

# Successful Contrastive Pretraining of Set Transformer

**Divyashree Koti**

Middle Tennessee State University

*dsk2v@mtmail.mtsu.edu*

*Advisor:*

**Dr. Joshua Lee Phillips**

August 22, 2023



Computational Finance

# Presentation Overview

## ① Introduction

Data

## ② Background

## ③ Methods

## Experimental Setup - Hyperparameters

## Standard Set Transformer Algorithm

## Contrastive Learning using Set Transformer

## Fine Tuning

## Generalization

## Visualization

## 4 Results

## Set Transformer

## Contrastive Learning using Set Transformer

## Fine Tuning

## Generalization

## Visualization

## ⑤ Conclusion



# Introduction

- Transformer - To process long sequences of data in NLP.
- Vision Transformer - Processes images by creating patches and the corresponding vector.
- Set Transformer - A permutation invariant model to process unordered sets, which is independent of positional embeddings makes it permutation equivariant, in contrast to Transformer and Vit.
- Contrastive Learning with Set Transformer - a unique approach to enhance the Set Transformer performance, which has been successfully contributed in Transformer and Vision Transformer models.
- Generalization
  - Weak generalization
  - Strong generalization



Computational Science

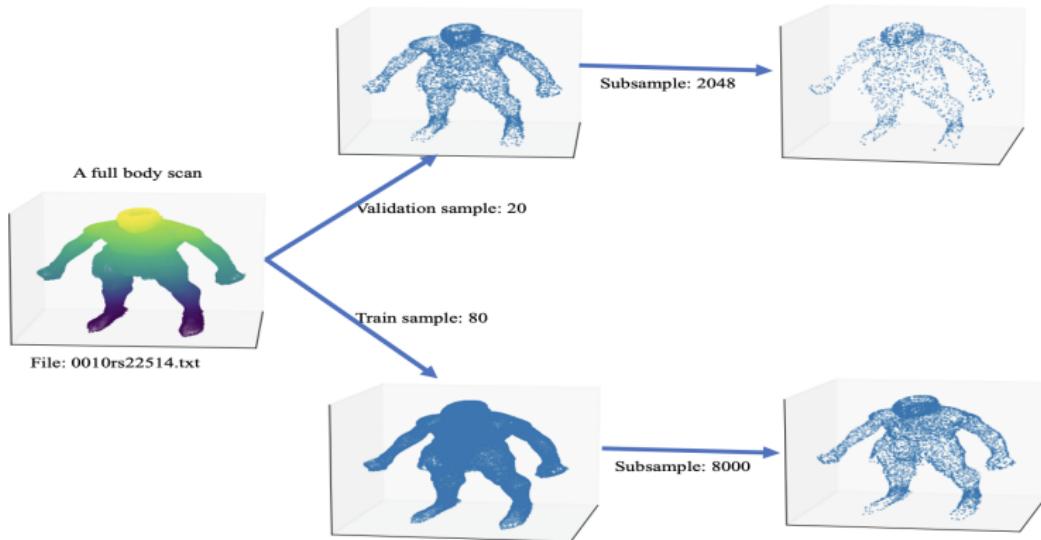
# Data

- Point cloud 3D Body scan
- Collected by Dr. Frederick Steven Cottle
- KX-16 Body Scanner [1]
- Dataset includes 75 Participants' unlabeled (unavailable metadata) data files, with point cloud count ranging between 51000 to 63000.
- As a preliminary step, this unlabelled dataset is used for the classification, a self-identifying task.



# Data

- Figure 1 shows the initial body scan, train to validation split of point clouds, and subsampling during each batch iteration.



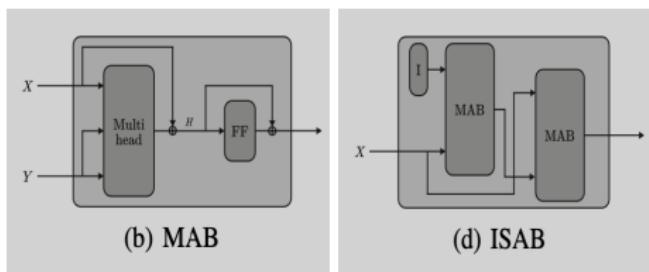
**Figure 1:** Train dataset 3D Body Scan



Computational Science

# Background

- Set Transformer - Lee, J. et. al. (2018) [2]
  - Induced Set Attention Block(ISAB) - speeds up computation ( $\mathcal{O}(nm)$ ) and extracts meaningful features.
  - Pooling layer - parameterized aggregation function, helps to capture the varying contribution of the instances, for better aggregation. Pooling by attention operation makes the model permutation invariant.



