Interpretable3D: An Ad-Hoc Interpretable Classifier for 3D Point Clouds

Tuo Feng, Ruijie Quan, Xiaohan Wang, Wenguan Wang, Yi Yang

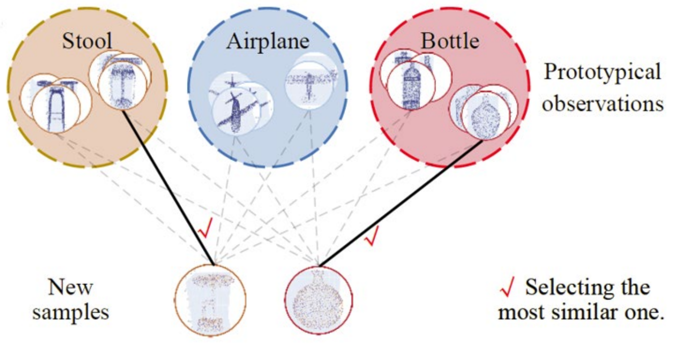
Problems of existing studies

1) Existing explanation studies on 3D models have been conducted with *post-hoc* explanations. However, *post-hoc* explanations are problematic:

- ☐ requiring **a separate modeling effort**
- ☐ **varying explanations** for different explanation models
- ☐ **cannot provide a reasoning process**
- ☐ **not interpretable to humans**

2) Parametric softmax classifier **learns highly abstract parameters** and **lacks a direct and intuitive interpretation**.

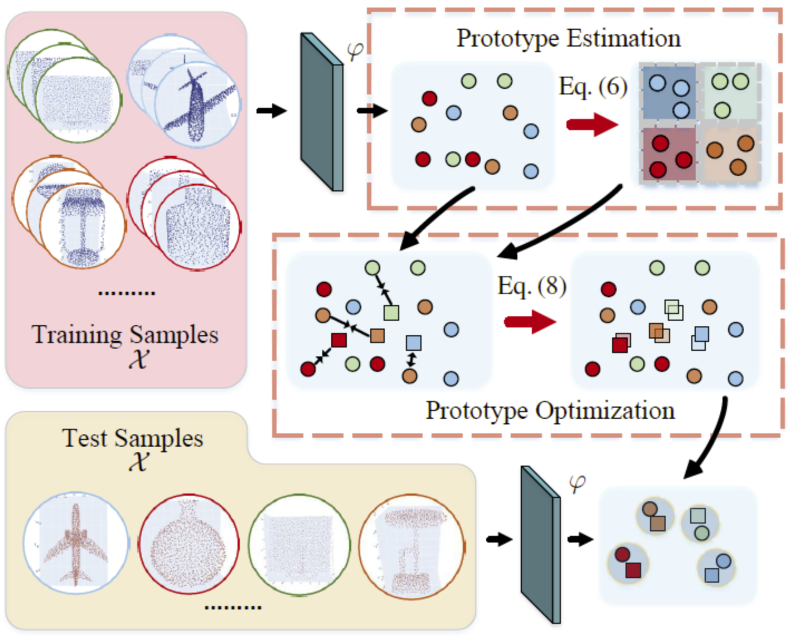
Why Interpretable3D



Interpretable3D

- ☐ provides **self-explanation** without *post-hoc* analysis and
- ☐ achieves **comparable performance** compared to softmax-based models.

The overview of our Interpretable3D



Case-based paradigm:
Prototypes can be interpreted as typical observations.

Training:
Two iterative training steps:

- ☐ Prototype Estimation
- ☐ Prototype Optimization

Testing:
Classifying new samples according to prototypes.

Prototype Estimation

Assignment matrix $\mathbf{A}^l$ is obtained by optimizing the similarity $\mathbf{Q}^l$ between the features and cluster centers:

$$\boldsymbol{A}^{l*} = \arg\min_{\boldsymbol{A}^l \geq 0} \langle \boldsymbol{Q}^l, \boldsymbol{A}^l \rangle_F,$$

The entropic regularization of this problem can be formulated as:

$$\min_{\boldsymbol{A}^l \geq 0} \langle \boldsymbol{Q}^l, \boldsymbol{A}^l \rangle_F - \zeta H(\boldsymbol{A}^l)$$

It can be solved by Sinkhorn-Knopp.

Prototype Optimization

To pursue more representative prototypes, the winning prototypes $\mathbf{M}^w$ are altered:

$$\boldsymbol{M}^w \leftarrow \boldsymbol{M}^w + \eta\,\psi(l, \hat{l}_w)(\boldsymbol{F}^l - \boldsymbol{M}^w),$$

$$\psi(l, \hat{l}_w) = \begin{cases} +1 & \text{if } l = \hat{l}_w \\ -1 & \text{else} \end{cases},$$

If $\mathbf{M}^w$ predicts correctly, it is rewarded with $\mathbf{F}^l$. Conversely, $\mathbf{M}^w$ is moved away from $\mathbf{F}^l$.

A momentum update strategy is applied to update the prototype, specifically, updating with the average of embeddings of each sub-class. In the last few epochs, we update prototypes with their most similar observations.

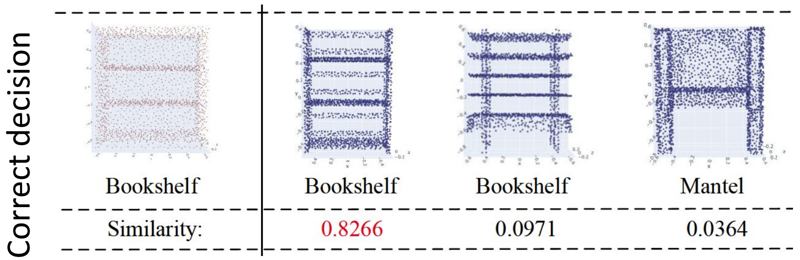
Classification results

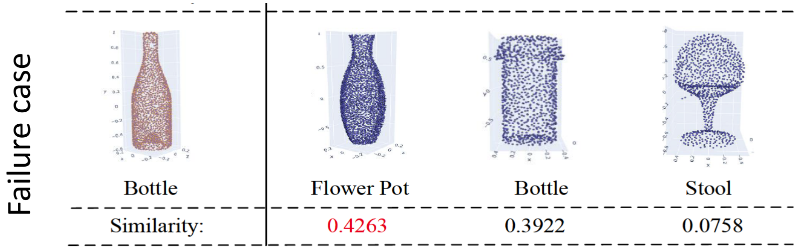
Our algorithm shows comparable performance:

Method	OA(%)	mAcc(%)
PointNet (Qi et al. 2017a)	89.2	86.0
PointNet++ (Qi et al. 2017b)	90.7	-
PointNet2 (Yan 2019)	92.2	-
PointNet2 + **Ours**	93.2	89.3
DGCNN (Phan et al. 2018)	92.9	90.2
DGCNN + **Ours**	93.5	90.3
PointMLP (Ma et al. 2021)	94.1	91.3
PointMLP + **Ours**	94.1	92.0
PointNeXt (Qian et al. 2022)	94.0	91.1
PointNeXt + **Ours**	94.3	91.8

How Interpretable3D makes decision

The decision-making mode is straightforward for users.





Codes and pre-trained models are at
https://github.com/FengZicai/ Interpretable3D