# S1. Supplementary Material

## S1.1. Non-generative AM and baselines

In this section we provide additional descriptions for Fig. 1. Consider the general term of AM updating according to the gradient as

$$x \leftarrow r_\theta \left( x + \eta \frac{\partial a_i}{\partial x} \right) \tag{S1}$$

where $x$ and $a_i$ denote the input to be optimized and the target activation respectively, $\eta$ is the learning rate and $r_\theta$ indicates the regularization term. The non-prior refers to AM without any regularization ($r_\theta = 1$). L2 Norm avoids the dominance of a few extreme pixel values in the whole image, which changes $r_\theta$ to $x \cdot (1 - \theta_{decay})x$ to penalize those extreme pixels. Gaussian blur replaces $r_\theta$ with a blur kernel

$$Gauss(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \tag{S2}$$

in order to penalize high frequency noise, where $\sigma$ denotes the standard deviation. The Total variation regularization adds a term $\frac{\partial TV(x)}{\partial x}$ that minimizes the total variation in the gradient which updates $x$, where

$$TV(x) = \sup_P \sum_{i=0}^{n_p - 1} |x_{i+1} - x_i| \tag{S3}$$

The Average initialization takes the mean value of the whole dataset as the initial example, thereby simplifying the optimization difficulty and allowing the generated examples to be closer to the real data. To show more clearly the defects of traditional regularization in the point clouds, we illustrate additional results in Fig. S1.

In table 1 we compare two categories of baselines, initialization and regularization. The initialization baseline only adjusts the initial position of AM without interfering with the optimization process. The *zero* baseline is the one that takes an all-zero matrix as the initial input. The *random* baseline is generated by sampling random points in a proper value domain ($[-1, 1]$ for ModelNet40). The *average* baseline utilizes with the aforementioned Average initialization as input. The *instance* baseline selects a random instance of the same class as the target activation from the dataset as input. For regularization baselines, we start with zero as the initial input and regularize using *L2 Norm*, *Gaussian blur* and *TV* as described above, respectively.

## S1.2. Detailed Architecture

The detailed structures of AE, AED and NAED are illustrated in Figs. S4, S5 and S6, respectively.

## S1.3. More qualitative AM results

As a complement to Fig. 3, Fig. S2 demonstrates more AM qualitative results.

## S1.4. Qualitative results for ShapeNet

Figure A exhibits examples of AM of the class "aircraft" generated by the proposed methods. S3

## S1.5. Ablation Study

To confirm the effectiveness, we perform ablation studies on each module. Compared to AE, AED has two main additional modules, i.e. the discriminator $D$ and the hidden distance loss $L_F$. On the basis of AED, NAED involves two additional modules, which are Gaussian noise added during training $Z_D$ and the second latent distance loss $L_{F2}$. Table S1 reports the results of the ablation tests for the four modules mentioned above. With similar point-wise distances, the utilization of $L_f$ tends to generate examples that are richer in diversity, while the employment of $D$ prefers examples that are monotonic and closer to real instances. On the other hand, incorporating Gaussian noise $Z_D$ at training significantly improves diversity, while inserting an additional hidden distance loss $L_{F2}$ improves the quality of the examples to some extent. Note that the ablation studies may not accurately reflect the real performance of the modules for the following two reasons:

- In the same way that perturbation-based explainability methods suffer from the bias towards highly correlated features, there are interactions between the modules in the model. The ablation of a single module neglects the correlation with other modules and therefore bias may also exist. Unfortunately, as neural networks are complex black boxes, it is impossible to disentangle how the modules affect each other at this stage.

- The definition of "perceptibility" is not unique. PC-AMs involves diversity, representativeness, latent & point-wise distances, whereas their weights are not uniform. The reason is that explanations are human-oriented and subjective. For example, regarding an explaining example, one part considers the representation to be more intuitive to enhance the comprehension of the model while the other believes that the explanations can only be recognized if they are closer to real objects.

