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

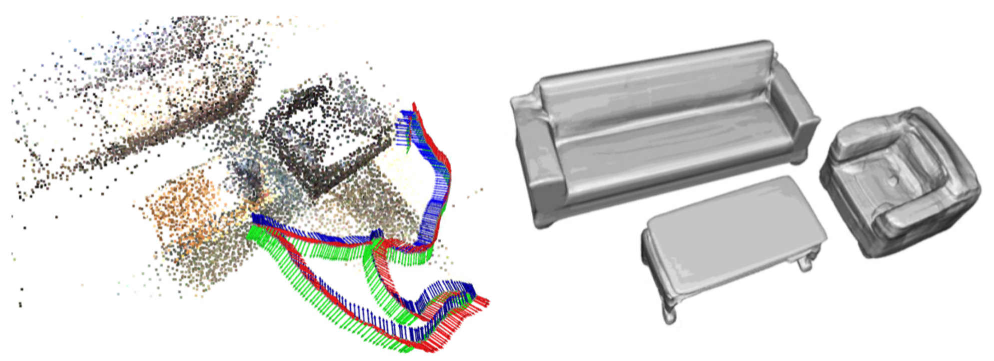
Motivations:

- Maps offering geometric and semantic information are useful and understandable for humans, allowing specification of symbolic tasks in terms of object entities.
- Striking the right balance between a faithful object reconstruction and a compact object representation remains an open research problem.

Contributions:

- A bi-level object model consisting of coarse and fine levels, which enables joint optimization of object pose and shape. The coarse-level uses a primitive shape for robust pose and scale initialization, and the fine-level model uses SDF residual directly to allow accurate shape modeling. The two levels are coupled via a shared latent space.
- 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 $\boldsymbol{c} \triangleq (\nu, \boldsymbol{z}, f_{\boldsymbol{\theta}}, g_{\boldsymbol{\phi}})$, where $\nu \in \mathbb{N}$ is the class identity, e.g., chair, table, sofa, and $\boldsymbol{z} \in \mathbb{R}^d$ is a latent code vector, encoding the average class 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 $g_{\boldsymbol{\phi}} : \mathbb{R}^d \mapsto \mathbb{R}^3$ 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 $f_{\boldsymbol{\theta}} : \mathbb{R}^3 \times \mathbb{R}^d \mapsto \mathbb{R}$ with parameters $\boldsymbol{\theta}$.
- An *object instance* of class $\boldsymbol{c}$ is a tuple $\boldsymbol{i} \triangleq (\boldsymbol{T}, \delta\boldsymbol{z})$, where $\boldsymbol{T} \in SIM(3)$ specifies the transformation from the global frame to the object instance frame, and $\delta\boldsymbol{z} \in \mathbb{R}^d$ is a deformation of the latent code $\boldsymbol{z}$, specifying the average shape of class $\boldsymbol{c}$.

Error Functions:

- Error function $e_{\boldsymbol{\phi}}$ measures the discrepancy between a distance-labelled point $(\boldsymbol{x}, d) \in \mathcal{X}_k(\boldsymbol{p})$ observed close to the instance surface and the coarse shape $\mathcal{E}_{\boldsymbol{u}}$ provided by $\boldsymbol{u} = g_{\boldsymbol{\phi}}(\boldsymbol{z})$. Error function $e_{\boldsymbol{\theta}}$ 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.
- Overall error function is defined as:

