

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

Rohith Agaram<sup>1</sup>      Shaurya Dewan<sup>1</sup>      Rahul Sajnani<sup>2</sup>      Adrien Poulenard<sup>3</sup>

Madhava Krishna<sup>1</sup>      Srinath Sridhar<sup>2</sup>

<sup>1</sup>RRC, IIIT-Hyderabad      <sup>2</sup>Brown University      <sup>3</sup>Stanford University

[ivl.cs.brown.edu/projects/canonicalfields](http://ivl.cs.brown.edu/projects/canonicalfields)

## A. Preprocessing Density Field

Before querying NeRF for occupancy, we do not have the extremities of the object. We query a uniform grid within the unit cube to obtain a  $32^3$  grid density field to find the object bounds and the center of the object. The density values are normalized to range  $[0, 1]$  (see page 3 in the main paper). We then perform an initial K-means clustering with  $K = 2$  to separate the foreground and background clusters. Note that this clustering is on the entire scene and is different from the clustering we perform while computing the losses. The clustering that we perform while computing the loss is within the object bounding box and not for the entire scene. The mean of query coordinate locations in the foreground cluster gives us the center of the object  $\mathbf{c}$ . We then obtain the extremities from the foreground cluster and find its maximal diagonal length ( $l$ ). We re-sample the region within the bounding cube (at center  $\mathbf{c}$  side length  $l$ ) and use it as input to CaFi-Net.

## B. Equivariance Properties of CaFi-Net

### B.1. Averaging Equivariant Signals is Equivariant

A tensor field of type  $\ell$  ( $\ell$ -field) is a map  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^{2\ell+1}$ . We have an action of  $\text{SO}(3)$  on any  $\ell$ -field  $f$  given by  $(R.f)(x) := D^\ell(R)f(R^{-1}x)$  for any rotation  $R \in \text{SO}(3)$  and  $x \in \mathbb{R}^3$  where  $D^\ell(R) \in \text{SO}(2\ell+1)$  is the Wigner matrix of type  $\ell$ . We say that transformation  $F$  transforming tensor fields of type  $p$  into tensor fields of type  $q$  if it commutes with the action of  $\text{SO}(3)$ , i.e. for all type  $p$ -field  $f$  and rotation  $R \in \text{SO}(3)$  we have  $F(R.f) = R.F(f)$ . We will call such transformations  $(p, q)$ -equivariant transformations.

**Lemma 1** For any  $r > 0$  the local average operator  $\text{mean}_r$  defined over  $p$ -fields by:

$$\text{mean}_r(f)(x) := \int_{B(x,r)} f(y)dy$$

where  $B(x, r) \subset \mathbb{R}^3$  is the open ball of radius  $r$  centered at  $x \in \mathbb{R}^3$  is a  $(p, p)$ -equivariant transform of fields.

**Proof:** Let  $x \in \mathbb{R}^3$  and  $R \in \text{SO}(3)$  we have:

$$\begin{aligned} \text{mean}_r(R.f)(x) &= \int_{B(x,r)} (R.f)(y)dy \\ &= \int_{B(x,r)} D^p(R)f(R^{-1}y)dy \\ &=_{u=R^{-1}y} D^p(R) \int_{R^{-1}B(x,r)} f(u)du \\ &= D^p(R) \int_{B(R^{-1}x,r)} f(u)du \\ &= (R.\text{mean}_r(f))(x) \end{aligned}$$

### B.2. Locally Averaged Density Weighted Equivariant Vectors are Equivariant

**Lemma 2** Scaling an equivariant field with density is a  $(1, 1)$ -equivariant transformation.

**Proof:** Let  $\sigma$  be a type-0 density field and  $f$  be a type-1 field at the corresponding location, weighing  $f$  by the average of  $\sigma$  is:

$$\begin{aligned} \text{mean}_r(\sigma)(x) \cdot f(x) &:= \left( \int_{B(x,r)} \sigma(y)dy \right) f(x) \\ \text{mean}_r(R \cdot \sigma)(x) \cdot (R \cdot f)(x) &= \left( \int_{B(x,r)} (R \cdot \sigma)(y)dy \right) D^1(R)(f)(R^{-1}x) \\ &= \left( \int_{B(x,r)} (\sigma)(R^{-1}y)dy \right) D^1(R)(f)(R^{-1}x) \\ &= \left( I \cdot \int_{R^{-1}B(x,r)} (\sigma)(u)du \right) D^1(R)(f)(R^{-1}x) \\ &= D^1(R) \cdot \text{mean}_r(\sigma)(x) \cdot (f)(R^{-1}x) \end{aligned}$$

Thus,  $\text{mean}_r(R \cdot \sigma)(x) \cdot f(x) = D^1(R) \cdot \text{mean}_r(\sigma)(x) \cdot f(R^{-1}x)$  proving the result that scaling equivariant features by average density do not break the equivariance property.

