Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields

Rohith Agaram[1] Shaurya Dewan[1] Rahul Sajnani[2] Adrien Poulenard[3]
Madhava Krishna[1] Srinath Sridhar[2]
[1]RRC, IIIT-Hyderabad [2]Brown University [3]Stanford University
ivl.cs.brown.edu/projects/canonicalfields

Figure 1. We present Canonical Field Network (**CaFi-Net**), a self-supervised method for 3D position and orientation (pose) canonicalization of objects represented as neural fields. We specifically focus on neural radiance fields (NeRFs) fitted to raw RGB images of arbitrarily posed objects (left), and obtain a **canonical field** (fixed novel view shown on right) where all objects in a category are consistently positioned and oriented.

Abstract

*Coordinate-based implicit neural networks, or neural fields, have emerged as useful representations of shape and appearance in 3D computer vision. Despite advances, however, it remains challenging to build neural fields for categories of objects without datasets like ShapeNet that provide "canonicalized" object instances that are consistently aligned for their 3D position and orientation (pose). We present Canonical Field Network (**CaFi-Net**), a self-supervised method to canonicalize the 3D pose of instances from an object category represented as neural fields, specifically neural radiance fields (NeRFs). CaFi-Net directly learns from continuous and noisy radiance fields using a Siamese network architecture that is designed to extract equivariant field features for category-level canonicalization. During inference, our method takes pre-trained neural radiance fields of novel object instances at arbitrary 3D pose and estimates a canonical field with consistent 3D pose across the entire category. Extensive experiments on a new dataset of 1300 NeRF models across 13 object categories show that our method matches or exceeds the performance of 3D point cloud-based methods.*

1. Introduction

Neural fields [59]—coordinate-based neural networks that implicitly parameterize signals—have recently gained significant attention as representations of 3D shape [6, 22, 28], view-dependent appearance [24, 42], and motion [25]. In particular, neural radiance fields (NeRFs) [24], have been successfully used in problems such as novel view synthesis [3, 4, 66], scene geometry extraction [55, 61], capturing dynamic scenes [19, 29, 30, 33, 52], 3D semantic segmentation [53, 68], and robotics [1, 13, 21].

Despite the progress, it remains challenging to build neural fields that represent an entire category of objects. Previous methods sidestep the problem by overfitting on a single instance [24], or learning [22, 28, 63] on datasets like ShapeNet [5] that contain objects that are manually **canonicalized** – oriented consistently for 3D position and orientation (3D pose) across a category. This strong supervision makes it easier to learn over categories, but limits their application to data that contain these labels. Recent work has proposed methods for self-supervised learning of 3D pose canonicalization [40, 43, 47], however, these operate on 3D point clouds, meshes, or voxels – but not neural fields.

In this paper, we present Canonical Field Network (**CaFi-Net**), a self-supervised method for category-level canonicalization of the 3D position and orientation of objects represented as **neural fields**, specifically neural radiance fields. Canonicalizing neural fields is challenging because, unlike 3D point clouds or meshes, neural fields are **continuous**, noisy, and hard to manipulate since they are parameterized as the weights of a neural network [60]. To address these challenges, we first extend the notion of **equiv-**

ariance to continuous vector fields and show how networks for processing 3D point clouds [50] can be extended to operate directly on neural radiance fields. We design CaFi-Net as a Siamese network that contains layers to extract equivariant features directly on vector fields. These field features are used to learn a canonical frame that is consistent across instances in the category. During inference, our method takes as input neural radiance fields of object instances from a category at arbitrary pose and estimates a **canonical field** that is consistent across the category. To handle noise in radiance fields from NeRF, our method incorporates density-based feature weighting and foreground-background clustering.

Our approach learns canonicalization without any supervision labels on a new dataset of 1300 pre-trained NeRF models of 13 common ShapeNet categories in arbitrary 3D pose (see Figure 1). We introduce several self-supervision loss functions that encourage the estimation of a consistent canonical pose. In addition, we present extensive quantitative comparisons with baselines and other methods on standardized canonicalization metrics [37] over 13 object categories. In particular, we show that our approach matches or exceeds the performance of 3D point cloud-based methods. This enables the new capability of directly operating on neural fields rather than converting them to point clouds for canonicalization. To sum up, we contribute:

- Canonical Field Network (**CaFi-Net**), the first method for self-supervised canonicalization of the 3D position and orientation (pose) of objects represented as neural radiance fields.
- A Siamese neural network architecture with equivariant feature extraction layers that are designed to directly operate on continuous and noisy radiance fields from NeRF.
- A public dataset of 1300 NeRF models from 13 ShapeNet categories including posed images, and weights for evaluating canonicalization performance.

2. Related Work

We focus our review of related work on neural fields, supervised canonicalization, equivariant neural network architectures, and self-supervised canonicalization.

Neural Fields: Neural fields are emerging as useful representations for solving problems such as novel view synthesis [3, 4, 15, 24, 66], shape encoding and reconstruction [22, 28], dynamic reconstruction [19, 29, 30], appearance modeling [23, 36, 45], and human motion modeling [8, 18, 46, 51]. Generalization of neural fields to object categories remains difficult, but some methods have used pre-canonicalized datasets to circumvent this problem [63]. Neural fields are also used outside of computer vision, for instance in reconstructing proteins [69], physics [35], or audio [11]. Please see [49, 59] for more details. In this paper, our focus is on canonicalizing for the 3D pose of neural fields, specifically, neural radiance fields (NeRF) [24].

Supervised Canonicalization: Datasets such as ShapeNet [5] and ModelNet40 [58] have 3D shapes that are manually pre-canonicalized. This inductive bias aids category-level generalization in problems such as 3D reconstruction [12, 22, 28, 48]. We can also use these datasets to formulate canonicalization as a supervised learning problem [54] enabling applications such as 6 degree-of-freedom object pose estimation, multi-view reconstruction [17, 44], and reconstruction of articulating [20, 65] and non-rigid objects [64]. However, our goal is to canonicalize without using manual pose labels.