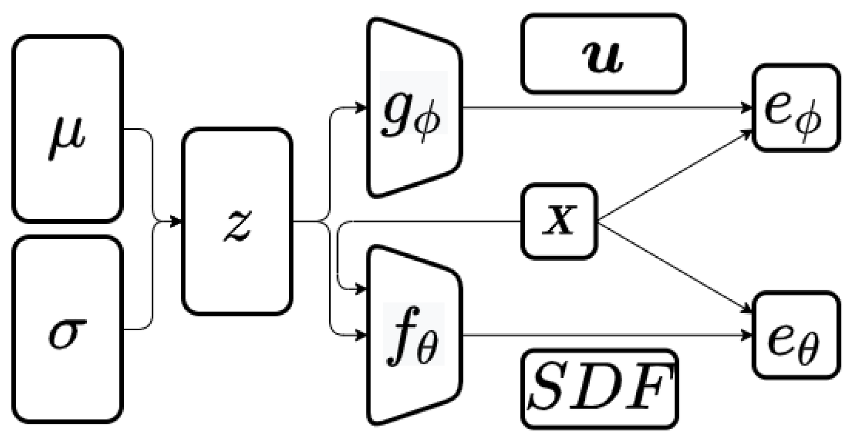
$$\alpha e_r(\delta \boldsymbol{z}) + \sum_k \sum_{\boldsymbol{p}} \sum_{(\boldsymbol{x},d)} \beta e_{\boldsymbol{\theta}}(\boldsymbol{x}, d, \boldsymbol{T}, \delta\boldsymbol{z}) + \gamma e_{\boldsymbol{\phi}}(\boldsymbol{x}, d, \boldsymbol{T}, \delta\boldsymbol{z}).$$

Object Pose and Shape Optimization

We distinguish between a training phase, where we optimize the parameters $\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}$ of an object class using offline data from instances with known mesh shapes, and a testing phase, where we 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:

- Latent shape code shared by coarse shape decoder g_ϕ and fine shape decoder f_θ:



- Fine-level shape error function $e_{\boldsymbol{\theta}}(\boldsymbol{x}, d, \boldsymbol{T}, \delta\boldsymbol{z}) \triangleq \rho(s f_{\boldsymbol{\theta}}(\boldsymbol{P}\boldsymbol{T}\underline{\boldsymbol{x}}; \boldsymbol{z} + \delta\boldsymbol{z}) - d)$. Coarse-level shape error $e_{\boldsymbol{\phi}}(\boldsymbol{x}, d, \boldsymbol{T}, \delta\boldsymbol{z}) \triangleq \rho(sh(\boldsymbol{P}\boldsymbol{T}\underline{\boldsymbol{x}}, g_{\boldsymbol{\phi}}(\boldsymbol{z} + \delta\boldsymbol{z})) - d)$.

Joint Pose and Shape Optimization:

- The Jacobian of $e_{\boldsymbol{\theta}}$ wrt transformation perturbation is:

$$\frac{\partial e_{\boldsymbol{\theta}}}{\partial \boldsymbol{\xi}} = \frac{\partial \rho(r)}{\partial r} \left(\hat{s}[\mathbf{0}_6, 1] f_{\boldsymbol{\theta}}(\boldsymbol{x}, \delta\hat{\boldsymbol{z}}) + \hat{s} \nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}(\boldsymbol{x}, \delta\hat{\boldsymbol{z}})^\top \boldsymbol{P} \left[\hat{\boldsymbol{T}}\underline{\boldsymbol{x}}\right]^{\odot} \right)$$

$$\frac{\partial e_{\boldsymbol{\theta}}}{\partial \delta\tilde{\boldsymbol{z}}} = \frac{\partial \rho(r)}{\partial r} \hat{s} \nabla_{\boldsymbol{z}} f_{\boldsymbol{\theta}}(\boldsymbol{x}, \delta\hat{\boldsymbol{z}}).$$

