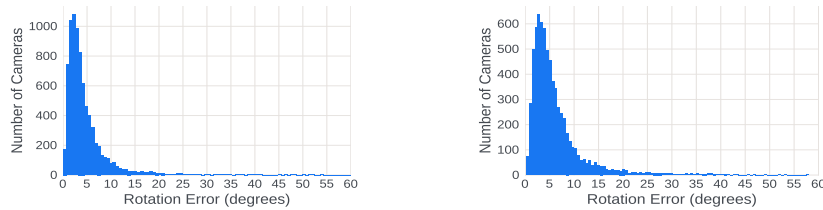


Fig. 4. Histogram of rotation errors for all test instances of airplanes (left, training class) and cars (right, unseen class).

Comparison to reconstruction methods. For the Glass Figurines dataset, we also compare our method to GNeRF [34], a recent deep method that can optimize per-scene camera poses without accurate initialization. While GNeRF is built for color images, the authors also showed an example result on silhouettes. We evaluate both masked color images and silhouettes in Table 4. For this dataset, GNeRF performs much worse in pose estimation. This is understandable due to the limited size (10 images) and pose distribution of each image set. The NeRF is accordingly unable to generalize to novel viewpoints, especially for silhouettes where cross-view occupancy constraints must be leveraged. GNeRF also must train on a single image set at a time and takes hours to converge. In contrast, our network runs in a single pass without any additional training.

We also note in Table 4 whether COLMAP’s SfM algorithm [48] could process the masked color images. When COLMAP succeeded, its poses tended to be very accurate (see supplementary). However, due to the lack of a consistent object appearance or background, COLMAP failed to reconstruct 5 of the 11 objects.

Pairwise angular distances. We conducted a small experiment to compare our method against the method of Littwin *et al.* [31], which is the only method we are aware of that can jointly estimate multi-view camera poses (albeit only relative rotations) for a collection of causally captured object silhouettes. We unfortunately were unable to obtain a copy of their implementation or data, and

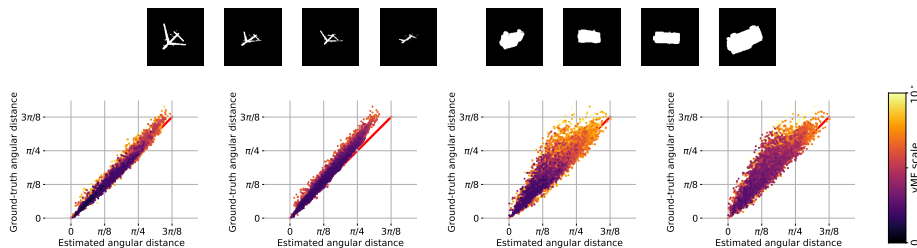


Fig. 5. Predicted versus ground-truth two-view angular distances for a validation class instance (airplane, left two columns) and an unseen class instance (car, right two columns), both from top-down viewpoints. Each point represents an image pair from 100 total images of the object, colored according to the image in the pair with the higher uncertainty (vMF angular spread at 95% of the vMF CDF). Top row: Example silhouettes. Second row: (1, 3) Result from running our network on 2000 samples of 10 images, showing the average error and lowest confidence per pair. (2, 4) Result from running our network once with all 100 images as input.


so we instead provide a qualitative comparison of Fig. 5 versus Fig. 2 in [31]. In the first and third graphs in Fig. 5, we have rendered 100 images of an object and from this sampled 2000 sets of 10 images. We plot the average estimated angular distance over all samples in which that pair appeared together, and we compare this to the ground-truth distance. Compared to [31], our estimates are much less noisy, and they match the ground truth with at least as much accuracy as [31] for a novel class.

We also show our confidence estimates in Fig. 5 and observe that they correlate to prediction accuracy and ground-truth distance, with nearer relative poses having higher confidence. Although confidence κ (Eq. (1)) is difficult to interpret directly in our loss formulation, we provide a rough sense of its scale by converting to an angular “spread” of the vMF distribution. Specifically, we consider the vMF CDF and, for a given value of κ , compute the angle $\arccos(\mathbf{r}_i^T \mathbf{x})$ that covers 95% of the distribution over the surface of the sphere. Put more simply, a darker color in the plot indicates a tighter distribution and higher confidence.

Many network inputs. As evidenced by our DTU experiments, our network generalizes to more inputs than it was trained on. In second and fourth graphs in Fig. 5, we take this to the extreme and provide our network with all 100 views of the object. Surprisingly, our network easily handles this configuration, producing similar error distributions to our 10-image samples. Our confidence predictions also have a qualitatively higher sensitivity in this scenario.