Equivariant Neural Networks: Pose-equivariant networks are equivariant to input pose [9, 50, 56, 57] by design. Some of these methods use Spherical Harmonic functions [10, 50, 56, 57], or vector neurons [9] to extract equivariant features. Equivariance is closely related to canonicalization since a canonical pose is also equivariant. Thus, previous methods have used pose-equivariance for canonicalization [37, 41, 43, 47]. However, these methods have thus far been limited to 3D point clouds, voxels, or meshes.

Self-Supervised Canonicalization: Recent research has shown that self/weak supervision is sufficient for learning pose canonicalization on point clouds [37, 38, 43, 47], 2D key-points [26], and images [27]. None of these previous self-supervised methods can operate directly on neural fields – to the best of our knowledge, ours is the first.

3. Background

We first provide background on the essential components and key ideas behind our method.

Neural Radiance Fields (NeRF): Given posed RGB images, NeRF and its variants [23, 24] learn a neural network to synthesize novel views of complex scenes. At inference, for any 3D location and viewing direction, it estimates a direction-independent density value σ, and a direction-dependent color c. Although designed for novel view synthesis, NeRF implicitly learns the 3D geometry and view-dependent appearance of scenes making it a useful representation of shape and appearance. While NeRFs excel at fitting a single object instance or scene, it struggles to generalize to object categories partly because objects can be in arbitrary 3D poses – thus, existing methods use pre-canonicalized datasets [63]. In this paper, we provide a way to canonicalize 3D pose without any supervision for objects represented as neural radiance fields. While our method is aimed at NeRFs, it can be extended to the 3D pose canonicalization of any neural field.

Equivariance and Canonicalization: A function Γ over x is said to be **equivariant** to a group operation $\mathcal{O}$ if its output changes by a fixed mapping M for any $\mathcal{O}$ operating on

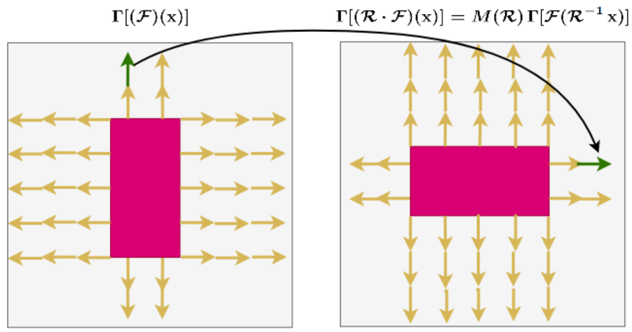

Figure 2. **Rotation Equivariance in Vector Fields.** Rotation equivariance in vector fields $\mathcal{F}$ (here, a simulated 2D magnetic field) require two operations: (1) update the position of the green arrow (the vector), and (2) rotating the green arrow at its new position. A function Γ operating on $\mathcal{F}$ is rotationally equivariant over the field if and only if $\Gamma[(\mathcal{R}\cdot\mathcal{F})(\mathbf{x})] = M(\mathcal{R})\,\Gamma[\mathcal{F}(\mathcal{R}^{-1}\mathbf{x})]$, where $\mathcal{R}$ is an $SO(2)$ rotation, and M is a rotation-dependent mapping function.

the input x, *i.e.*, $\Gamma(\mathcal{O}x) = M(\mathcal{O})\,\Gamma(x)$. In this paper, we are interested in equivariance to the $SE(3)$ group denoting 3D position and orientation (pose) – in particular we focus on the rotation group $SO(3)$ since 3D translation equivariance is readily achieved through mean centering [26]. **Pose canonicalization** is closely related to equivariance – our goal is to find an *equivariant* canonical 3D rigid transformation that maps the neural field of an object in an arbitrary pose to a canonical pose. We call the outputs of our method **canonical fields** – neural fields that are consistently oriented across a category of shapes.

Rotation Equivariance in Vector Fields: Unlike 3D point clouds, rotation equivariance in fields is more involved. Consider a 2D vector field $\mathcal{F}$ defined $\forall \mathbf{x} \in \mathbb{R}^2$ (see Figure 2). Rotating this field requires two operations: (1) updating the positions of vectors in the field to new rotated positions, and (2) rotating the directions of the vectors. A function Γ operating on $\mathcal{F}$ is rotationally equivariant over the field if and only if $\Gamma[(\mathcal{R}\cdot\mathcal{F})(\mathbf{x})] = M(\mathcal{R})\,\Gamma[\mathcal{F}(\mathcal{R}^{-1}\mathbf{x})]$, where $\mathcal{R}$ is an $SO(2)$ rotation, and M is a rotation-dependent mapping function. In this paper, we operate on NeRF's density field (a scalar field) but also extract equivariant features that can be vector-valued fields.

Tensor Field Networks (TFNs): Many methods have been proposed for using equivariant features for rotation canonicalization of 3D data, for instance, spherical CNNs [7, 43], vector neurons [9], or capsule networks [47, 67]. However, these and other methods [38] are limited to 3D point cloud or voxel representations. In this paper, we extend Tensor Field Networks [34, 50] pose canonicalization directly on samples from NeRF's density field. TFNs operate by computing Type-ℓ real-valued spherical harmonic functions that are rotation equivariant: $D^\ell(R)Y^\ell(X) = Y^\ell(Rx)$. Where D^ℓ : $\mathrm{SO}(3) \to \mathrm{SO}(2\ell+1)$ is the Wigner matrix of type ℓ (please see [34, 50] for details). For a 3D point set $P \in \mathbb{R}^{N\times 3}$, a rotation-equivariant TFN convolution layer (EQConv) at point $p \in P$ is defined as: $\mathrm{EQConv}^J(p) = Q^{(n,\ell),J}\left(\sum_{y\in {}^{2R}\mathcal{N}} k_r^n(p-y)\otimes s^\ell[y]\right)$. Where ${}^{2R}\mathcal{N}$ is a set containing neighbors of point p at twice the resolution and $Q^{(n,\ell),J}(\cdot)$ is the Clebsh Gordon decomposition to combine type-n kernel k_r^n and type-ℓ equivariant signal s^ℓ. The $\mathrm{EQConv}(\cdot)$ layer aggregates type-ℓ feature s^ℓ at multiple receptive fields along with learnable weights to perform equivariant convolution. Our method uses the EQConv layer as the fundamental building block to canonicalize *continuous* fields.