## S1.6. Evaluation metrics for Point Clouds AM

For space limitation, we demonstrate the deficiency of FID in measuring the latent distance of point clouds in figure S7, as the argument of section 4.2.

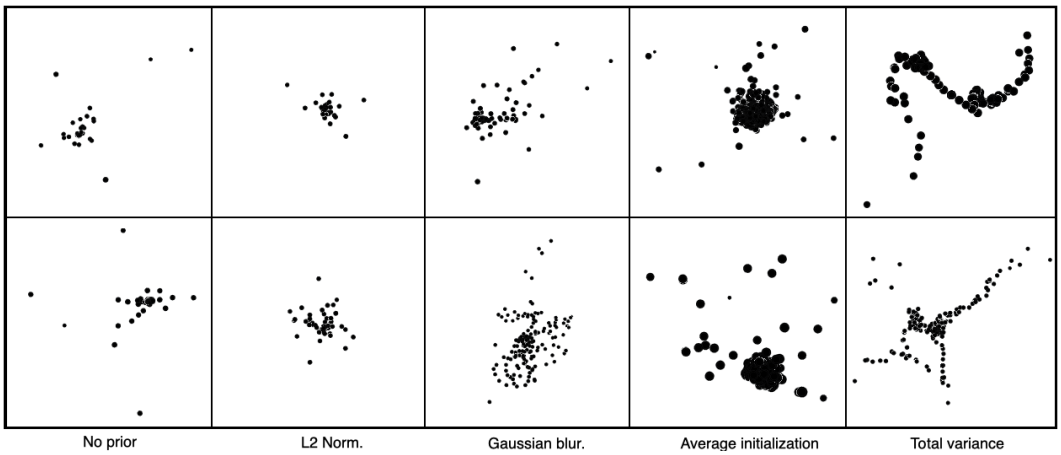


Figure S1. Additional demonstration of non-generative AM for point clouds. The target classes for the first and second line are "airplane" and "table", respectively. As a conclusion, the non-generative regularization fails to facilitate the point cloud AM to yield perceptible examples.

