# ELLIPSDF: Joint Object Pose and Shape Optimization with a Bi-level Ellipsoid and Signed Distance Function Description

Mo Shan, Qiaojun Feng, You-Yi Jau, Nikolay Atanasov
University of California San Diego

## Motivations & Contributions

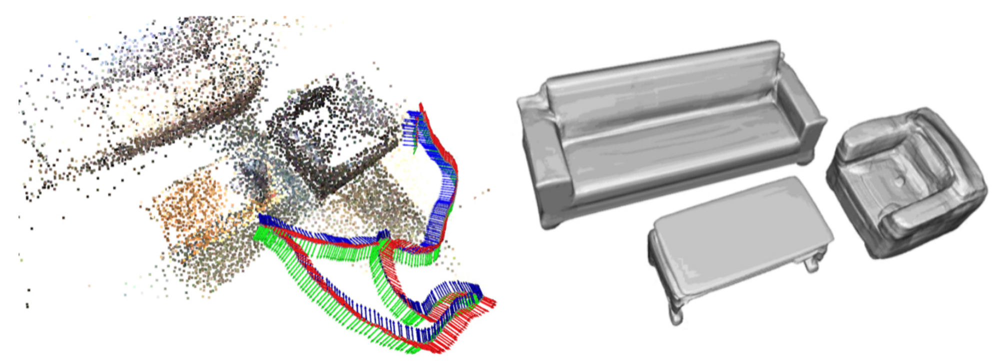
**Motivations:**

- Build maps that offer geometric and semantic information useful and understandable for humans, allowing specification of tasks in terms of object entities.
- Strike the right balance between a faithful object reconstruction and a compact object representation.

**Contributions:**

- A **bi-level object model** with coarse and fine levels, to enable joint optimization of object pose and shape. The two levels are coupled via a shared latent space.
  - **Coarse-level** uses a primitive shape for robust pose and scale initialization.
  - **Fine-level** uses SDF residual directly to allow accurate shape modeling.
- A cost function to measure the mismatch between the bi-level object model and the segmented RGB-D observations in the world frame.

**Overview:** We propose ELLIPSDF, an expressive yet compact model of object pose and shape, and an associated optimization algorithm to infer an object-level map from multi-view RGB-D camera observations.



## Problem Formulation

**Definitions:**

- An *object class* is a tuple $\boldsymbol{c} \triangleq (\nu, \boldsymbol{z}, f_{\boldsymbol{\theta}}, g_{\boldsymbol{\phi}})$
  - $\nu \in \mathbb{N}$ is the class identity, e.g., chair, table, sofa.
  - $\boldsymbol{z} \in \mathbb{R}^d$ is latent code encoding average shape.
- Shape is represented in a canonical coordinate frame at two levels of granularity: coarse and fine.
  - Coarse shape is specified by an **ellipsoid** $\mathcal{E}_{\boldsymbol{u}}$ with semi-axis lengths $\boldsymbol{u} = g_{\boldsymbol{\phi}}(\boldsymbol{z})$ decoded from the latent code $\boldsymbol{z}$ via a function $g_{\boldsymbol{\phi}}$ with parameters $\boldsymbol{\phi}$.
  - Fine shape is specified by the **signed distance** $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ from any $\boldsymbol{x} \in \mathbb{R}^3$ to the average shape surface, decoded from the latent code $\boldsymbol{z}$ via a function $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$.
- An *object instance* of class $\boldsymbol{c}$ is a tuple $\boldsymbol{i} \triangleq (\boldsymbol{T}, \delta\boldsymbol{z})$.
  - $\boldsymbol{T} \in SIM(3)$ specifies the transformation from the global frame to the object instance frame.
  - $\delta\boldsymbol{z} \in \mathbb{R}^d$ is a deformation of the latent code $\boldsymbol{z}$, specifying the average shape of class $\boldsymbol{c}$.