4. Method

Our method assumes that we have pre-trained NeRF models of scenes containing single object instances at arbitrary poses from common shape categories. Our scenes contain primarily the object, but also some background – perfect foreground-background segmentation is not available. Our goal is to build a method that estimates a canonicalizing $SE(3)$ transformation that aligns these object NeRFs to a consistent category-level **canonical field**.

Rather than operate on 3D point clouds, meshes, or voxels, our method (see Figure 3) operates directly on samples from the continuous neural radiance field. During training, we learn **canonical fields** from a pre-trained NeRF dataset in a fully **self-supervised** manner. We use a Siamese network architecture that extracts equivariant field features for canonicalization (see Figure 3). During inference, given a previously unseen NeRF model of a new object instance in an arbitrary pose, our method estimates the transformation that maps it to the canonical field.

4.1. NeRF Sampling

For a pre-trained neural radiance field that maps 3D position (x, y, z) and direction θ, ϕ to color and density $(\mathbf{c}, \sigma)$, the input to our method consists of samples on the density field. We do not use color because it is direction-dependent in NeRF and also varies for different instances.

Preprocessing: During both training and inference, we uniformly sample the pre-trained NeRF density field to find the object center and its bounding box (see supplementary document for details). We then sample a uniform grid in the density field of the object's bounding box. We empirically choose uniform grid sampling over fully random sampling since we found no difference. The uniformly sampled density grid is then normalized to obtain density values $\boldsymbol{\sigma}_D$ in the continuous range $[0, 1]$, *i.e.*,

$$\boldsymbol{\sigma}_D := \{1 - exp(-d\cdot\Gamma_r^\sigma(x)) \mid x \in \boldsymbol{X}\}, \qquad (1)$$

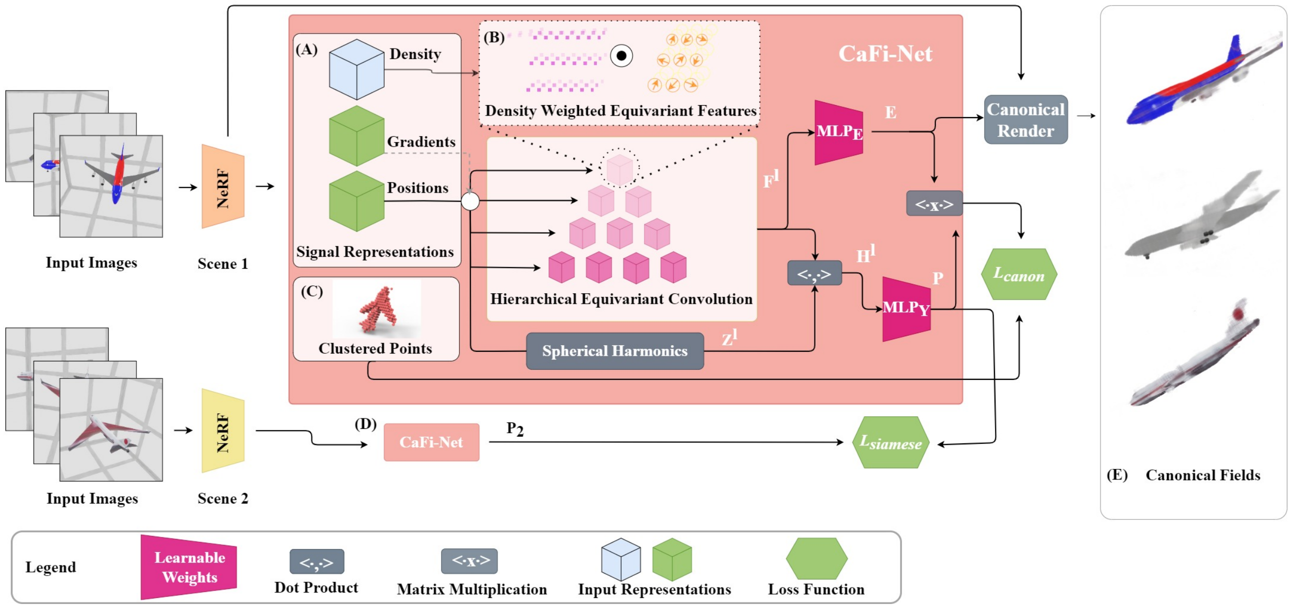

Figure 3. **CaFi-Net** samples a density field from NeRF and uses its **density**, **position** and **density gradients** as input signals (A) to canonicalize the field. We predict rotation equivariant features and weigh them by density (B) to guide our learning from the occupied regions of the scene. We then compute an invariant embedding by taking a dot product between equivariant features. This invariant embedding is used to canonicalize the field that enables rendering all the objects in the canonical frame (E). Our method also applies an inter-instance consistency loss (D) that aligns different instances of the same category in the canonical frame. **We do not assume pre-canonicalized fields, and canonicalize in a self-supervised manner**.

where $\boldsymbol{X}$ is the regularly sampled grid within the object bounding box and d is the depth step size used in the coarse NeRF sampling. **Translation** and **scale** canonicalization is simply achieved by moving the bounding box's center [26] and scaling it, so our next concern is on rotation canonicalization.

4.2. CaFi-Net

CaFi-Net (Canonical Field Network) is our method for 3D rotation canonicalization using NeRF's density field (red box in Figure 3). We now describe its components.

Signal Representation: At each of the sampled density field locations, we have several choices of signals to use as the input for CaFi-Net. For instance, we can (1) use the NeRF density value directly, (2) use the xyz location of the sampled field, or (3) use the gradient of the density field. Densities alone do not capture the position and directional components of the field, so we use a combination of densities and gradients of densities at query locations of the grid as our input signal for CaFi-Net. Gradients of density capture the object surface and help in canonicalization (see ablation study in Section 5).