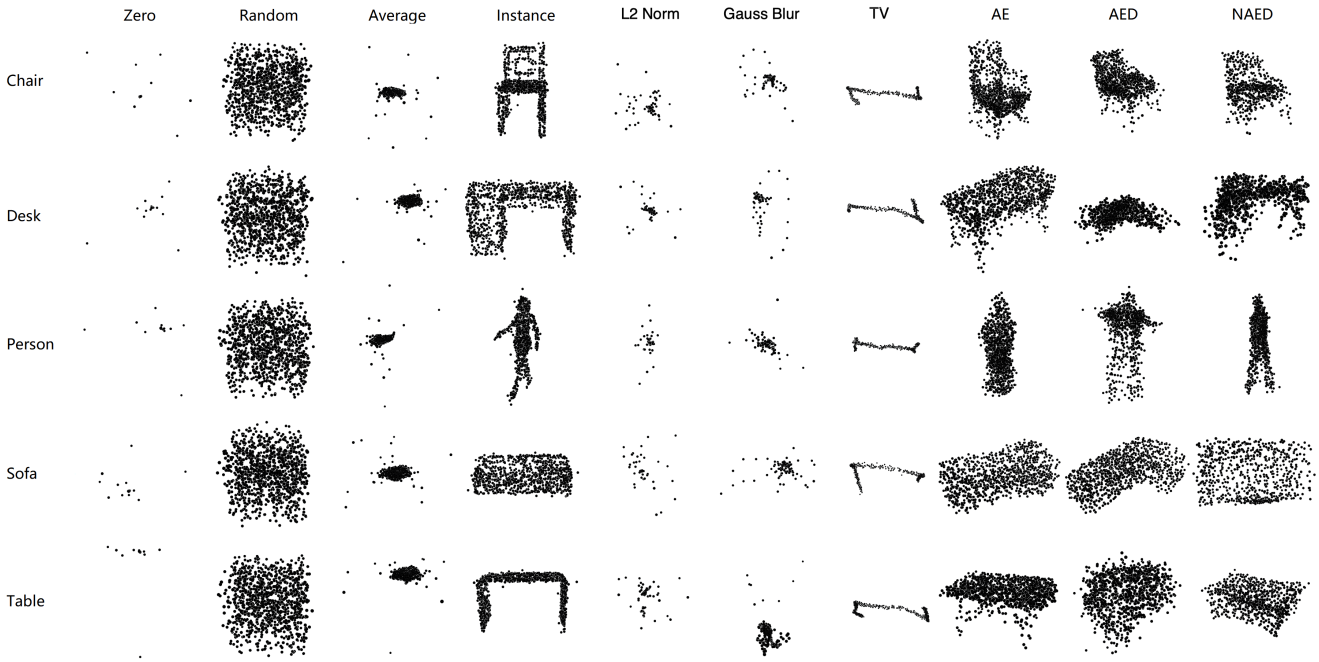


Figure S2. Additional AM results of different approaches. From left to right: Zero initialization, random initialization, initialized with the average of the test data per class, initialized from a specific instance, regularizations via L2 Norm, Gaussian Blur and Total Variation, and our proposed AE, AED and noisy NAED. Apparently, except for the instance initialization, the non-generative model-based approaches suffer from serious flaws in perceivability of AM examples (the first three columns). However, the AM example initialized from a certain instance lacks the "global" property, and the generated examples are unrepresentative.

## S1.7. Data reviewing with AM

We present an example of data reviewing with AM for ModelNet40 in figure S8. The AM instances in the middle are generated by NAED.

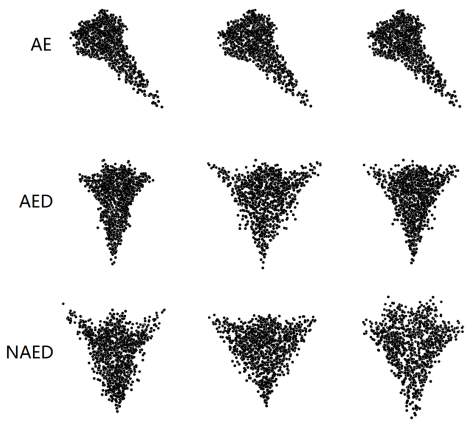


Figure S3. AM examples from AE, AED and NAED respectively of class "airplane" in ShapeNet. The qualitative performances of AE, AED and NAED are comparable to those on ModelNet40.

| | Module | m-IS | FID | CD | PC-AMs |
|---|---|---|---|---|---|
| AED | $\overline{\mathcal{L}_D}$ | 1.333 | 0.026 | 0.074 | 4.44 |
| | $\overline{L_F}$ | 1.077 | **0.010** | 0.089 | 4.56 |
| | $All_{AED}$ | 1.124 | 0.018 | 0.086 | 4.37 |
| NAED | $\overline{Z_D}$ | 1.204 | 0.021 | **0.059** | 4.52 |
| | $\overline{L_{F2}}$ | 1.410 | 0.019 | 0.069 | 4.71 |
| | $All_{NAED}$ | **1.461** | 0.014 | 0.074 | **4.89** |

Table S1. Ablation study for AED and NAED. The upper and lower halves are the ablation tests for the AED and NAED, respectively. For AED, the models from top to bottom are, in order: AED without discriminator $D$, AED without latent distance loss $L_F$, and AED with both. For NAED, the models are, in order, NAED without Gaussian noise, NAED without latent distance loss $L_F2$, and NAED with both.