Additional results. We include a number of experiments in our supplementary, including complete results on the 3D Warehouse, RealScan, and Glass Figurines datasets; images of the RealScan IOU errors; and a visualization of the network satisfying epipolar constraints even for a failure case. We also include qualitative results on two real-world scenarios of a single object photographed in different environments: (1) a transparent swan sculpture with manually segmented masks, and (2) a chair with masks segmented via Mask R-CNN [15]. The

Table 4. Example images and mean camera pose accuracy for the Glass Figurines dataset. We compare our method to GNeRF [34] with color images and with silhouette inputs ([34]-S), and we note if SfM [48] succeeded or failed for the dataset. GNeRF failures are marked with dashes. See the supplementary for more information.



Object	SfM [48]	R [°]			t_s [ratio]			t_d [°]		
		[34]	[34]-S	Ours	[34]	[34]-S	Ours	[34]	[34]-S	Ours
brown sq.	✓	18.71	–	3.74	0.04	–	0.03	7.12	–	2.59
dog (c.)	✓	14.00	18.43	6.92	0.13	0.12	0.04	3.02	2.19	4.88
dog (p.)	✗	17.56	–	8.14	0.05	–	0.02	9.36	–	7.32
dolphin	✗	24.70	–	5.10	0.03	–	0.03	7.52	–	3.01
flamingo	✓	21.05	–	6.13	0.10	–	0.05	2.97	–	2.02
flower	✗	–	17.27	3.94	–	0.06	0.02	–	16.79	3.33
frog	✓	21.66	15.48	3.59	0.02	0.07	0.02	2.40	11.41	0.91
parrot	✓	27.40	11.92	21.16	0.05	0.04	0.03	7.24	8.98	1.46
penguin	✗	20.04	–	9.09	0.03	–	0.04	3.38	–	1.97
rabbit	✗	25.32	20.75	5.38	0.02	0.05	0.02	6.38	1.96	3.21
snake	✓	29.85	24.59	11.29	0.05	0.05	0.02	7.85	4.24	1.90

latter case is a promising example of providing automatically extracted masks to our network. Finally, we include three proof-of-concept results of IDR [66] applied to our pose estimates for DTU. These results indicate that our approach has sufficient accuracy to initialize state-of-the-art reconstruction methods.

5 Conclusion

The experimental results above support our hypothesis that neural networks can be trained to regress relative pose information, as well as pose confidences, for a given set of silhouette images of an unknown object. Our network model generalizes well to novel object classes and from the synthetic to the real domain. Although we train on a fixed number of 10 images, we observe that our network can capably regress poses for many more inputs at a time.

While the benefit of silhouette constraints for pose estimation has long been recognized, our work shows that silhouette cues on their own can effectively initialize pose estimates for state-of-the-art 3D reconstruction methods on untextured objects. Our work also suggests that permutation-equivariant processing may prove to be an invaluable tool in many-view object reconstruction pipelines, and that multi-view reasoning in neural networks (*e.g.*, aggregating features over all inputs in our pipeline) can yield more-robust estimates compared to two-view methods for 6DOF pose regression, particularly if confidence is also captured.

One limitation of our current work is its reliance on pre-segmented masks. Since our network takes masks as input, however, it could be integrated into an end-to-end pipeline that starts with an object segmentation network, applies a silhouette-based pose estimation, and then performs additional color-image-based pose refinement. Other future work includes leveraging symmetries and texture to resolve silhouette ambiguities when they arise. Also, while we show promising results on medium-baseline views, more work is needed to achieve full generalization w.r.t rotation, *e.g.*, in scenarios where object is viewed from opposite sides, or where the images have substantial relative roll.