**Figure 2:** Multi Head Attention Block(MAB)[2]

**Figure 3:** Induced Set Attention Block(ISAB)[2]

# Background

- Simple framework for Contrastive Learning [3] - Chen, T. et. al. (2020): ResNet-50 on ImageNet dataset, Positive pair/ Negative pair. Using SimCLR self-supervised method shows similar performance to Resnet-50 supervised learning.
- Generalization
  - Kullback-Leibler divergence(KL divergence) [4]

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \quad (1)$$



Computational Science

# Methods

## General Experimental Setup

- Data Loader
- Batch size - train:4, validation:4
- Sub sample size - train:8000, validation:2048
- Train : Validation = 80:20
- Monte Carlo Simulation - 10
- Optimizer - Adam
- Loss - Sparse categorical cross entropy



# Experimental Setup

## Hyperparameters

### Standard Set Transformer

Embedded dimensions - 64, 32  
Number of heads - 32, 16  
Induce points - 128, 64  
stack - 3  
Dropout - 0.05  
Learning rate - 1e-03  
Number of epochs - 250

### Contrastive Pretraining

Experimented batch size - 32, 16  
Point cloud - 1024, 2048  
Temperature - 0.5  
Number of epochs - 200

### Fine Tuning

Linear layer - 128, 1024  
Non-linear activation function - LeakyReLU  
Dropout - 0.1  
The rest of the hyperparameters are  
maintained as same as Standard Set Transformer



Computational Science

# Standard Set Transformer Algorithm

---

## Set Transformer architecture

---

- 1:  $y = \text{Linear}(3, \text{embed\_dim})$
- 2: **for**  $j = 1$  to  $\text{stack}$  **do**
- 3:    $y = \text{InducedSetAttentionBlock}(y)$
- 4: **end for**
- 5:  $y = \text{Dropout}(0.05)(y)$
- 6:  $y = \text{PoolingByMultiHeadAttention}(y)$
- 7:  $y_{\text{embedding}} = \text{Dropout}(0.05)(y)$
- 8:  $y = \text{FinalDense}(\text{numberofclasses})(y_{\text{embedding}})$
- 9: *return*  $y, y_{\text{embedding}}$