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

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

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

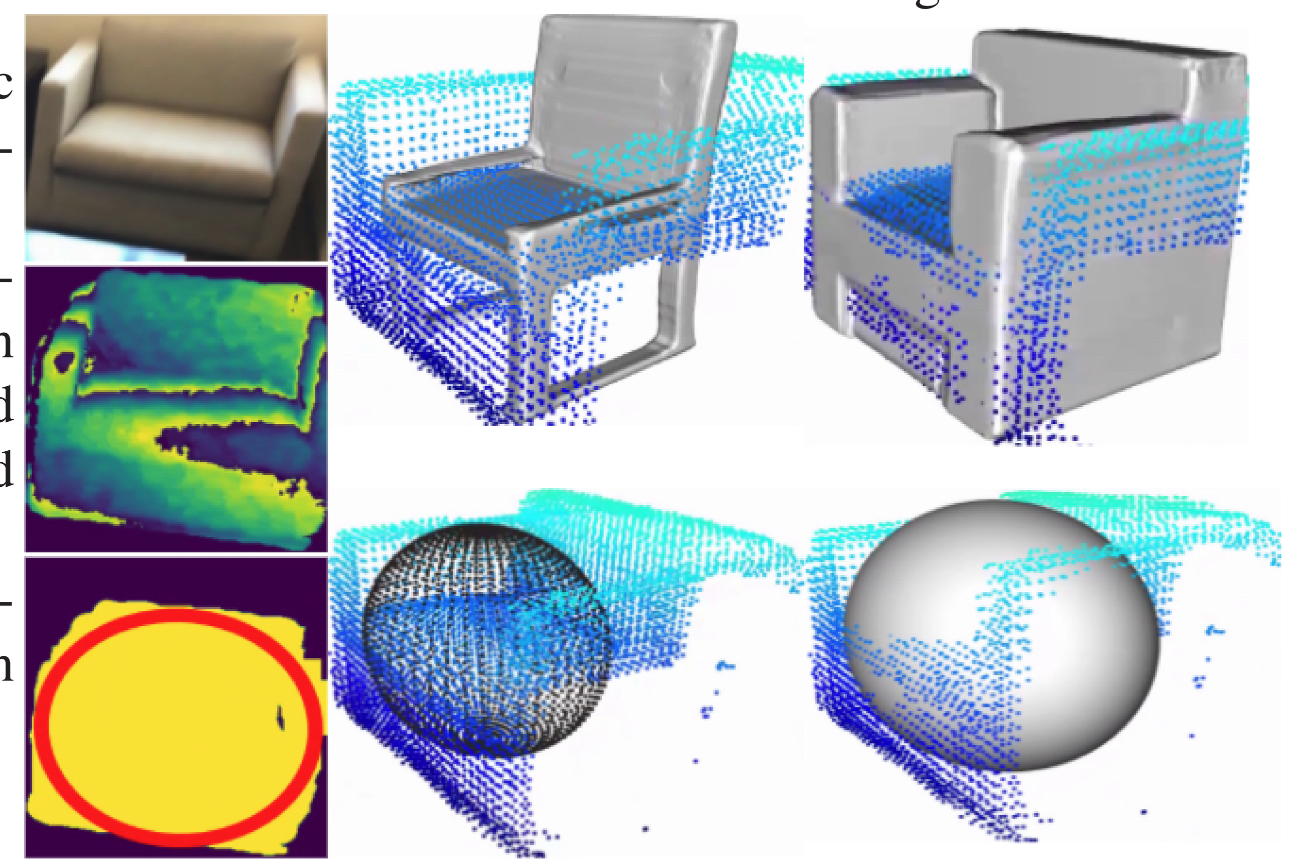
Experiments & Results

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. We segment out the objects from the scene-level mesh using provided instance labels and sample points from the object meshes to generate the point observations.

