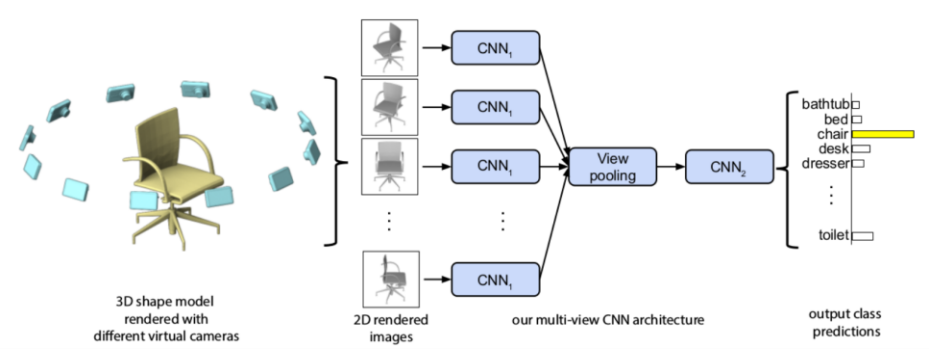
Exploit Clues from Views: Self-Supervised and Regularized Learning for Multiview Object Recognition

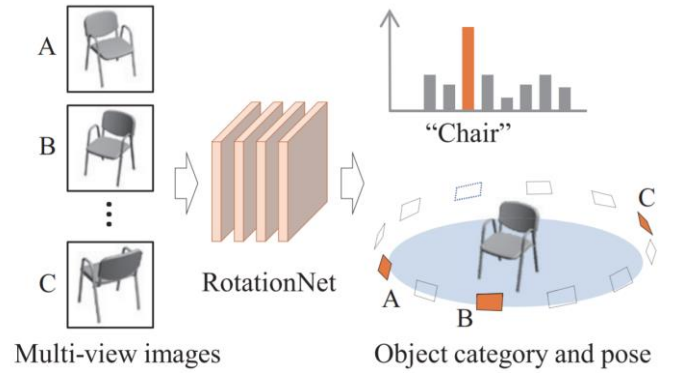
Chih-Hui Ho, Bo Liu, Tz-Ying Wu, Nuno Vasconcelos
University of California, San Diego

Introduction

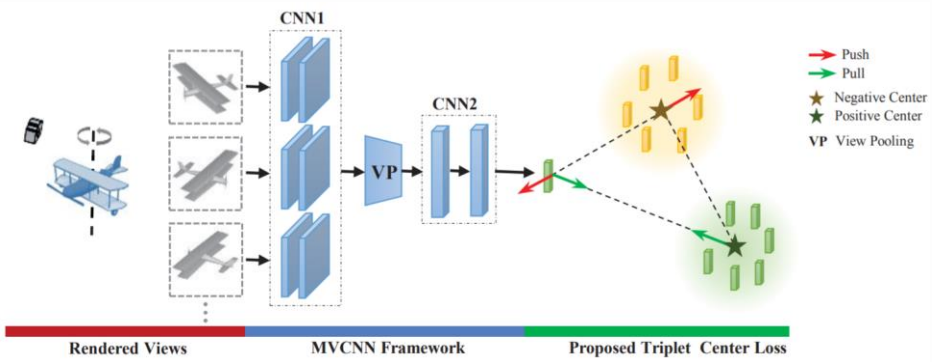
- Multiview recognition has been well studied in the literature and achieves decent performance in object recognition and retrieval task. For example,