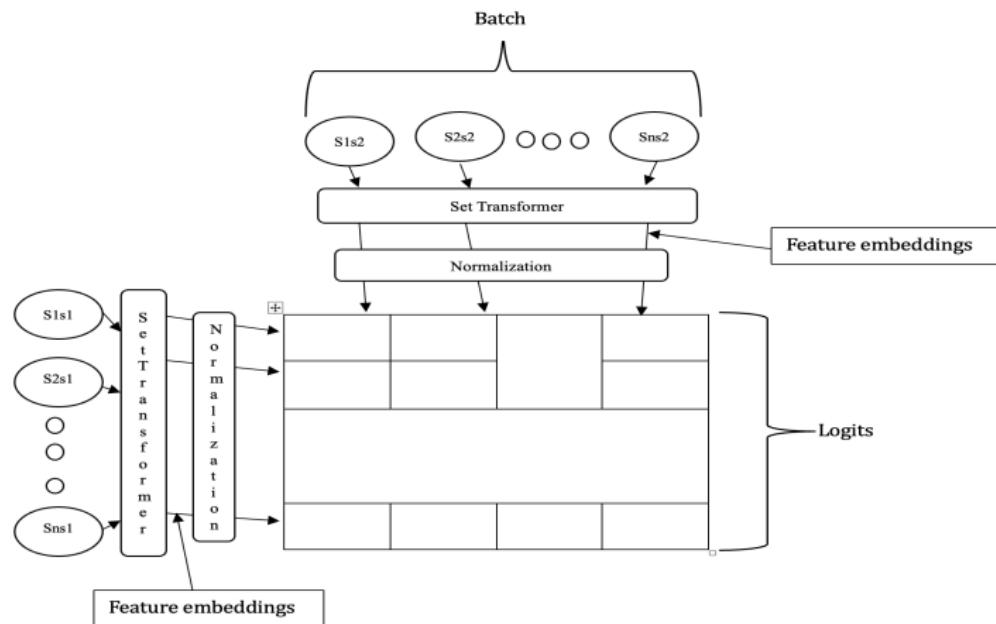
---



Computational Science

# Contrastive Learning using Set Transformer

Visual representation of contrastive learning



**Figure 4:** Contrastive Learning using Set Transformer architecture



Computational Science

# Contrastive Learning using Set Transformer

## Algorithm

---

### Contrastive Learning architecture

---

```
1:  $y_1, y\_embedding1 = \text{SetTransformer}(\text{batch1})$ 
2:  $y_2, y\_embedding2 = \text{SetTransformer}(\text{batch2})$ 
3:  $y\_embedding1 = \text{Linear}(\text{embed\_dim}, \text{projection\_dim})(y\_embedding1)$ 
4:  $y\_embedding2 = \text{Linear}(\text{embed\_dim}, \text{projection\_dim})(y\_embedding2)$ 
5:  $y\_embedding1 = \text{Norm}(y\_embedding1)$ 
6:  $y\_embedding2 = \text{Norm}(y\_embedding2)$ 
7:  $y = \text{Mul}(y\_embedding1, y\_embedding1.T) * \text{Temperature}$ 
```

---



Computational Finance

# Generalization

- Weak Generalization - Has knowledge about all the category data at the time of training.
- Strong Generalization(leave one out) - Is not aware of a particular class data at a given training period.
- Probability matrix -  $75 \times 75$
- Reassign all diagonal elements in weak generalization 0.
- Normalize both of the matrices
- Apply KL Divergence



# Visualization

- Get feature embeddings for both standard Set transformer and Contrastive pre-trained model.
- 10 vectors for each sample.
- Applied PCA with 2 principal components to visualize in 2D.
- Applied t-SNE with a 2D component with the perplexity of 20.
- Evaluate along the 2D axis if, distribution is relatable to human distinguish.