**Error Functions:**

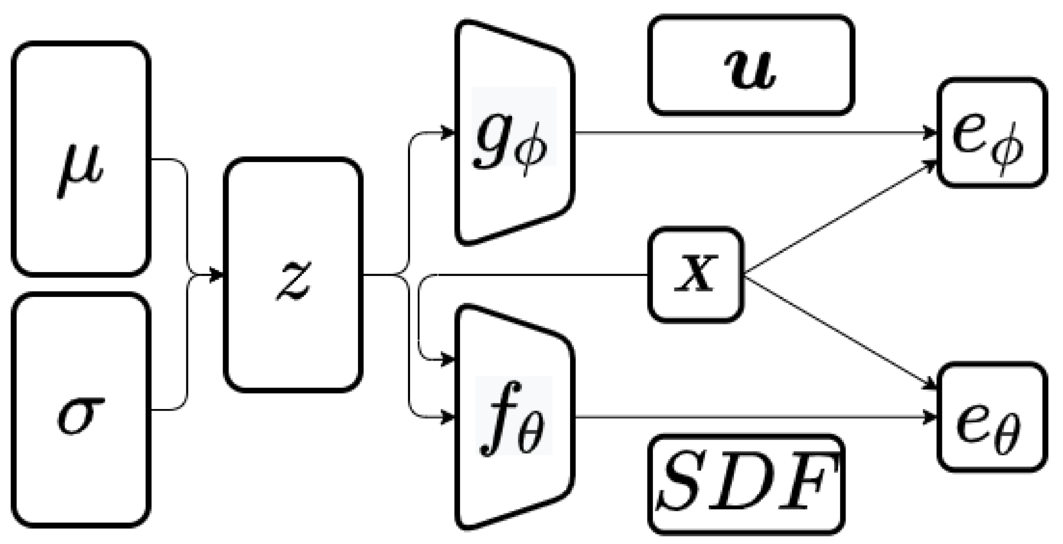
- $e_{\boldsymbol{\phi}}$ measures the discrepancy between a distance-labelled point $(\boldsymbol{x}, d) \in \mathcal{X}_k(\boldsymbol{p})$ observed close to surface and the coarse shape $\mathcal{E}_{\boldsymbol{u}}$ provided by $\boldsymbol{u} = g_{\boldsymbol{\phi}}(\boldsymbol{z})$.
- $e_{\boldsymbol{\theta}}$ is used for the difference between $(\boldsymbol{x}, d)$ and the SDF value $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ predicted by the fine shape model.

## Object Pose and Shape Optimization

- **Training phase**: optimize parameters $\boldsymbol{z}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ of object class using offline data, from instances with known meshes.
- **Testing phase**: optimize the pose $\boldsymbol{T}$ and shape deformation $\delta\boldsymbol{z}$ of a previously unseen instance from the same category using online distance data from an RGB-D camera.

**Training an ELLIPSDF Model:**

- Learn latent shape code shared by coarse shape decoder $g_{\phi}$ and fine shape decoder $f_{\theta}$.




**Joint Pose and Shape Optimization:**

- Given initial object transformation and shape deformation, solve joint object pose and shape optimization via **gradient descent**:

$$\boldsymbol{T}^{i+1} \triangleq \exp\left(-\eta_1 \frac{\partial e(\boldsymbol{T}, \delta\boldsymbol{z}, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*; \{\mathcal{X}_k(\boldsymbol{p})\})}{\partial \boldsymbol{\xi} \boldsymbol{i}}\right) \boldsymbol{T}^i,$$

$$\delta\boldsymbol{z}^{i+1} \triangleq \delta\boldsymbol{z}^i - \eta_2 \left(\frac{\partial e(\boldsymbol{T}, \delta\boldsymbol{z}, \boldsymbol{\theta}^*, \boldsymbol{\phi}^*; \{\mathcal{X}_k(\boldsymbol{p})\})}{\partial \delta\boldsymbol{z}}\right).$$