Equivariant Convolution: To canonicalize for rotation, we leverage equivariance properties of spherical harmonic functions to obtain rotation-equivariant learnable features on the field [34, 50]. Type-ℓ spherical harmonic functions (Y^ℓ) are degree-ℓ polynomials of points on the sphere that are equivariant to $SO(2\ell+1)$. The densities sampled from NeRF reconstructions are often noisy and contain outliers. To promote our method to learn only from occupied regions, we weigh each equivariant feature type by its density. This weighing has no effect on the equivariance of features (see supplementary document for proof). We have the option to weigh equivariant features with the local average of the density to smoothen the signal or to weigh directly by the density. We found weighing by density worked better than taking the local average density (see Section 5). We learn equivariant features of type ℓ as:

$$ {}^{R}s^{\ell} := f_w({}^{R}x)\left(W^{\ell}(\mathrm{EQConv}^{\ell}(\nabla(\boldsymbol{\sigma}_D({}^{R}x)))) + \delta_{0\ell}b\right), \quad (2) $$

$$ f_w({}^{R}x) := \begin{cases} \mathrm{mean}_r(\boldsymbol{\sigma}_D({}^{R}x)), & or \\ \boldsymbol{\sigma}_D({}^{R}x), \end{cases} \quad (3) $$

where, W^ℓ and b are the weights and biases, ${}^{R}x$ is a 3D point queried at resolution R and $\mathrm{mean}_r(\boldsymbol{\sigma}_D({}^{R}x))$ is the average density at location ${}^{R}x$ from points sampled at radius r. We hierarchically aggregate features at resolutions $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$ to obtain global equivariant features. The EQConv^{ℓ} convolution layer is the same as defined in Section 3. We employ non-linearities from [34] that preserve equivariance and capture better equivariant features. We also compute the max-pool of point-wise features of the last layer to obtain global type-ℓ equivariant features F^ℓ.

Canonicalization: After obtaining the global type-ℓ

equivariant feature F^ℓ, we compute its dot product with point-wise spherical harmonics Z^ℓ scaled by their norms $Z^\ell(\boldsymbol{X}) = ||\boldsymbol{X}||_2 Y^\ell(\boldsymbol{X}/||\boldsymbol{X}||_2)$ to obtain a pose-invariant embedding H^ℓ [37]. As the two vectors F^ℓ and Z^ℓ are both equivariant to input rotation, their dot product is invariant. We use a linear layer on top of the pose-invariant embedding to estimate grid coordinates in the canonical frame ($P \in \mathbb{R}^{H\times W\times D\times 3}$) for each point in $\boldsymbol{X}$ and M equivariant transformations $E \in \mathbb{R}^{M\times 3\times 3}$ [37] that orient the canonical coordinates to the input coordinates. We then choose the best canonicalizing transform E_b and penalize it for canonicalization. The invariant embedding H, grid coordinates P and equivariant rotation E are given as,

$$H^\ell(\boldsymbol{\sigma}_D, \boldsymbol{X}) := \langle F^\ell(\boldsymbol{\sigma}_D), Z^\ell(\boldsymbol{X})\rangle, \tag{4}$$

$$P := \text{MLP}_Y(H(\boldsymbol{\sigma}_D, \boldsymbol{X})), \tag{5}$$

$$E := \text{MLP}_E(F(\boldsymbol{\sigma}_D, \boldsymbol{X})). \tag{6}$$

Here, we have dropped the type-ℓ notation for H and F as we concatenate all equivariant feature types.

Siamese Network Architecture: Experimentally, canonicalizing noisy density fields is difficult without guidance on shape similarity within a category. We therefore use a Siamese training strategy [26] where two different object instances are forced to be consistently canonicalized (see the lower branch in Figure 3).

Clustering: CaFi-Net predicts a canonical coordinate for every grid point, but unlike point clouds, fields have many unoccupied regions that do not provide any information. We penalize the network on foreground regions only by performing K-means clustering on the densities within the object bounding box with $K = 2$ and choose the cluster $\boldsymbol{C}_f$ with a higher mean. Note that we still operate on continuous fields, but only cluster samples to guide our training.

4.3. Training

CaFi-Net is trained on a large dataset of pre-trained NeRF models over 13 object categories (see Section 5). We use the following loss functions to train our model.

Canonicalization Loss: To self-supervise our learning, we transform the predicted canonical grid coordinates to the input grid using the canonical rotation E_b and compute a point-wise L2 distance between them. This loss forces the predicted invariant grid coordinates P to reconstruct the shape and forces the predicted canonicalizing transform to be equivariant to the input transformation. The final loss is computed over the canonicalizing transform that minimizes this loss:

$$E_b := \min_j \left(\text{mean}_{i\in\boldsymbol{C}_f} ||\boldsymbol{X}_i - E_j P_i||_2^2\right), \tag{7}$$

$$\mathcal{L}_{canon} := \text{mean}_{i\in\boldsymbol{C}_f} ||\boldsymbol{X}_i - E_b P_i||_2^2, \tag{8}$$

where C_f is a set of points belonging to the foreground.

Orthonormality Loss: The predicted equivariant transformations E should have orthonormal vectors and we use the following to force orthonormality using:

$$U_j, D_j, V_j := \text{SVD}(E_j), \tag{9}$$

$$\mathcal{L}_{ortho} := \frac{1}{M}\sum_j ||E_j - U_j V_j^T||_2, \tag{10}$$

where $\text{SVD}(E_j)$ computes the singular value decomposition of the j^{th} equivariant transformation.