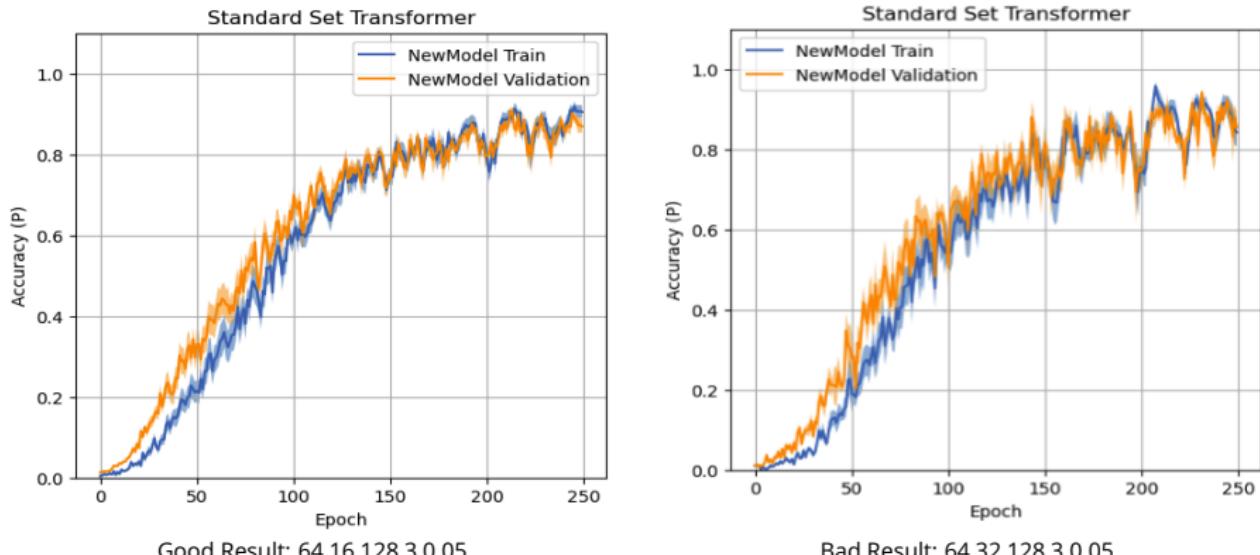


Computational Science

# Results

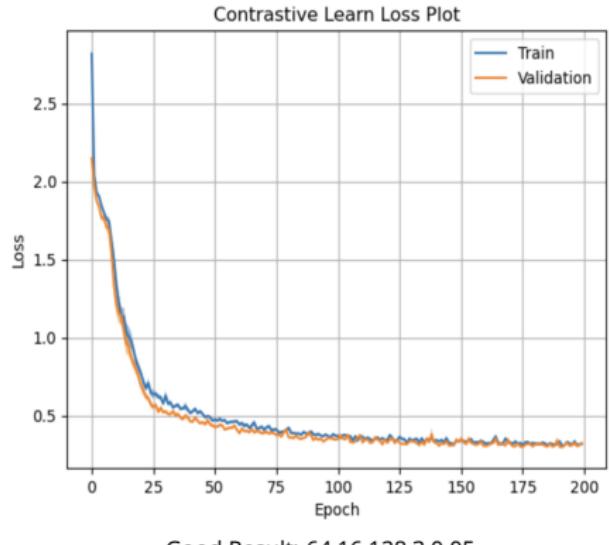


# Standard Set Transformer

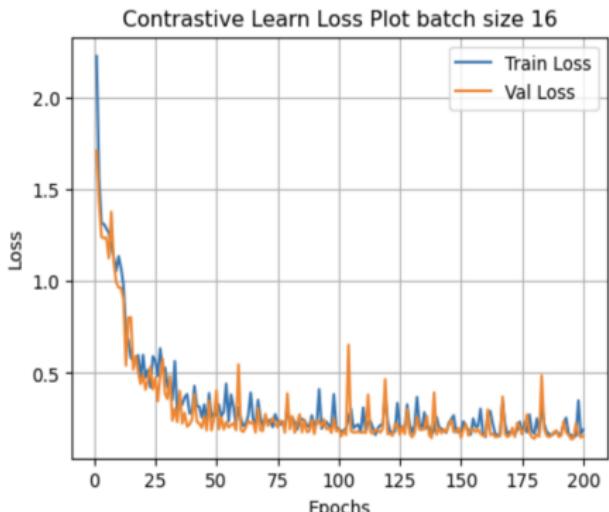


**Figure 5:** Monte Carlo simulated Standard Set Transformer Accuracy

# Contrastive Learning using Set Transformer



**Figure 6:** Monte Carlo simulated Contrastive Learning using Set Transformer Loss

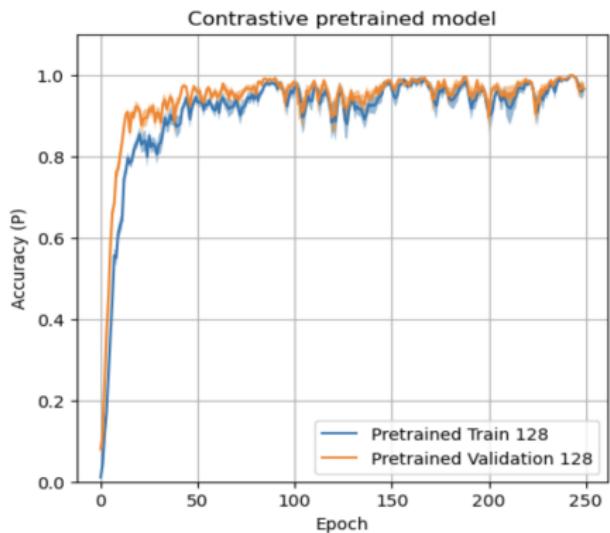


Bad Result: 64,16,128,3,0.05,batch size: 16

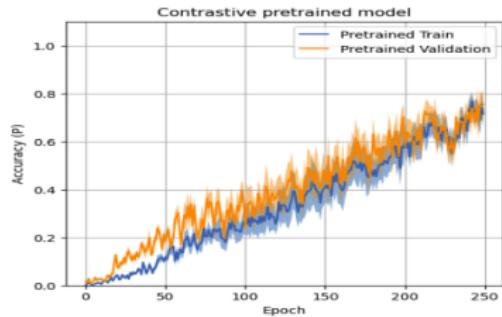


Computational Science

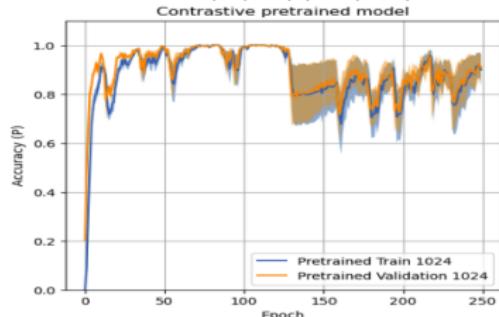