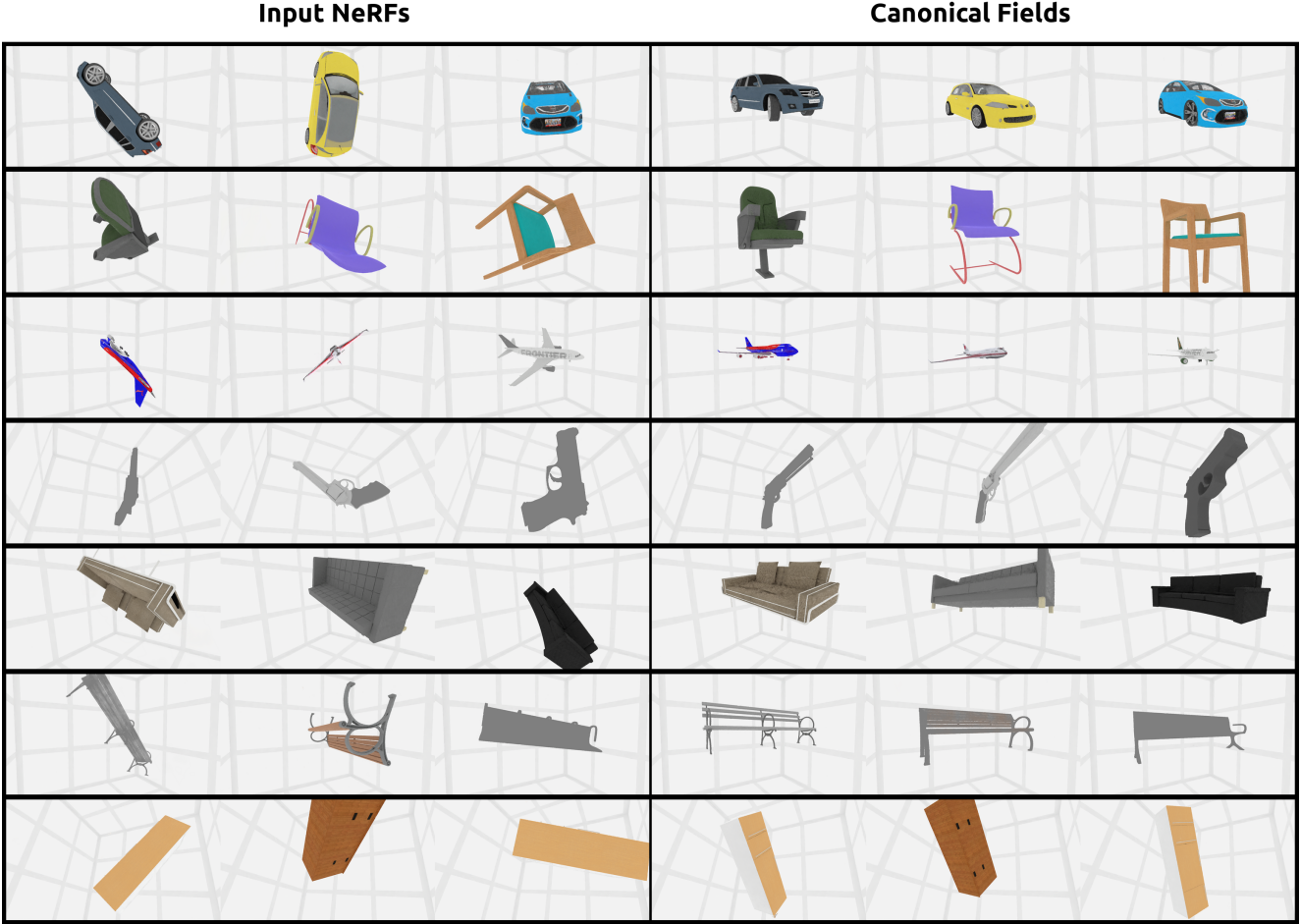


Figure 1. CaFi-Net qualitative canonicalization results for 7 categories (see following pages for more results).

### B.3. Gradient is Type-1 Equivariant

**Lemma 3** *The gradient operator is a  $(0,1)$ -equivariant transformation.*

**Proof:** Let  $f$  be a 0-type field, by the chain rule of differentiation, for any  $x, h \in \mathbb{R}^3$  and  $R \in \text{SO}(3)$  we have

$$\begin{aligned}
 \langle \nabla(R.f), h \rangle &= D_x(R.f)(h) \\
 &= D_x f \circ R^{-1}(h) \\
 &= D_{R^{-1}x} f \circ D_x R^{-1}(h) \\
 &= \langle \nabla_{R^{-1}x} f, R^{-1}h \rangle \\
 &= \langle R \nabla_{R^{-1}x} f, h \rangle \\
 &= \langle R \nabla_{R^{-1}x} f, h \rangle
 \end{aligned}$$

thus  $\nabla(R.f) = R \cdot \nabla f$  which concludes the proof.