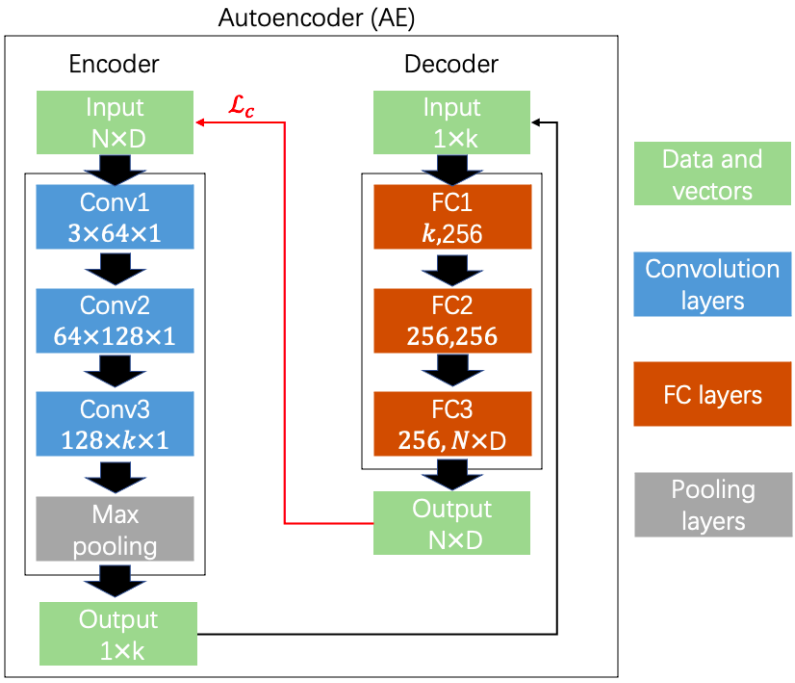


Figure S4. Architecture of AE. In the box are the components of AE, and on the right is the legend. In the plot, the black arrows represent the data flow and the red arrows represent the calculation process of the loss.

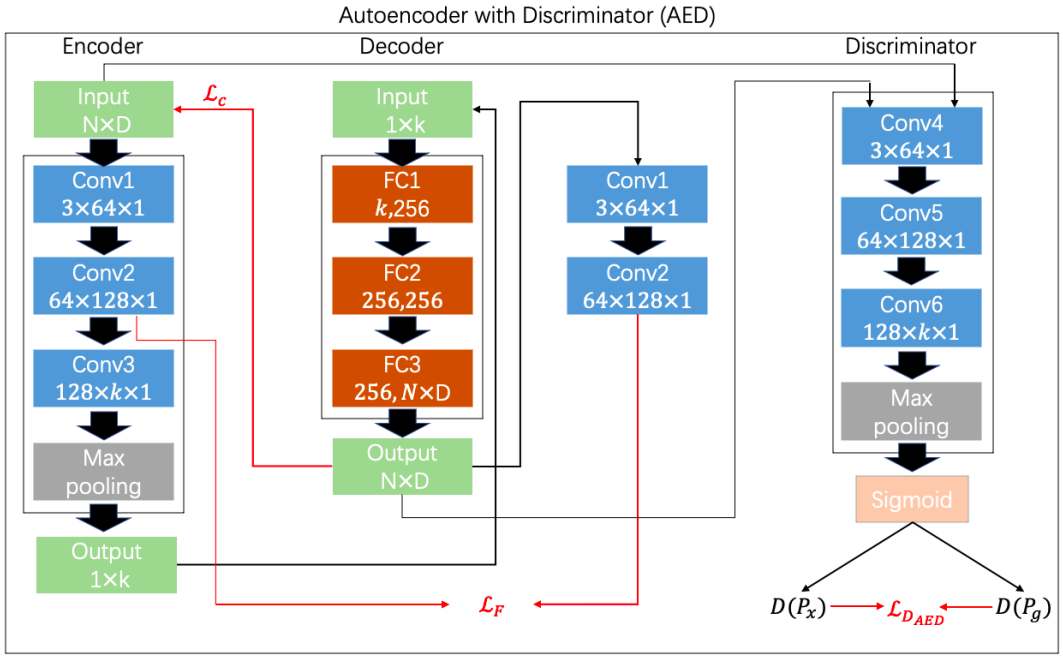


Figure S5. Architecture of AED.

