Planes vs. Chairs: Category-guided 3D shape learning without any 3D cues

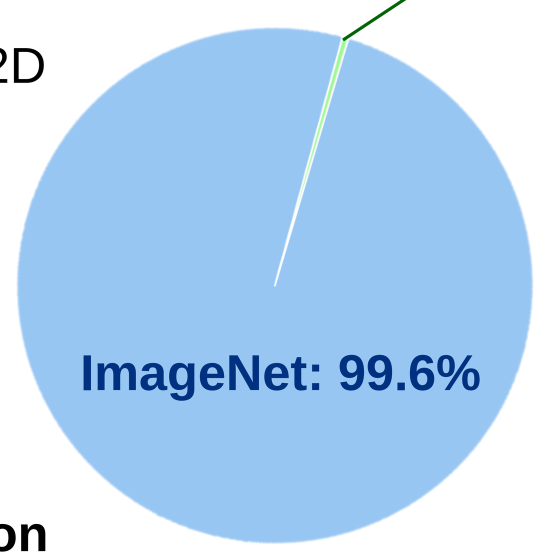
Zixuan Huang[1], Stefan Stojanov[1], Anh Thai[1], Varun Jampani[2], James M. Rehg[1]

[1]Georgia Institute of Technology, [2]Google Research

ECCV TEL AVIV 2022

Code available

Motivation



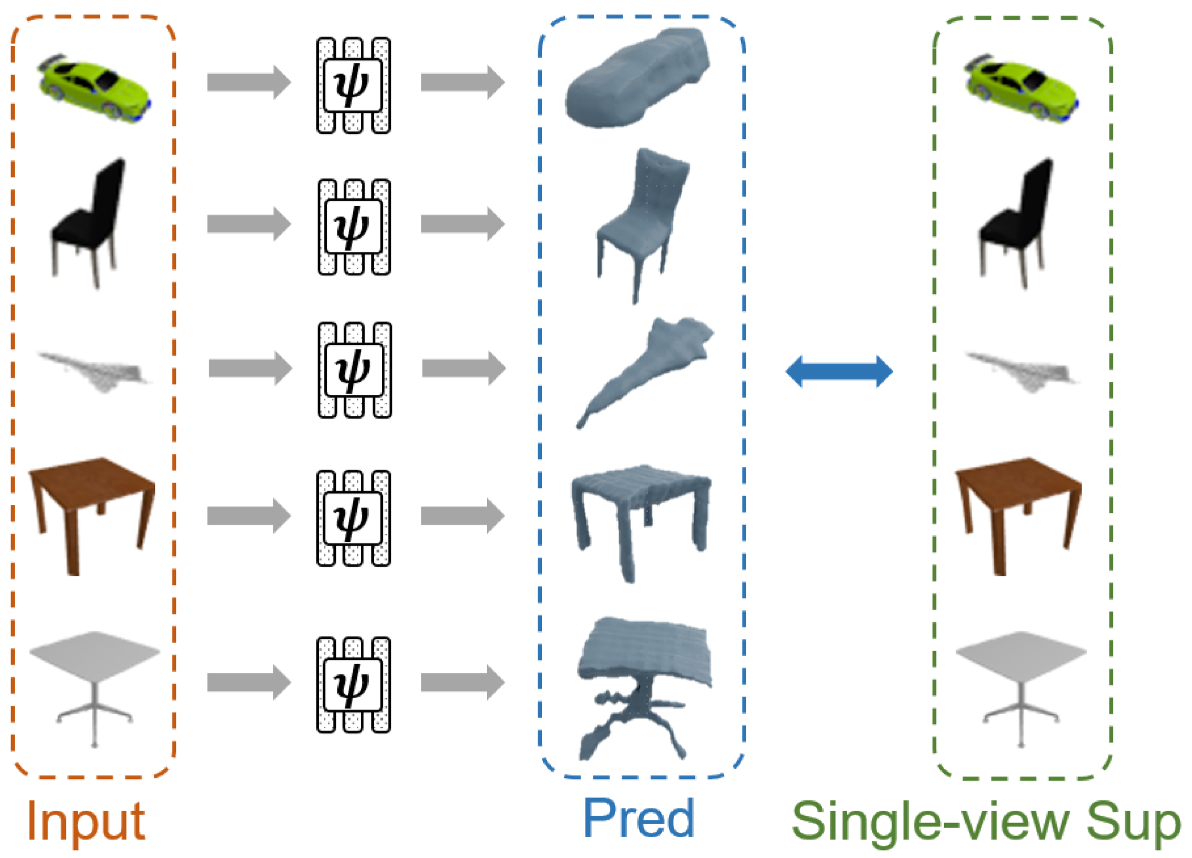
Large-scale annotated data empowered the great success of learning-based method in 2D computer vision tasks.