# Fine Tuning



Good Result: 64,16,128,3,0.05, 128, 0.1



Bad Result: 64,32,128,3,0.05, 128, 0.1



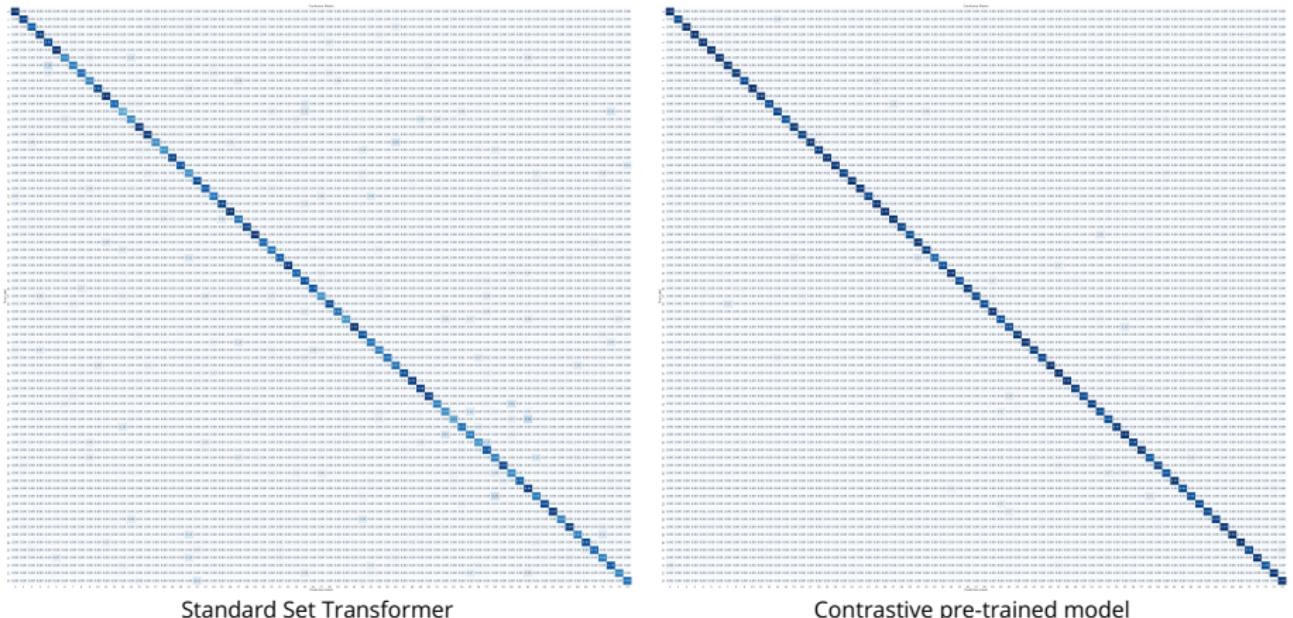
Bad Result: 64,16,128,3,0.05, 1024, 0.1

**Figure 7:** Monte Carlo simulated Contrastive pre-trained model Accuracy



Computational Science

# Weak Generalization



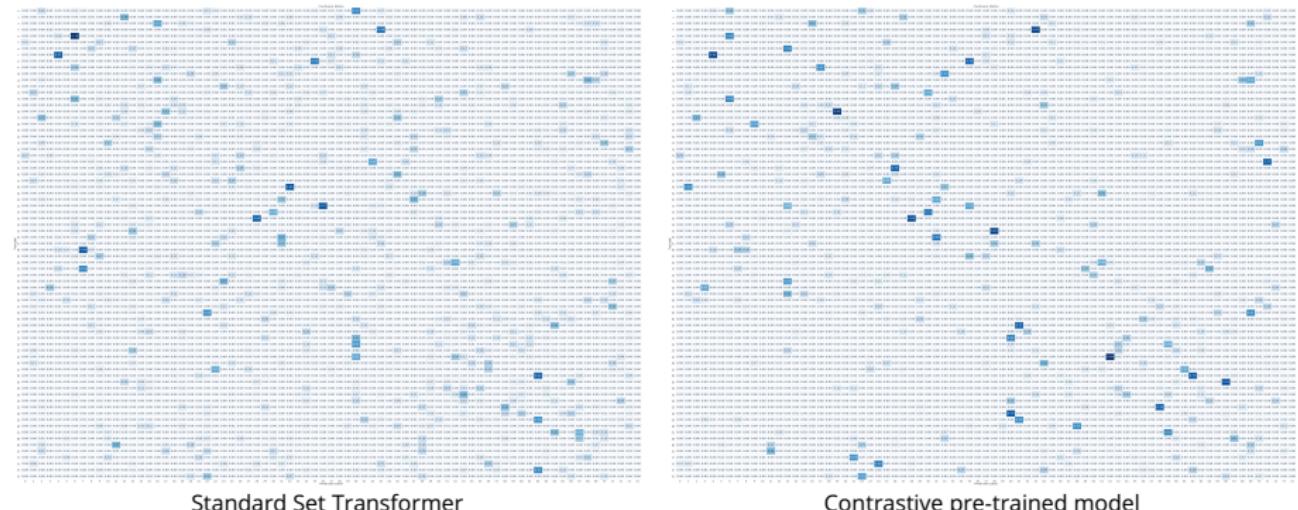
**Figure 8:** Weak Generalization averaged across Monte Carlo simulation

- The higher the variation in blue shades the less confident the model was each time of simulation.
- Uniform diagonal signifies less confusion in each run in self-identifying.