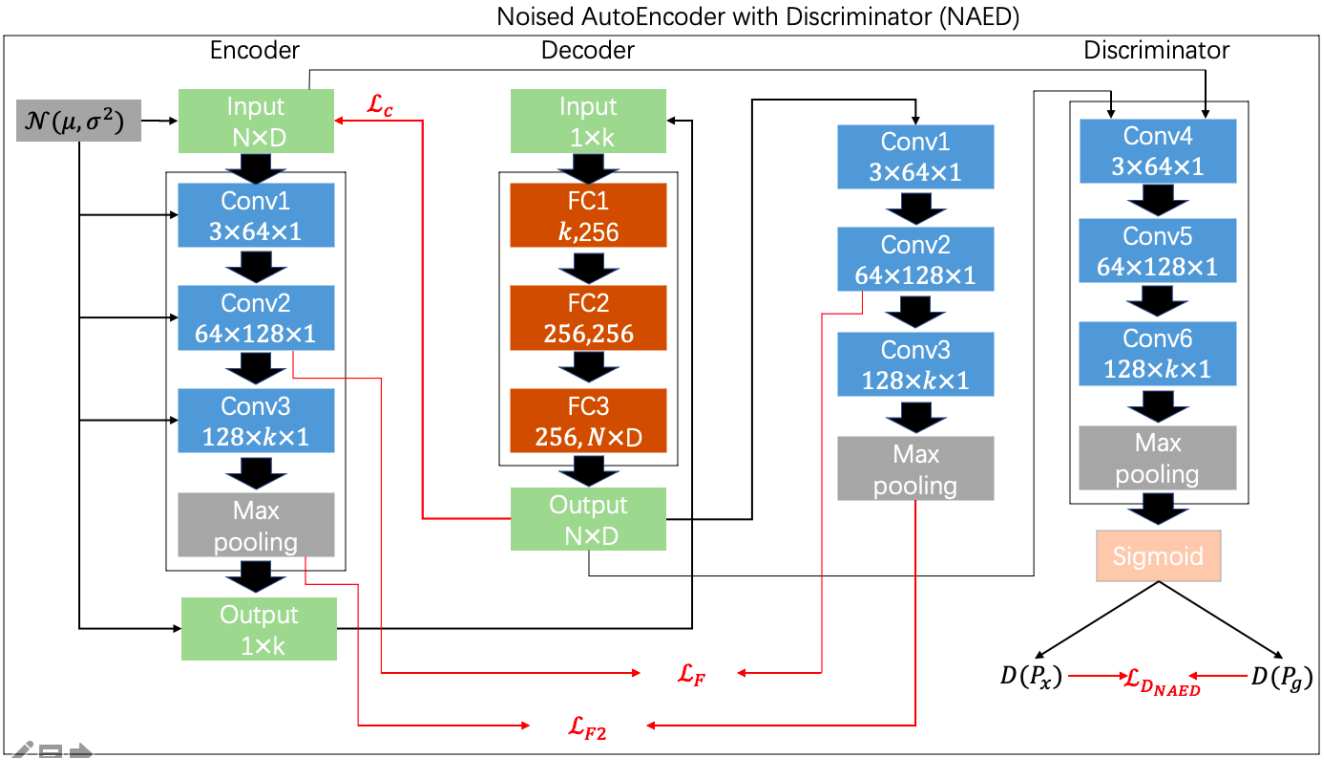


Figure S6. Architecture of NAED. The $\mathcal{N}(\mu, \sigma^2)$ in the upper left represents Gaussian noise with mean and variance of $\mu$ and $\sigma^2$, respectively.