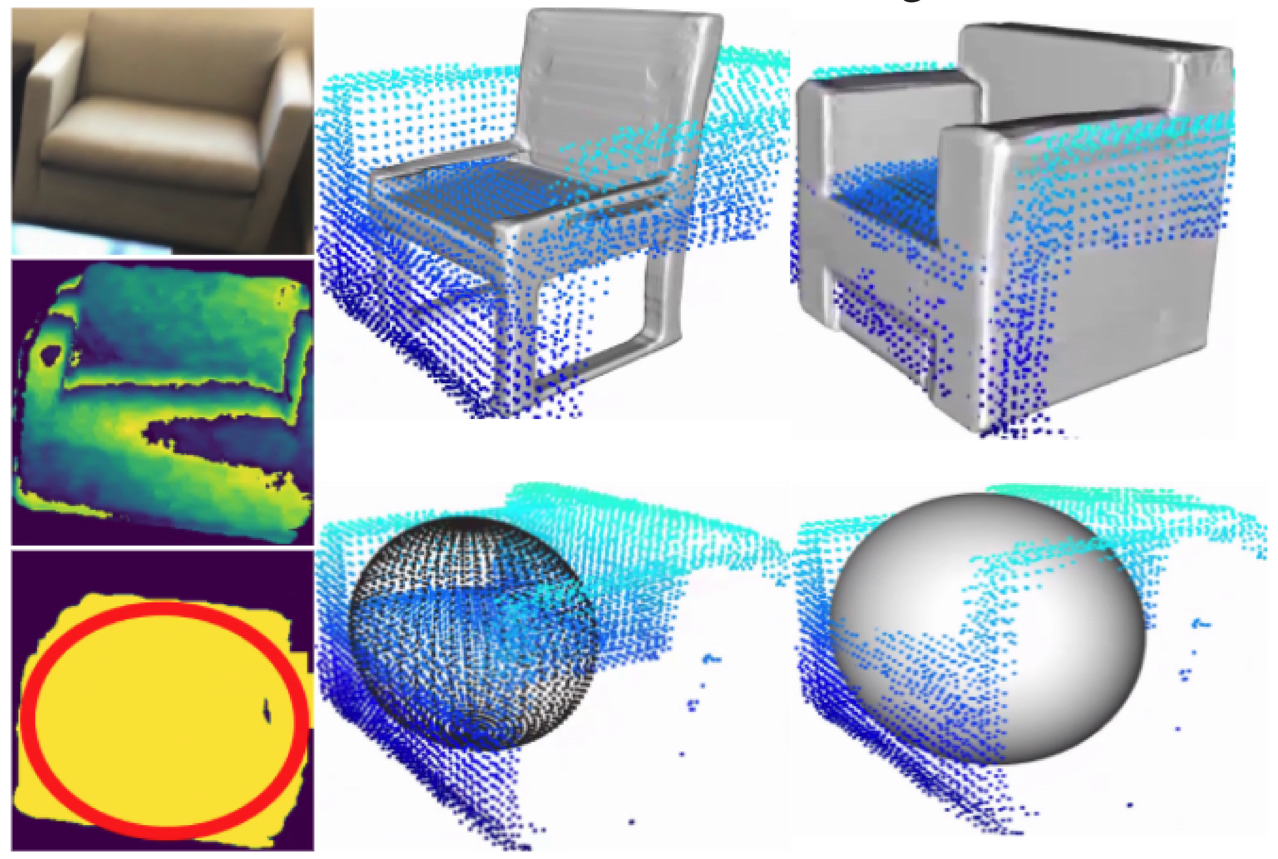
## Experiments & Results

- We evaluate ELLIPSDF on the **ScanNet dataset**, which provides 3D scans captured by a RGB-D sensor of indoor scenes with chairs, tables, displays, etc.