Computational Science

# Strong Generalization



**Figure 9:** Strong Generalization averaged across Monte Carlo simulation

**Table 1:** Results for Generalization using KL Divergence

Method	KL Divergence
Standard Set Transformer	6.60
Contrastive Pre-trained model	6.25



Computational Science

# Visualization

## PCA Visualization for Standard Set Transformer

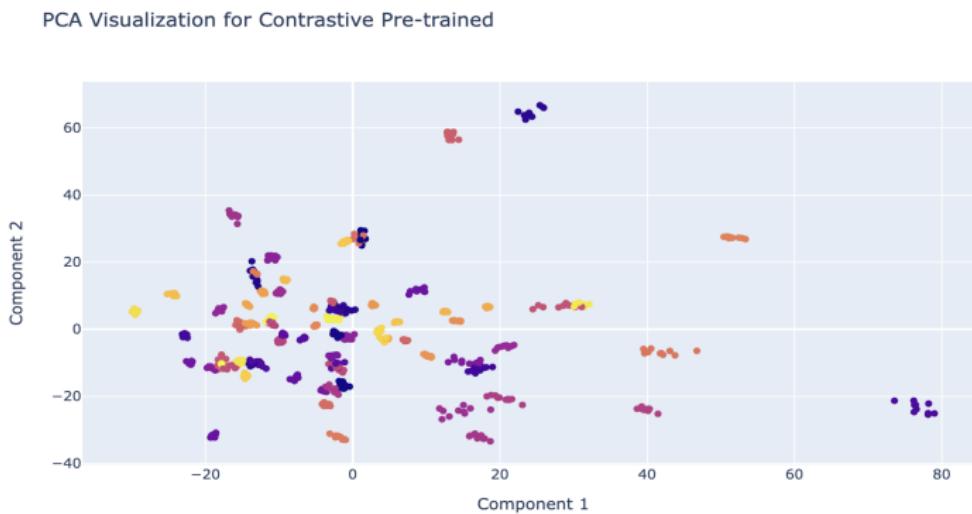


**Figure 10:** PCA Visualization for Standard Set Transformer

- Principal Component 1: 30.26%
- Principal Component 2: 24.38%
- Total Principal Component: 54.64%

# Visualization

## PCA Visualization for Contrastive Pre-trained Model

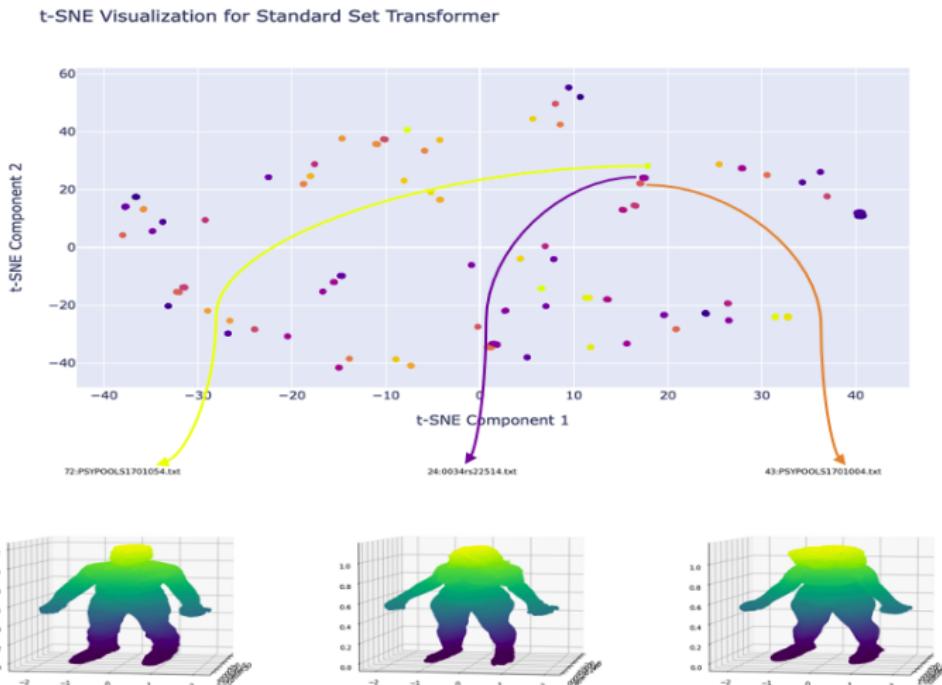


**Figure 11:** PCA Visualization for Contrastive Pre-trained

- Principal Component 1: 20.50%
- Principal Component 2: 18.35%
- Total Principal Component: 38.85%