Siamese Shape Loss: We use a Siamese loss that penalizes for consistency between different shapes belonging to the same category. Given two un-canonicalized fields in different frames of reference $\boldsymbol{\sigma}_D^1$ and $\boldsymbol{\sigma}_D^2$, we canonicalize both the fields to the canonical frame and compute chamfer distance between their predicted canonical coordinates for points in the foreground cluster C_f, *i.e.*,

$$\mathcal{L}_{siamese} := \text{CD}(P(\boldsymbol{\sigma}_D^1), P(\boldsymbol{\sigma}_D^2)). \tag{11}$$

This loss regularizes the training by ensuring that instances of the same category should be aligned in the canonical frame.

Architecture and Hyper-parameters: CaFi-Net predicts a 128 dimensional invariant embedding for each point in space and performs convolution at resolutions $1/2$, $1/4$, and $1/8$. Each equivariant convolution layer aggregates features from 512 neighboring points at twice the resolution. We then use equivariant non-linearities introduced in [34] by applying Batch Normalization [14] and ReLU activation [2] in the inverse spherical harmonic transform domain. We use a three-layer MLP with intermediate Batch Normalization and ReLU layers followed by a final linear layer to predict canonical coordinates. To train our network, we weigh the canonicalization loss ($\mathcal{L}_{canon}$) the highest with weight 2.0 and 1.0 for all the other loss functions. All our models are written in PyTorch [31] and are trained for 300 epochs with a batch size of 2 on an Nvidia 1080-Ti GPU. We use the Adam Optimizer [16] with an initial learning rate of $6e-4$ and L2 weight regularization of $1e-5$ for all our experiments. Please see the supplementary document for more details.

5. Experiments

We now provide details on our extensive experiments to evaluate the quality of our pose canonicalization. To our knowledge, we are the first method for pose canonicalization of neural radiance fields which makes direct comparisons with other work challenging. Nonetheless, we have designed experiments to compare our method with three 3D point cloud-based canonicalization methods. Overall, we present three classes of experiments: (1) evaluation of the performance of our method on 13 different categories,

Table 1. This table compares the canonicalization performance of our method (F - operating on fields) with other 3D point cloud-based methods (P) on three standard metrics (IC, CC and GEC) on our dataset of 13 categories. We compare with PCA, Canonical Capsules (CaCa) [47] and ConDor [37]. All metrics are multiplied by 100 for ease of reading. The top two performing methods are highlighted in **magenta** (best) and **blue** (second best). We are better than SOTA [37] on the Ground Truth Equivariance Consistency (GEC) and Category-Level Consistency (CC) with lower mean and median canonicalization error. However, we perform on par with ConDor [37] on the Instance-Level Consistency.

	bench	cabinet	car	cellph.	chair	couch	firearm	lamp	monitor	plane	speaker	table	water.	avg.	med.
Instance-Level Consistency (IC) ↓															
PCA (P)	14.65	5.94	6.13	**0.67**	6.13	7.63	15.07	12.84	6.78	7.90	5.40	10.31	10.72	8.47	7.63
CaCa (P) [47]	2.89	2.09	2.01	2.81	1.08	1.90	**0.23**	**2.97**	2.35	1.51	2.21	2.42	**2.60**	2.08	2.21
ConDor (P) [37]	**0.56**	**1.27**	**0.36**	**0.53**	**0.96**	**0.32**	**0.60**	**3.78**	**0.44**	**0.64**	**1.97**	**1.32**	**2.15**	**1.15**	**0.64**
Ours (F)	**2.62**	**1.15**	**0.39**	1.28	**0.81**	**1.43**	0.74	3.87	**0.41**	**0.36**	**0.60**	**1.09**	2.74	**1.34**	**1.09**
Category-Level Consistency (CC) ↓															
PCA (P)	14.55	9.68	7.48	4.26	9.90	13.45	15.13	11.87	13.09	8.15	5.51	10.43	11.07	10.36	10.43
CaCa (P) [47]	2.79	**0.38**	1.05	**0.83**	1.82	**1.59**	1.31	**4.34**	**0.39**	1.83	**0.65**	**1.91**	**2.06**	**1.61**	1.59
ConDor (P) [37]	**1.98**	1.68	**0.47**	**1.04**	**1.28**	**0.88**	**0.96**	4.44	1.58	**1.26**	2.10	2.10	2.56	1.72	**1.58**
Ours (F)	**2.77**	**1.36**	**0.48**	1.26	**0.75**	**1.59**	**0.92**	**3.92**	**0.58**	**0.63**	**0.77**	**1.48**	**2.35**	**1.45**	**1.26**
Ground Truth Equivariance Consistency (GEC)↓															
PCA (P)	14.40	5.86	6.32	1.64	6.32	8.12	15.01	12.49	6.82	8.16	5.32	10.17	11.27	8.61	8.12
CaCa (P) [47]	3.65	2.28	2.40	2.74	1.90	2.22	1.28	4.69	2.48	2.09	**2.22**	3.01	**2.79**	2.59	2.40
ConDor (P) [37]	**2.12**	**1.79**	**0.50**	**1.17**	**1.34**	**0.93**	**1.04**	**4.58**	**1.59**	**1.31**	**2.22**	**2.24**	**2.67**	**1.81**	**1.59**
Ours (F)	**3.26**	**1.51**	**0.54**	**1.43**	**0.94**	**1.90**	**1.01**	**4.37**	**0.63**	**0.66**	**0.81**	**1.63**	3.02	**1.67**	**1.43**

(2) comparison with other work, and (3) ablations to justify method design choices.