**Visualizations of Intermediate Results:**

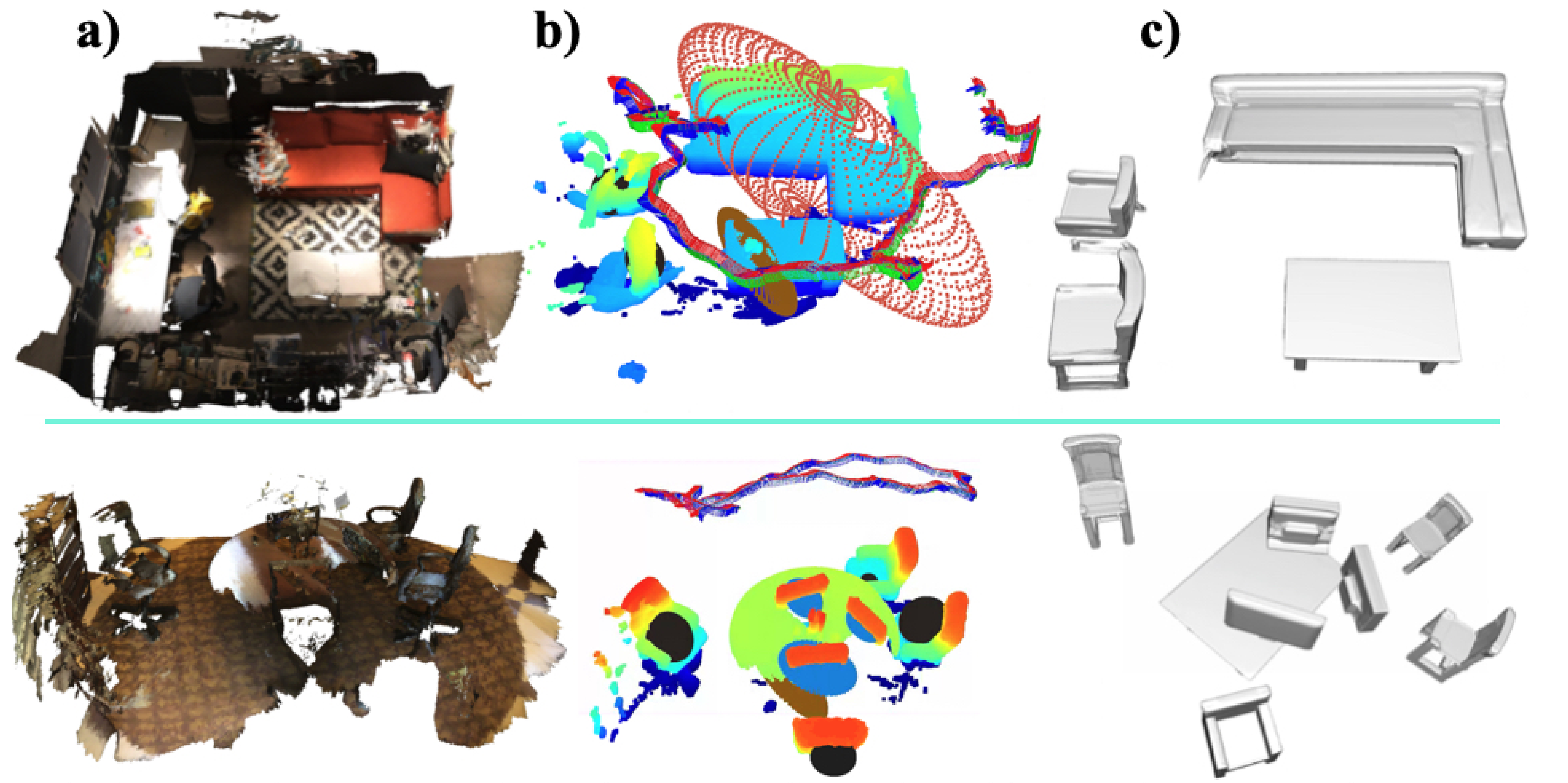
- The ELLIPSDF decoder model is trained on synthetic CAD models from **ShapeNet**.
- From left to right:
  - RGB image, depth image, instance segmentation (yellow), fitted ellipse (red) for a chair.
  - Mean shape and ellipsoid with initial pose.
  - Optimized fine-level and coarse-level shapes with optimized pose.
- Optimization step improves the scale and shape estimates notably, e.g. by transforming the four-leg mean shape into an armchair.

Intermediate ELLIPSDF stages.



**Qualitative Results on a larger scale:**

Column a): Ground-truth scene in ScanNet Sequences. Column b): The ellipsoids (black for chair, red for sofa, blue for monitor, brown for table) are the initialized objects. Column c): Reconstructed meshes using ELLIPSDF.




**Quantitative results for pose estimation on ScanNet:**

| Scan2CAD | Vid2CAD | ELLIPSDF (init) | ELLIPSDF (opt) |
|---|---|---|---|
| 31.7 | 38.3 | 31.5 | **39.6** |

**Quantitative results for shape evlaution on ScanNet:**

| Method<br># intances | cabinet<br>132 | chair<br>820 | display<br>209 | table<br>146 | avg.<br>327 |
|---|---|---|---|---|---|
| ELLIPSDF (fine) | 88.4 | 88.3 | 90.6 | 76.2 | 85.9 |
| ELLIPSDF (coarse+fine) | **91.0** | **90.6** | **96.9** | **77.3** | **89.0** |

**Comparison of 3D detection results on ScanNet:**

| mAP @ IoU=0.5 | Chair | Table | Display |
|---|---|---|---|
| FroDO | 0.32 | 0.06 | 0.04 |
| MOLTR | 0.39 | 0.06 | 0.10 |
| ELLIPSDF (fine) | 0.42 | 0.26 | 0.25 |
| ELLIPSDF (coarse+fine) | **0.43** | **0.27** | **0.31** |