# Visualization

## t-SNE Visualization for Standard Set Transformer



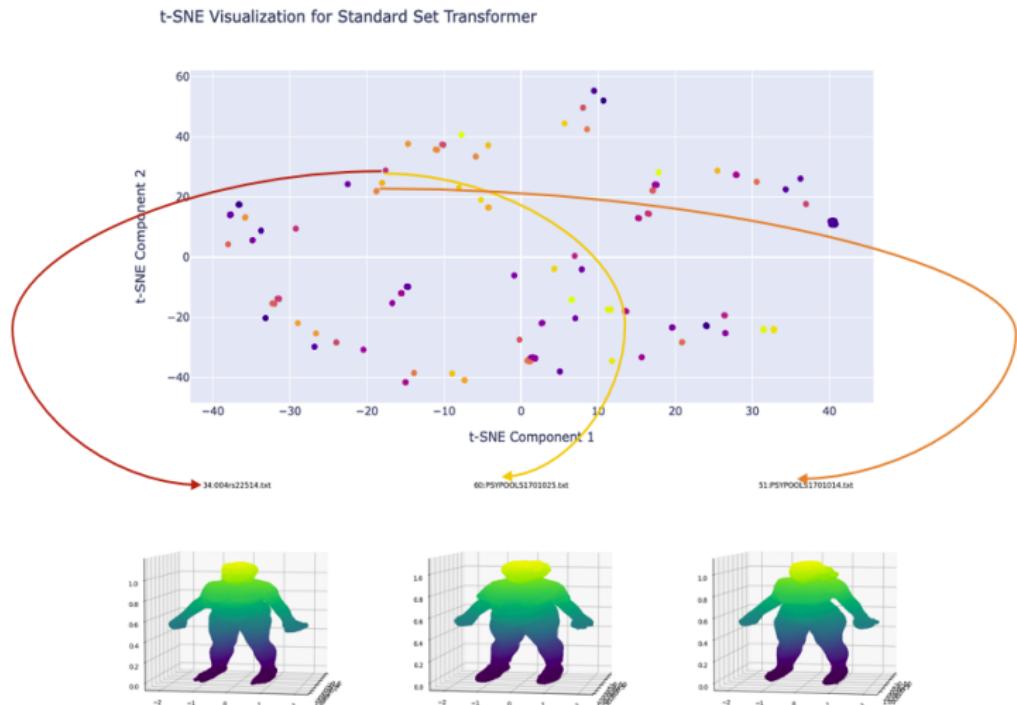
**Figure 12:** t-SNE Visualization for Standard Set Transformer



Computational Science

# Visualization

## t-SNE Visualization for Standard Set Transformer



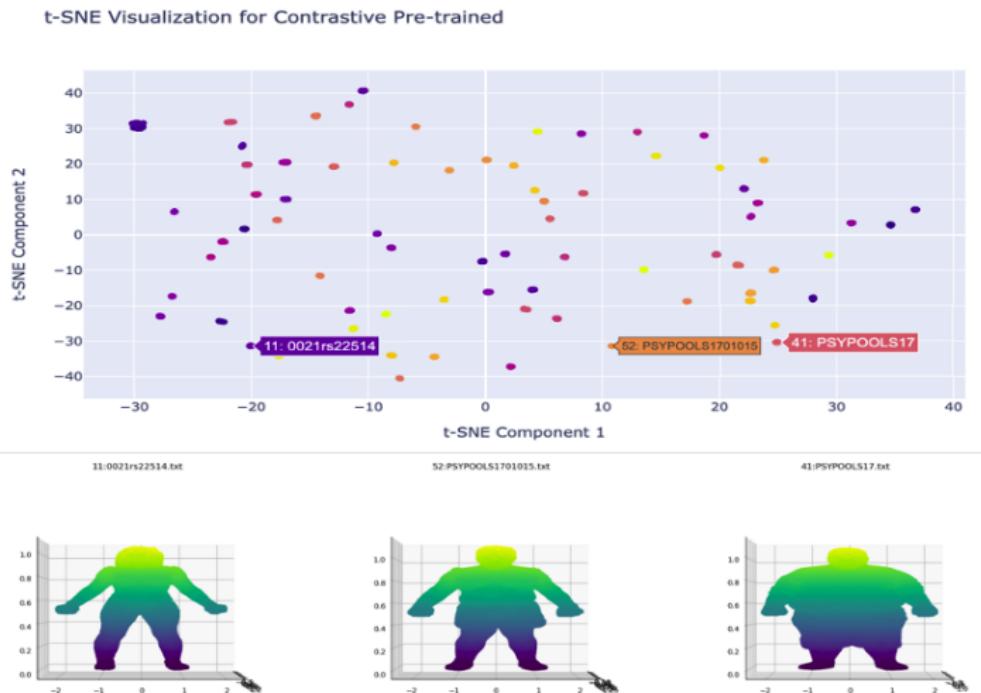
**Figure 13:** t-SNE Visualization for Standard Set Transformer