However, 3D reconstruction from single images is still quite challenging
- ➢ One key reason is the lack of annotated data at scale.

How can we learn 3D shape reconstruction in a more scalable way?

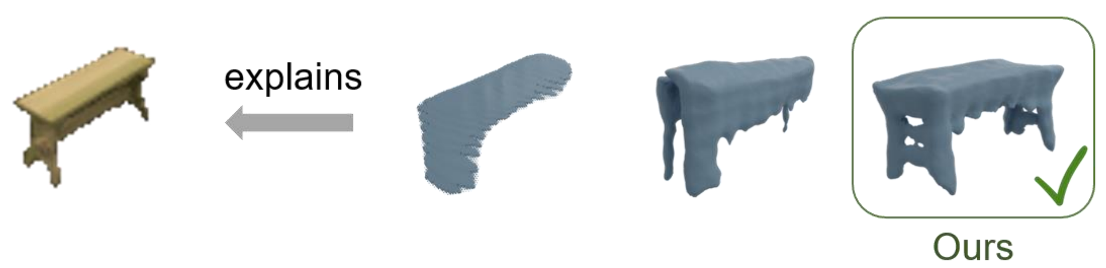
Problem Framing

We propose learning under Multi-Category Single-View (MCSV) setting:
- ➢ No 3D cues such as viewpoints or multiple views for supervision;
- ➢ Learn a single unified model that works for all the categories.



MCSV learning
- ➢ is more scalable
- ➢ enables data pooling to learn category-agnostic features

But this learning setting makes the shape-viewpoint entanglement problem even harder to solve:

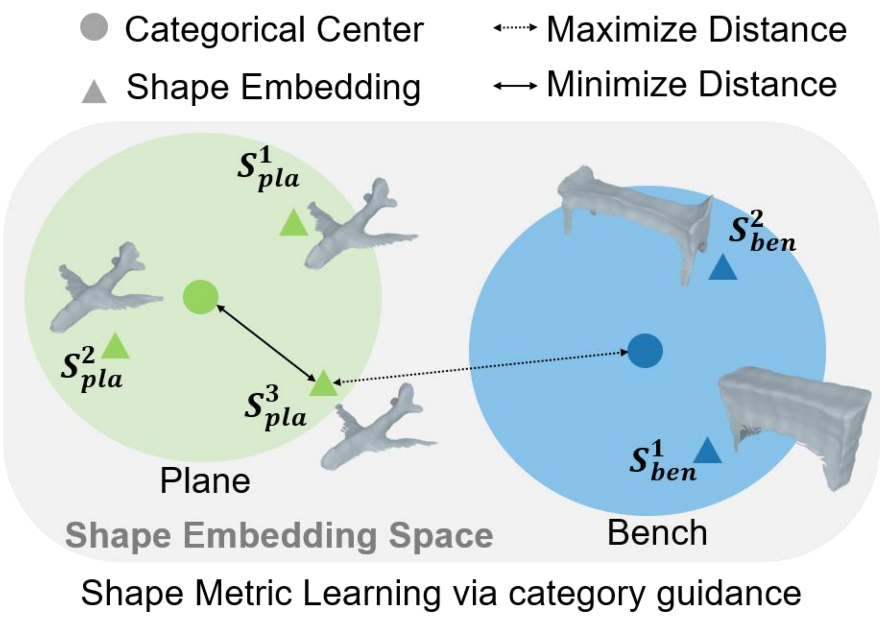


Can we better constrain the shape learning?

A 3D-unsupervised model that learn shapes of multiple object categories at once.

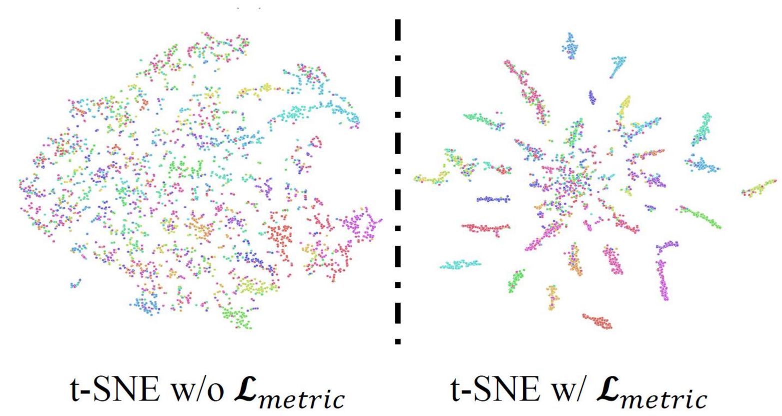
Project Page: https://zixuanh.com/multiclass3D