Visualizations of Intermediate Results:

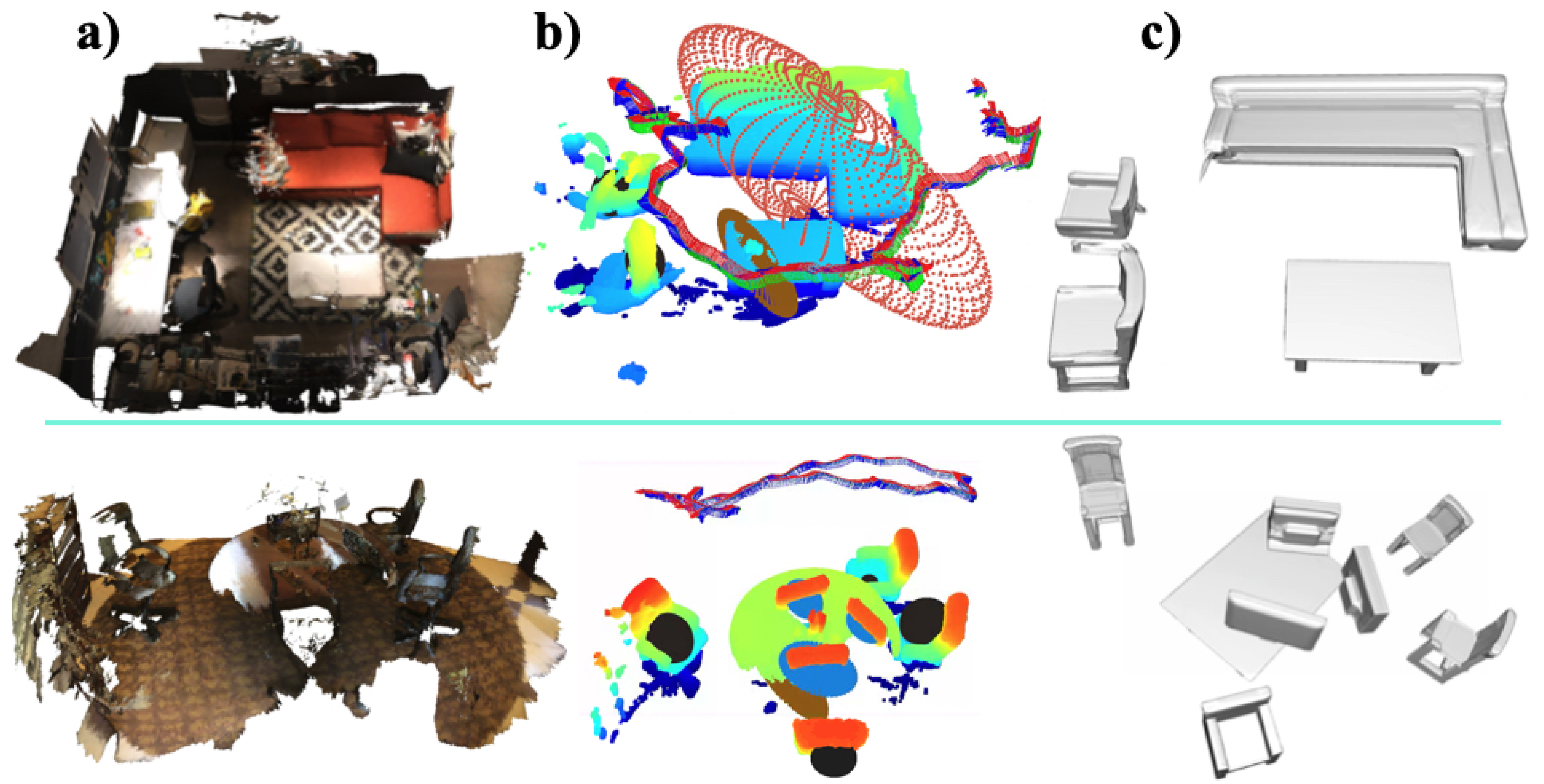
- The ELLIPSDF decoder model is trained on synthetic CAD models from ShapeNet. Each model's scale is normalized to be inside a unit sphere.
- First column: RGB image, depth image, instance segmentation (yellow), fitted ellipse (red) for a chair in ScanNet scene 0461. Second column: mean shape and ellipsoid with initialized pose. Third column: 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 # intances	cabinet 132	chair 820	display 209	table 146	avg. 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**