Computational Science

# Visualization

## t-SNE Visualization for Contrastive Pre-trained Model



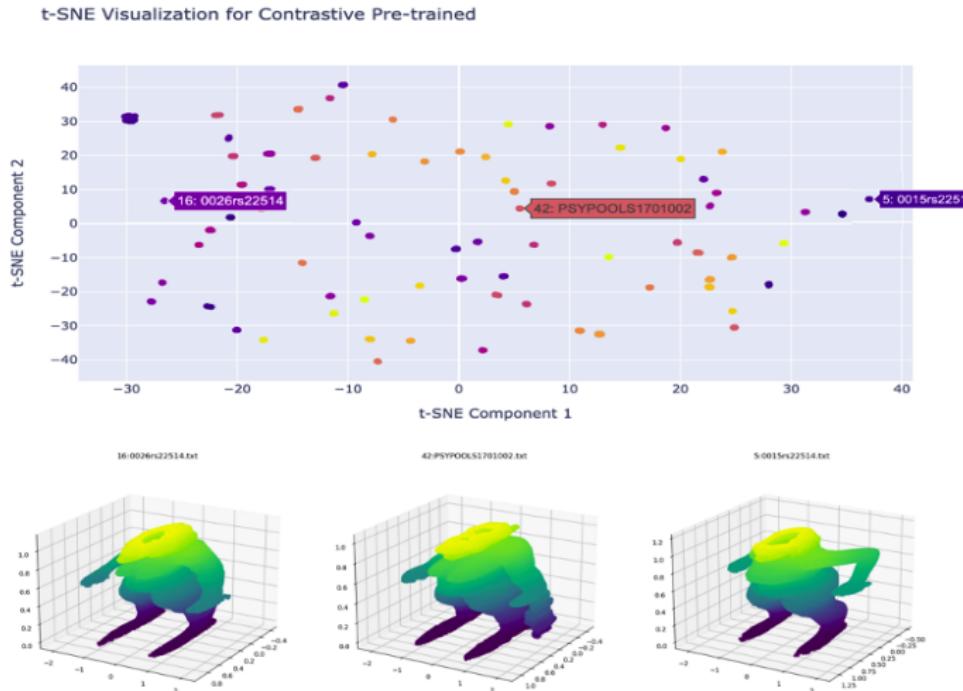
**Figure 14:** t-SNE Visualization for Contrastive Pre-trained Model



Computational Science

# Visualization

## t-SNE Visualization for Contrastive Pre-trained Model



**Figure 15:** t-SNE Visualization for Contrastive Pre-trained Model



Computational Science

# Conclusion

- Standard Set transformer and Contrastive Pre-trained model for self-identification task.
- Weak and strong generalization for efficiency of the model.
- Contrastive Pre-trained model excelled Standard Set Transformer performance in terms of accuracy, speed, and stability.
- Contrastive pre-trained model showed improved generalization than the standard Set Transformer.



# Conclusion

## Future work

- Contrastive Pre-trained model can be used as a solid foundation for extended work like Generative Adversarial Set Transformer(GAST).
- GAST approach is to generate point cloud data to form a full or a partial 3D body scan.
- This approach is in the notion to provide support in case of missing scanned parts or lesser point cloud data that helps with increased precision.



# Reference

- [TC]2 Introduces KX-16 Body Scanner. Textile World. (2012, March 20). <https://www.textileworld.com/textile-world/new-products/2012/03/tc2-introduces-kx-16-body-scanner-3/>
- Lee, J., Lee, Y., Kim, J., Kosiorek, A. R., Choi, S., & Teh, Y. W. (2018). Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. ArXiv. /abs/1810.00825
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. ArXiv. /abs/2002.05709
- Wikimedia Foundation. (2023, July 16). Kullback–Leibler divergence. Wikipedia. [https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler\\_diverge](https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence)

# Thank you