Categorical Metric Learning



Shape Metric Learning via category guidance

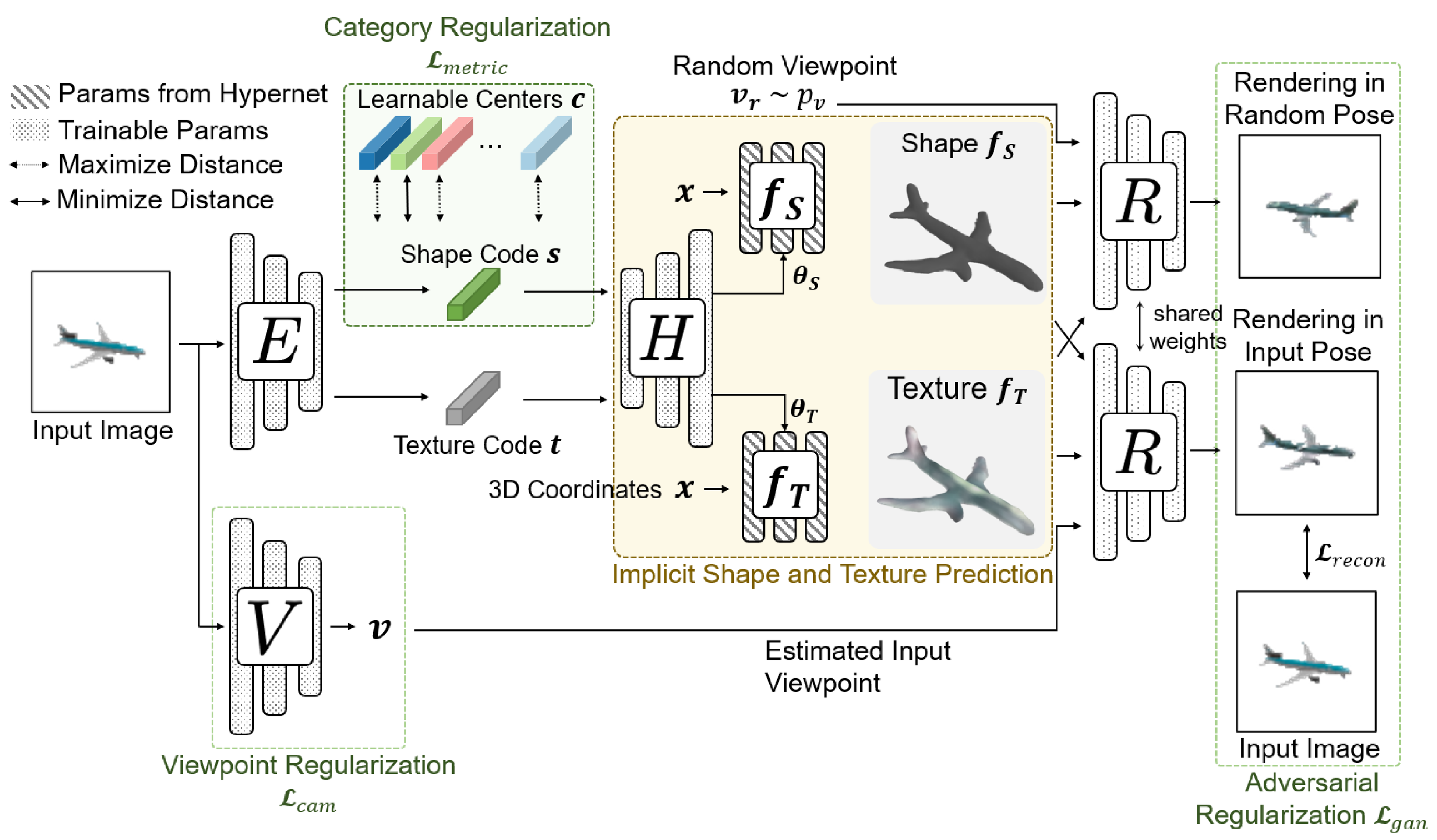
We regularize the shape learning via category labels.



$$\mathcal{L}_{metric} = -\sum_{i=0}^{N} log \frac{exp(d(\boldsymbol{s}_i, \boldsymbol{c}_{y_i})/\tau)}{\sum_{k \in C} exp(d(\boldsymbol{s}_i, \boldsymbol{c}_k)/\tau)}$$

- ➢ Introduce a distance prior that eliminates erroneous shapes.
- ➢ Facilitate the supervision received by a particular instance to affect its neighbors.