References

1. Billings, G., Johnson-Roberson, M.: SilhoNet: An RGB method for 6d object pose estimation. *IEEE Robotics and Automation Letters* **4**(4), 3727–3734 (2019)
2. Cai, M., Reid, I.: Reconstruct locally, localize globally: A model free method for object pose estimation. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 3153–3163 (2020)
3. Cashman, T.J., Fitzgibbon, A.W.: What shape are dolphins? Building 3d morphable models from 2d images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **35**(1), 232–244 (2012)
4. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In: *European Conference on Computer Vision (ECCV)*. pp. 628–644. Springer (2016)
5. Do, T.T., Pham, T., Cai, M., Reid, I.: Real-time monocular object instance 6d pose estimation. In: *British Machine Vision Conference (BMVC)* (2019)
6. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 605–613 (2017)
7. Forbes, K., Nicolls, F., De Jager, G., Voigt, A.: Shape-from-silhouette with two mirrors and an uncalibrated camera. In: *European Conference on Computer Vision (ECCV)*. pp. 165–178. Springer (2006)
8. Forbes, K., Voigt, A., Bodika, N., et al.: Using silhouette consistency constraints to build 3d models. In: *Pattern Recognition Association of South Africa (PRASA)*. pp. 33–38 (2003)
9. Garrido-Jurado, S., Munoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R.: Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern Recognition* **51**, 481–491 (2016)
10. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: *European Conference on Computer Vision (ECCV)*. pp. 484–499. Springer (2016)
11. Glover, J., Popovic, S.: Bingham procrustean alignment for object detection in clutter. In: *International Conference on Intelligent Robots and Systems*. pp. 2158–2165. IEEE (2013)
12. Govindu, V.M.: Combining two-view constraints for motion estimation. In: *Computer Vision and Pattern Recognition (CVPR)*. vol. 2, pp. II–II. IEEE (2001)
13. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: *European Conference on Computer Vision (ECCV)*. pp. 408–421. Springer (2010)
14. Gwak, J., Choy, C.B., Chandraker, M., Garg, A., Savarese, S.: Weakly supervised 3d reconstruction with adversarial constraint. In: *International Conference on 3D Vision (3DV)*. pp. 263–272. IEEE (2017)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: *International Conference on Computer Vision (ICCV)*. pp. 2961–2969 (2017)
16. Hejrati, M., Ramanan, D.: Analysis by synthesis: 3d object recognition by object reconstruction. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 2449–2456 (2014)
17. Hernández, C., Schmitt, F., Cipolla, R.: Silhouette coherence for camera calibration under circular motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **29**(2), 343–349 (2007)

18. Huang, P.H., Lai, S.H.: Contour-based structure from reflection. In: *Computer Vision and Pattern Recognition (CVPR)*. vol. 1, pp. 379–386. IEEE (2006)
19. Huang, P.H., Lai, S.H.: Silhouette-based camera calibration from sparse views under circular motion. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1–8. IEEE (2008)
20. Insafutdinov, E., Dosovitskiy, A.: Unsupervised learning of shape and pose with differentiable point clouds. In: *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 2802–2812 (2018)
21. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 406–413. IEEE (2014)
22. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5846–5854 (2021)
23. Joshi, T., Ahuja, N., Ponce, J.: Structure and motion estimation from dynamic silhouettes under perspective projection. In: *International Conference on Computer Vision (ICCV)*. pp. 290–295. IEEE (1995)
24. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1966–1974 (2015)
25. Kasten, Y., Geifman, A., Galun, M., Basri, R.: Algebraic characterization of essential matrices and their averaging in multiview settings. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)*
26. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 3907–3916 (2018)
27. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D making RGB-based 3d detection and 6d pose estimation great again. In: *International Conference on Computer Vision (ICCV)*. pp. 1521–1529 (2017)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
29. Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., et al.: The digital Michelangelo project: 3d scanning of large statues. In: *Conference on Computer Graphics and Interactive Techniques*. pp. 131–144 (2000)
30. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405* (2021)
31. Littwin, E., Averbuch-Elor, H., Cohen-Or, D.: Spherical embedding of inlier silhouette dissimilarities. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 3855–3863 (2015)
32. López-Sastre, R.J., Tuytelaars, T., Savarese, S.: Deformable part models revisited: A performance evaluation for object category pose estimation. In: *International Conference on Computer Vision (ICCV) Workshops*. pp. 1052–1059. IEEE (2011)
33. Maron, H., Litany, O., Chechik, G., Fetaya, E.: On learning sets of symmetric elements. In: *International Conference on Machine Learning (ICML)* (2020)
34. Meng, Q., Chen, A., Luo, H., Wu, M., Su, H., Xu, L., He, X., Yu, J.: GNeRF: GAN-based Neural Radiance Field without Posed Camera. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021)
35. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 4460–4470 (2019)

36. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
37. Moran, D., Koslowsky, H., Kasten, Y., Maron, H., Galun, M., Basri, R.: Deep permutation equivariant structure from motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5976–5986 (October 2021)
38. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: International Symposium on Mixed and Augmented Reality (ISMAR). pp. 127–136. IEEE (2011)
39. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Computer Vision and Pattern Recognition (CVPR) (2020)
40. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Teaching 3d geometry to deformable part models. In: Computer Vision and Pattern Recognition (CVPR). pp. 3362–3369. IEEE (2012)
41. Poirson, P., Ammirato, P., Fu, C.Y., Liu, W., Kosecka, J., Berg, A.C.: Fast single shot detection and pose estimation. In: International Conference on 3D Vision (3DV). pp. 676–684. IEEE (2016)
42. Prisacariu, V.A., Kähler, O., Murray, D.W., Reid, I.D.: Simultaneous 3d tracking and reconstruction on a mobile phone. In: International Symposium on Mixed and Augmented Reality (ISMAR). pp. 89–98. IEEE (2013)
43. Rad, M., Lepetit, V.: BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: International Conference on Computer Vision (ICCV). pp. 3828–3836 (2017)
44. Rezende, D.J., Eslami, S., Mohamed, S., Battaglia, P., Jaderberg, M., Heess, N.: Unsupervised learning of 3d structure from images. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 29, pp. 4996–5004 (2016)
45. Romero-Ramirez, F.J., Muñoz-Salinas, R., Medina-Carnicer, R.: Speeded up detection of squared fiducial markers. *Image and vision Computing* **76**, 38–47 (2018)
46. Schmitt, C., Donne, S., Riegler, G., Koltun, V., Geiger, A.: On joint estimation of pose, geometry and svbrdf from a handheld scanner. In: Computer Vision and Pattern Recognition (CVPR). pp. 3493–3503 (2020)
47. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV). pp. 501–518. Springer (2016)
48. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
49. Sinha, S.N., Pollefeys, M., McMillan, L.: Camera network calibration from dynamic silhouettes. In: Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. I–I. IEEE (2004)
50. Sinha, S.N., Steedly, D., Szeliski, R.: A multi-stage linear approach to structure from motion. In: European Conference on Computer Vision (ECCV). pp. 267–281. Springer (2010)
51. Song, C., Song, J., Huang, Q.: HybridPose: 6d object pose estimation under hybrid representations. In: Computer Vision and Pattern Recognition (CVPR). pp. 431–440 (2020)
52. Sra, S.: Directional statistics in machine learning: a brief review. *Applied Directional Statistics: Modern Methods and Case Studies* p. 225 (2018)

53. Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., Pollefeys, M.: Live metric 3d reconstruction on mobile phones. In: International Conference on Computer Vision (ICCV). pp. 65–72 (2013)
54. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6d object pose prediction. In: Computer Vision and Pattern Recognition (CVPR). pp. 292–301 (2018)
55. Tulsiani, S., Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In: Computer Vision and Pattern Recognition (CVPR). pp. 2897–2905 (2018)
56. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Computer Vision and Pattern Recognition (CVPR). pp. 2626–2634 (2017)
57. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF—: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
58. Wiles, O., Zisserman, A.: Silnet: Single-and multi-view reconstruction by learning from silhouettes. In: British Machine Vision Conference (BMVC) (2017)
59. Wong, K.Y., Cipolla, R.: Structure and motion from silhouettes. In: International Conference on Computer Vision (ICCV). vol. 2, pp. 217–222. IEEE (2001)
60. Wong, K.Y., Cipolla, R.: Reconstruction of sculpture from its profiles with unknown camera positions. IEEE Transactions on Image Processing **13**(3), 381–389 (2004)
61. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: MarrNet: 3d shape reconstruction via 2.5d sketches. In: Advances in neural information processing systems (NeurIPS). pp. 540–550 (2017)
62. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: European Conference on Computer Vision (ECCV). pp. 365–382. Springer (2016)
63. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
64. Xiao, Y., Qiu, X., Langlois, P.A., Aubry, M., Marlet, R.: Pose from shape: Deep pose estimation for arbitrary 3d objects. In: British Machine Vision Conference (BMVC) (2019)
65. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: Advances in neural information processing systems (NeurIPS). pp. 1696–1704 (2016)
66. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Ronen, B., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33 (2020)
67. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: iNeRF: Inverting neural radiance fields for pose estimation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021)
68. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. In: Advances in neural information processing systems (NeurIPS). pp. 3391–3401 (2017)
69. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Computer Vision and Pattern Recognition (CVPR) (2019)

70. Zhu, R., Kiani Galoogahi, H., Wang, C., Lucey, S.: Rethinking reprojection: Closing the loop for pose-aware shape reconstruction from a single image. In: International Conference on Computer Vision (ICCV). pp. 57–65 (2017)