Dataset: All our experiments use a new large synthetic NeRF dataset of shapes from 13 categories (that overlap with [12, 37, 47]) from the ShapeNet [5] dataset. For each category, we pick 100 instances from ShapeNet, rotate the shape randomly, and render 54 omnidirectional views sampled in a cube around the shape using Blender. Different from other datasets, we also simulate cluttered backgrounds. These posed views are used to train 100 NeRF models per category using the public PyTorch implementation of NeRF [62] for a total of **1300 NeRF models**. To our knowledge, this is one of the largest (compared to [11, 53]) 360° synthetic NeRF datasets – we will publicly release the simulator, raw posed images, and trained NeRF weights. Each NeRF model is trained for 400 epochs with 1024 randomly selected rays per iteration, a coarse sampling resolution of 64 points, and a fine sampling resolution of 128 points along each ray. Rather than train NeRF to maximize PSNR on a novel view, we fix the number of epochs for all models to enable a fair comparison between instances. As a consequence, some NeRF models may have a lower PSNR, but **our goal is not to increase NeRF quality but rather improve canonicalization quality** – hence we fix the number of epochs. For each category, we set aside 20% of the models for testing. **Metrics**: Since there are no known metrics for evaluating canonical fields, we resort to 3 metrics used for 3D point cloud canonicalization [37]: (1) **Instance-Level Consistency (IC)**, a metric that measures how consistently we can canonicalize the same instance in different poses, (2) **Category-Level Consistency (CC)** to measure how consistently we canonicalize across different instances in the same category, and (3) **Ground Truth Equivariance Consistency (GEC)**, a variant of the Ground Truth Consistency (GC) metric proposed in [37] to measure canonicalization performance compared to manual labels. We use the GEC because we observed that the original GC metric is prone to degeneracy when the canonicalizing transform is identity. We fix this issue by taking three point clouds P_i, P_j, and P_k from a canonicalized dataset $\mathcal{P}$ and rotating P_i by R_1, P_j by R_2 where both R_1 and R_2 are random rotations. Let $\mathcal{C}(P)$ predict a canonicalizing rotation for a point cloud P, we then compute the Ground Truth Equivariance consistency as:

$$\text{GEC} := \frac{1}{|\mathcal{P}|^3} \sum_{P_i, P_j, P_k \in \mathcal{P}} \text{CD}(\mathcal{C}(R_1 P_i) R_1 P_k, \mathcal{C}(R_2 P_j) R_2 P_k), \quad (12)$$

where CD refers to the Chamfer distance. Here, we apply canonicalizing transforms of shape P_i and P_j to the shape P_k. This modified metric will not be degenerate for identity canonicalizing transforms and is more rigorous in evaluating ground truth consistency. To compute our metrics, we used point clouds sampled from the Shapenet meshes [5] using poisson disk sampling.

5.1. Evaluation & Comparisons

We first analyze the performance of our method on the three metrics across 13 different categories. In Table 1, we highlight our method that operates on a field as **Ours (F)**. Our method performs well across all categories with low variance between different categories. We observe similar trends across metrics for a set of categories (*e.g.*, firearms,lamps) indicating that category shape plays a

Category	xyz vs. Gradient Signals	
	xyz	Gradient signal
Ground Truth Equivariance Consistency (GEC)↓		
bench	**2.17**	3.26
cellphone	1.84	**1.43**
chair	1.72	**0.94**
plane	2.07	**0.66**
Average	1.95	**1.57**
Instance-Level Consistency (IC)↓		
bench	**2.04**	2.62
cellphone	1.6	**1.28**
chair	1.4	**0.81**
plane	1.88	**0.36**
Average	1.73	**1.26**
Category-Level Consistency (CC)↓		
bench	**1.99**	2.77
cellphone	1.49	**1.26**
chair	1.52	**0.75**
plane	1.9	**0.63**
Average	1.72	**1.35**

(a) **Choice of Signal Representation** - Canonicalization metrics for using Gradients vs. xyz locations as input signal. Gradients capture the object surface that help in canonicalization.

Category	Local Average Density	
	w/o	w
Ground Truth Equivariance Consistency (GEC)↓		
bench	**3.26**	3.38
cellphone	**1.43**	1.63
chair	**0.94**	1.3
plane	0.66	**0.64**
Average	**1.57**	1.73
Instance-Level Consistency (IC)↓		
bench	**2.62**	2.72
cellphone	**1.28**	1.63
chair	**0.81**	2.64
plane	**0.36**	0.39
Average	**1.26**	1.85
Category-Level Consistency (CC)↓		
bench	**2.77**	2.82
cellphone	**1.26**	1.54
chair	**0.75**	1.07
plane	0.63	**0.51**
Average	**1.35**	1.49

(b) **Weighing Equivariant Signals by Local Average Density** deteriorates the performance by smoothing out important details of the shape. We show canonicalization with (**w**) and without (**w/o**) weighing by the local averaged density.

Category	Siamese Training	
	w/o siamese	with siamese
Ground Truth Equivariance Consistency (GEC)↓		
bench	3.52	**3.26**
cellphone	1.61	**1.43**
chair	1.02	**0.94**
plane	1.31	**0.66**
Average	1.86	**1.57**
Instance-Level Consistency (IC)↓		
bench	2.73	**2.62**
cellphone	1.3	**1.28**
chair	0.82	**0.81**
plane	1.1	**0.36**
Average	1.48	**1.26**
Category-Level Consistency (CC)↓		
bench	2.95	**2.77**
cellphone	1.44	**1.26**
chair	0.83	**0.75**
plane	1.26	**0.63**
Average	1.62	**1.35**

(c) **Siamese Training** improves performance on all canonicalization metrics on average. We show canonicalization performance *with siamese* and without **(w/o) siamese** training. The average of Ground Truth Equivariance Consistency *GEC* metric reduces to 1.57 from 1.86

Table 2. Ablation studies to justify three key design choices: (a) signal representation, (b) density weighting, and (c) Siamese architecture.

role in canonicalization performance.

Next, we compare our method against three benchmarks: Principal Component Analysis (PCA) [32], Canonical Capsules (CaCa) [47], and ConDor [37]. All prior methods canonicalize clean point clouds with all points on the surface. To have a fair comparison, we train the above methods on point clouds (P) sampled from NeRF scenes for points on and within the object. Most scenes have points in the range of 120–1200 when sampled uniformly on the 32^3 grid. We resample the scene when points are less than 1024 and train both Canonical Capsules and ConDor on the resulting point clouds. CaFi-Net estimates a canonicalizing transform directly from the density field that we use to transform its corresponding ShapeNet point cloud for comparison against point canonicalizers. We follow the same protocol as [37] to measure and benchmark canonicalization performance for all the methods.