Architecture



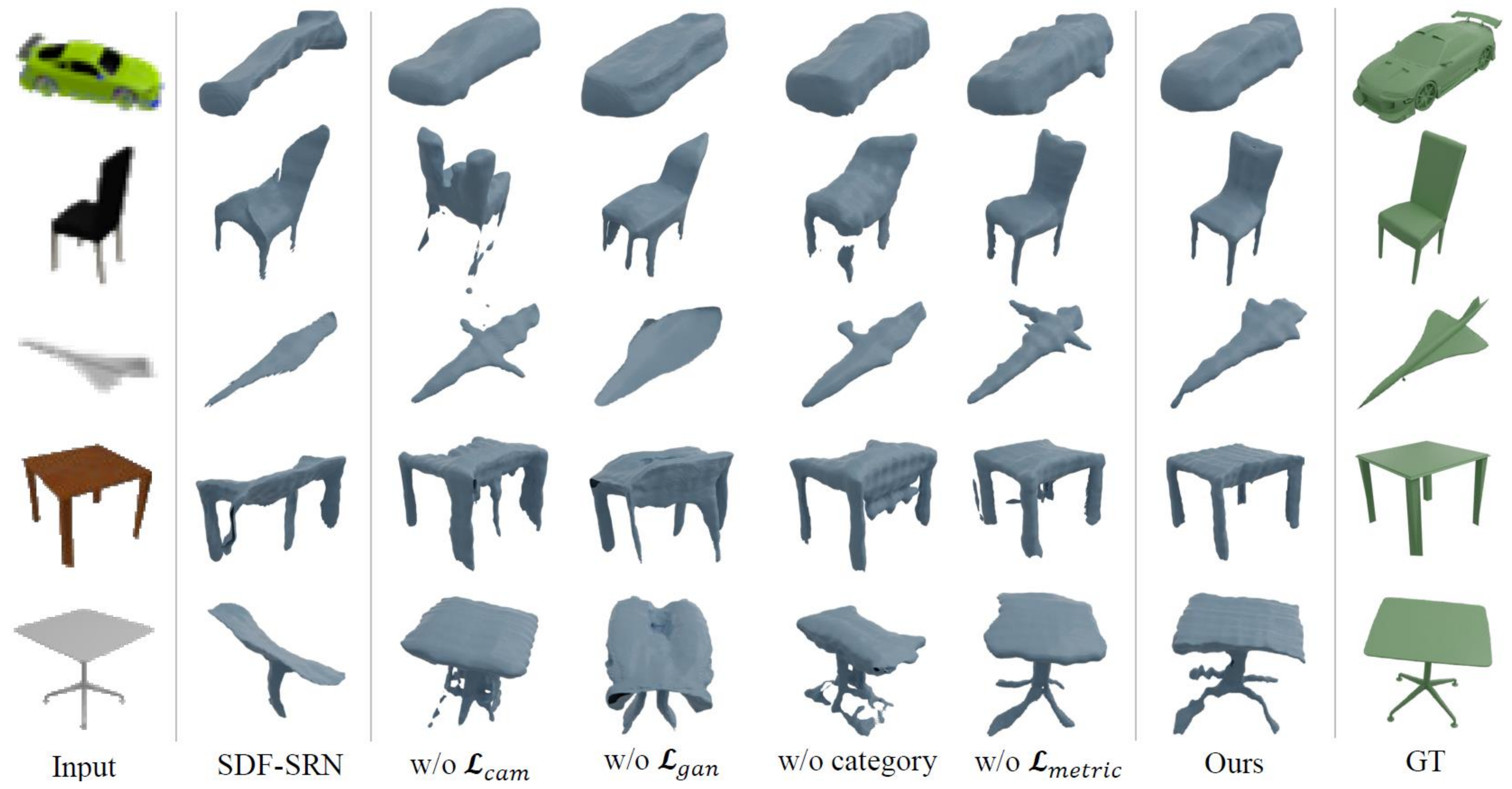
- ➢ *E:* image encoder
- ➢ f_S : implicit SDF MLP
- ➢ f_T : implicit RGB MLP
- ➢ *H:* hypernetwork
- ➢ *V:* viewpoint estimator
- ➢ *R:* learnable renderer

Results

Quantitative ablation and SOTA comparison on ShapeNet-13:

Methods	F-Score@1.0↑	F-Score@5.0↑	F-Score@10.0↑	CD↓
w/o category	0.1589	0.6261	0.8527	0.520
w/o $\mathcal{L}_{metric}$	0.1875	0.6864	0.8805	0.458
w/o $\mathcal{L}_{cam}$	0.1837	0.6741	0.8758	0.463
w/o $\mathcal{L}_{gan}$	0.1846	0.6437	0.8422	0.532
Ours	**0.2005**	**0.7168**	**0.8949**	**0.430**
SDF-SRN	0.1606	0.5441	0.7584	0.682

Qualitative ablation and SOTA comparison on ShapeNet-13:



Results on ShapeNet-55:

Results on Pascal3D+:

Limitation:
- ➢ training instability due to the adversarial regularization, particularly on real-world images with many categories