### C. Additional Training Details

To train CaFi-Net, we augment NeRF density fields with random rotations  $R_{rand}$  and canonicalize them at each

Equivariant Convolution Non-Linearities
InverseSphericalHarmonicTransform(sphere_samples=64)
BatchNorm(momentum=0.75)
ReLU
MLP( $F_{in}, F_{out}$ )
ForwardSphericalHarmonicTransform(sphere_samples=64)

Table 1. **Equivariant Convolution Non-Linearities** - We use the result in [2] and apply non-linearities after performing an inverse spherical harmonic transform to avoid breaking the equivariance of each layer. The  $sphere\_samples$  is the sphere sampling resolution to perform the Spherical Harmonic Transform and  $F_{in}, F_{out}$  are the input and output feature dimensions respectively.

training iteration. We can easily do this by sampling the NeRF model at location  $R_{rand}^{-1}x$  instead of  $x$  in a differentiable manner using [1]. Our method is built over equivariant layers and non-linearities from [2]. Table 1 shows the non-linearities and learning layers that are used after each equivariant convolution described in the main manuscript.

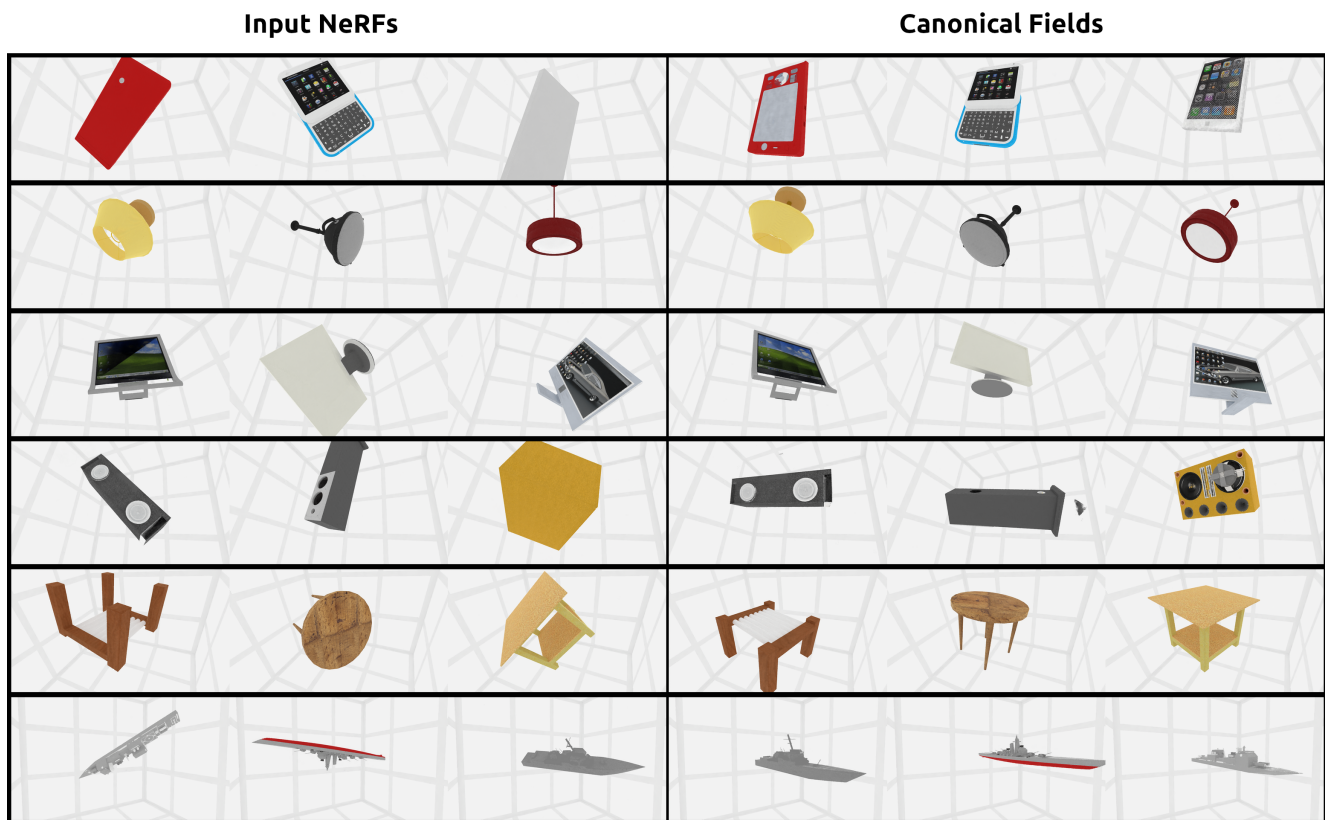


Figure 2. **CaFi-Net** qualitative canonicalization results for the remaining 6 categories.

## D. Qualitative Results

We illustrate qualitative results by rendering objects in the canonical frame for all the 13 categories in Figure 1 and Figure 2. Here, we fix a camera position and viewing direction in the canonical frame and render all objects from the same camera.

## References

- [1] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2
- [2] 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. 2