Analysis: Table 1 compares the metrics for PCA (P), CaCa (P), ConDor (P), and CaFi-Net [Ours (F)]. We list point cloud canonicalizers with (P) and field canonicalizer as (F) in the table. On average, our method performs better than point cloud-based methods in the CC and GEC metrics and is on par with the state-of-the-art method (ConDor) in the IC metric. This is despite the fact that our method directly operates on the field. We observe that PCA underperforms in all categories suggesting that the noisy point clouds make it unreliable. Similarly we observe that CaCa fails to canonicalize in categories where high noise is observed. Given that ConDor also uses TFNs in a different architecture, we conclude that the choice of TFNs makes canonicalization more robust to noise. Our method has the additional advantage of operating directly on the field. Please see Figure 4 for a qualitative comparison of the different methods.

5.2. Ablations

We conduct additional experiments to justify key design choices in our method. For all the ablation studies we use a smaller subset of our data consisting of 4 categories: bench, cellphone, chair, and airplane.

Choice of Signal Representation: To select the appropriate signal representation to use (see Section 4), we conducted an ablation study using just the density field xyz locations or its gradients. As seen in Table 2 (a), the average canonicalization error over 4 categories is lower when using gradient as compared to using density xyz locations on all three metrics. We hypothesize that gradient of density weighs equivariant features closer to the surface of the object higher. This improves the canonicalization performance.

Weighing Features by Local Average Density: Next, we justify why weighting the equivariant features using the density is helpful. Table 2 (b) shows results for weighting features by local average density compared to weighting features by density. Weighing features by local average smoothens out details that are necessary to canonicalize the volume. Our average ground truth canonicalization error

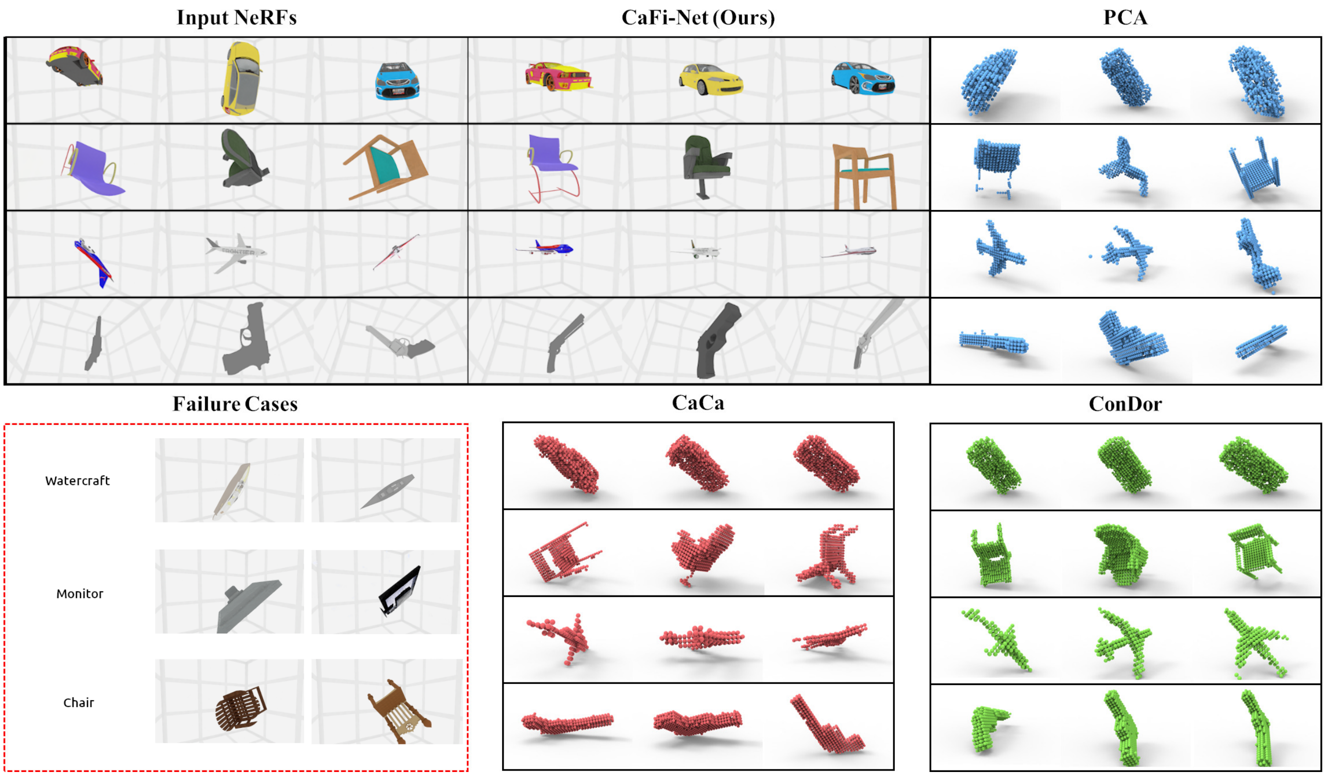

Figure 4. This figure visually compares the canonicalization performance of our method with 3D point cloud-based methods. **Input NeRFs** shows one of the RGB views used to learn our NeRF input (notice how orientations of instances are random). **CaFi-Net (Ours)** shows the results of our canonical field rendered from a novel view unseen during NeRF training. **PCA** (blue), **CaCa** [47] (red), and **ConDor** [37]'s (green) results are also presented for the same instances. Our method matches or exceeds ConDor. We also show some failure cases for several instances (red box) which occur either due to thin structures or symmetry.

increases from 1.57 to 1.73 if we weigh features by local averaged densities. In future works, we will delve into better methods for reducing input signal noise that does not smoothen out the geometry details of objects to further improve our canonicalization.