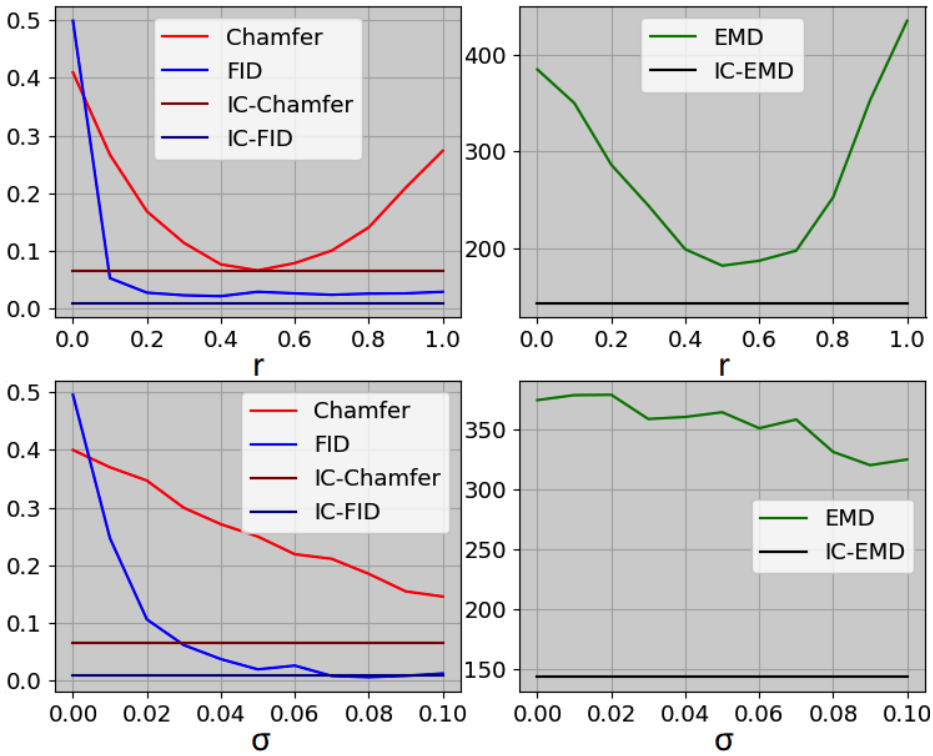


Figure S7. CD, FID and EMD metrics of instances generated by various parameters (the first two plots and the last two plots are uniform and Gaussian distributions respectively), where IC denotes intra-class, which is the average of corresponding distances between **real** objects in the dataset. $r$ denotes the interval parameter of uniform distribution and $\sigma$ denotes the variance of Gaussian distribution. The mean of Gaussian distribution is set to zero to ensure that the generated instances are symmetric about the zero point of the spatial coordinates. The larger the difference between the non-IC and the IC curves of the corresponding metrics, the better the method is capable of distinguishing random examples from real ones. Obviously, when $r > 0.1$ and $\sigma > 0.04$, the FIDs are almost indistinguishable from the averaged intra-class distance. However, it can be seen from CD and EMD that there are still significant discrepancies between the randomly distributed examples and the real ones.

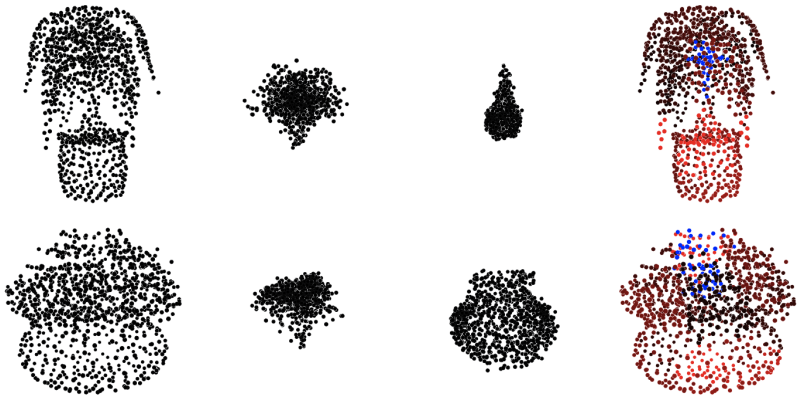

Figure S8. An example of reviewing the inaccuracies of the dataset. The first column shows the instances in the dataset that are labeled as "plant" but are classified as "vase". The second and third columns demonstrate the AM output for the categories "plants" and "vases" respectively. The last column exhibits an explanation generated from 3D LIME, where brighter red points represent more positive attributions while conversely brighter blue points represent more negative attributions, neutral attributed points are colored as black.