Siamese Architecture: Finally, we justify the need for a two-branch Siamese network architecture. Although our method can canonicalize even without this architecture, it makes it easier to learn over instances in a category as observed in [26]. Table 2 (c) compares our method on using a single branch architecture and a Siamese architecture. Clearly, the Siamese architecture helps improve results confirming our hypothesis. Enforcing different instances of the same category to align with each other reduces the Ground Truth Canonicalization error from 1.86 to 1.57. We also see a decrease in all the other metrics including Instance-Level Consistency.

6. Conclusion

We presented Canonical Field Network (CaFi-Net), a method for self-supervised category-level canonicalization of the 3D pose of objects represented as neural radiance fields. CaFi-Net consists of a Siamese network architecture with rotation-equivariant convolution layers to extract features for pose canonicalization. Our approach operates directly on NeRF's noisy but continuous density field. We train our method on a large dataset of 1300 NeRF models obtained for 13 common ShapeNet categories. During inference, our method is able to canonicalize arbitrarily oriented NeRFs. Experiments show that our method matches or outperforms 3D point cloud-based methods.

Limitations & Future Work: Our approach has several limitations that future work should investigate. First, we require a 360° NeRF model of objects and cannot handle partial views from front-facing NeRFs. Second, canonicalizing densities with uncertainty is more difficult as the uncertainty increases for very thin/small structures that can be missed (see chair in Figure 4). Furthermore, we also inherit the issues that TFNs have with symmetry (see monitor case in Figure 4). In future work, we plan to extend our approach to create a large real dataset of common object categories without requiring manual pose canonicalization and combat the symmetry issue in CaFi-Net as in [39]

Acknowledgments: This work was supported by NSF grant CNS-2038897, an AWS Cloud Credits award, NSF CloudBank, and a gift from Meta Reality Labs. We thank Chandradeep Pokhariya and Ishaan Shah.

References

[1] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022. 1

[2] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 5

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1, 2

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 2

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1, 2, 6

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[7] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. 3

[8] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. 2

[9] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12200–12209, October 2021. 2, 3

[10] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020. 2

[11] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2021. 2, 6

[12] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 2, 6

[13] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-nerf: Using a neural radiance field to grasp transparent objects. Oct 2021. 1

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 5

[15] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5

[17] Jiahui Lei, Srinath Sridhar, Paul Guerrero, Minhyuk Sung, Niloy Mitra, and Leonidas J Guibas. Pix2surf: Learning parametric 3d surface models of objects from images. In *European Conference on Computer Vision*, pages 121–138. Springer, 2020. 2

[18] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 2

[19] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 1, 2

[20] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3706–3715, 2020. 2

[21] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. In *Proceedings of Robotics: Science and Systems*, 2021. 1

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[23] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 2

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2

[25] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 1

[26] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7688–7697, 2019. 2, 3, 4, 5, 8

[27] David Novotny, Roman Shapovalov, and Andrea Vedaldi. Canonical 3D Deformer Maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction. In *NeurIPS*, 2020. 2

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 1, 2

[29] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[30] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 1, 2

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[32] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 7

[33] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1

[34] Adrien Poulenard and Leonidas J Guibas. A functional approach to rotation equivariant non-linearities for tensor field networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13174–13183, 2021. 3, 4, 5

[35] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. 2

[36] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[37] Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J Guibas, and Srinath Sridhar. Condor: Self-supervised canonicalization of 3d pose for partial shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16969–16979, 2022. 2, 5, 6, 7, 8

[38] Rahul Sajnani, AadilMehdi Sanchawala, Krishna Murthy Jatavallabhula, Srinath Sridhar, and K Madhava Krishna. Draco: Weakly supervised dense reconstruction and canonicalization of objects. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10302–10309. IEEE, 2021. 2, 3

[39] Ahyun Seo, Byungjin Kim, Suha Kwak, and Minsu Cho. Reflection and rotation symmetry detection via equivariant learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[40] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022. 1

[41] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. 2022. 2

[42] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[43] Riccardo Spezialetti, Federico Stella, Marlon Marcon, Luciano Silva, Samuele Salti, and Luigi Di Stefano. Learning to orient surfaces by self-supervised spherical cnns. *arXiv preprint arXiv:2011.03298*, 2020. 1, 2, 3

[44] Srinath Sridhar, Davis Rempe, Julien Valentin, Sofien Bouaziz, and Leonidas J Guibas. Multiview aggregation for learning category-specific shape reconstruction. *arXiv preprint arXiv:1907.01085*, 2019. 2

[45] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[46] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2

[47] Weiwei Sun, Andrea Tagliasacchi, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geoffrey Hinton, and Kwang Moo Yi. Canonical capsules: Unsupervised capsules in canonical pose. *arXiv preprint arXiv:2012.04718*, 2020. 1, 2, 3, 6, 7, 8

[48] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 2

[49] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering. Nov 2021. 2

[50] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 2, 3, 4

[51] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, October 2022. 2

[52] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Nonrigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 1

[53] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes, 2021. 1, 6

[54] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[55] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 2021. 1

[56] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *NIPS*, pages 10381–10392, 2018. 2

[57] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proc. CVPR*, pages 849–858, 2018. 2

[58] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proc. CVPR*, pages 1912–1920, 2015. 2

[59] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 1, 2

[60] Guandao Yang, Serge Belongie, Bharath Hariharan, and Vladlen Koltun. Geometry processing with neural fields. *Advances in Neural Information Processing Systems*, 34:22483–22497, 2021. 1

[61] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2021. 1

[62] Lin Yen-Chen. Nerf-pytorch. `https://github.com/yenchenlin/nerf-pytorch/`, 2020. 6

[63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 2

[64] Kevin Zakka, Andy Zeng, Johnny Lee, and Shuran Song. Form2fit: Learning shape priors for generalizable assembly from disassembly. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9404–9410. IEEE, 2020. 2

[65] Ge Zhang, Or Litany, Srinath Sridhar, and Leonidas Guibas. Strobenet: Category-level multiview reconstruction of articulated objects. *arXiv preprint arXiv:2105.08016*, 2021. 2

[66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 2

[67] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1018, 2019. 3

[68] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[69] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021. 2