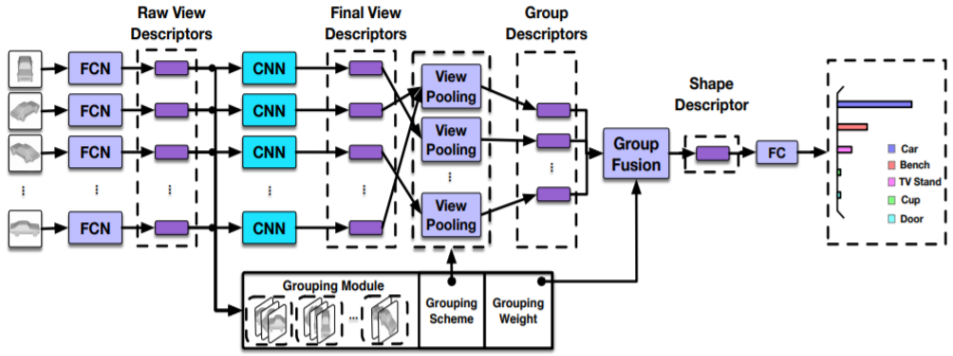
MVCNN



RotationNet



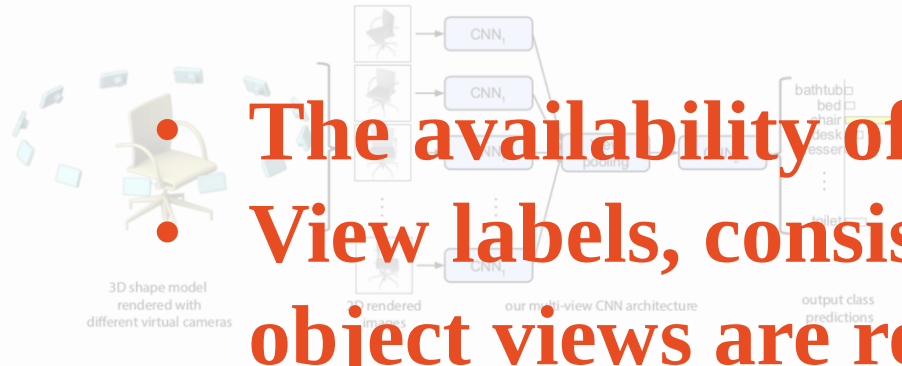
Triplet center loss



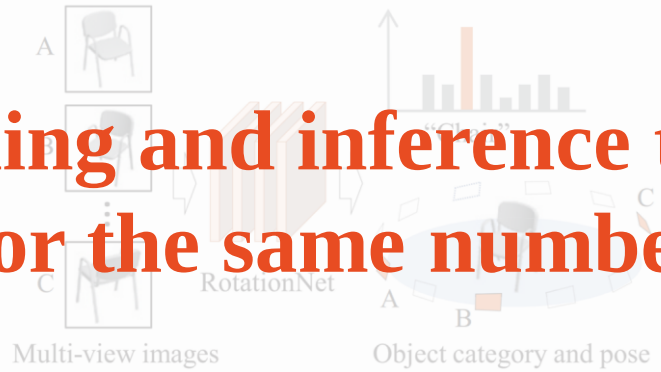
GVCNN

Introduction

- However, most previous works rely on some impractical assumptions.



MVCNN



RotationNet

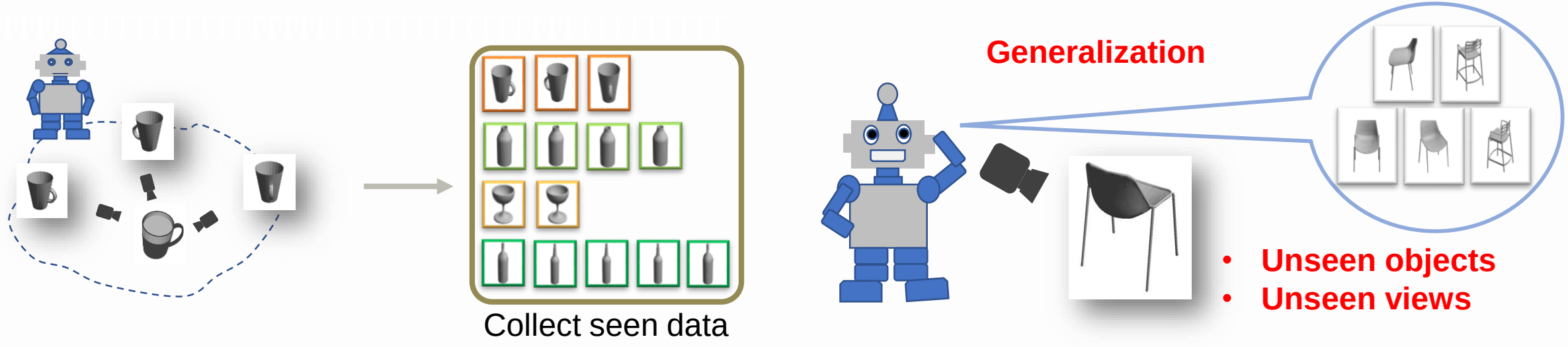
- **The availability of all views in training and inference time.**
- **View labels, consistent view angles or the same number of object views are required.**

Triplet center loss

GVCNN

Introduction

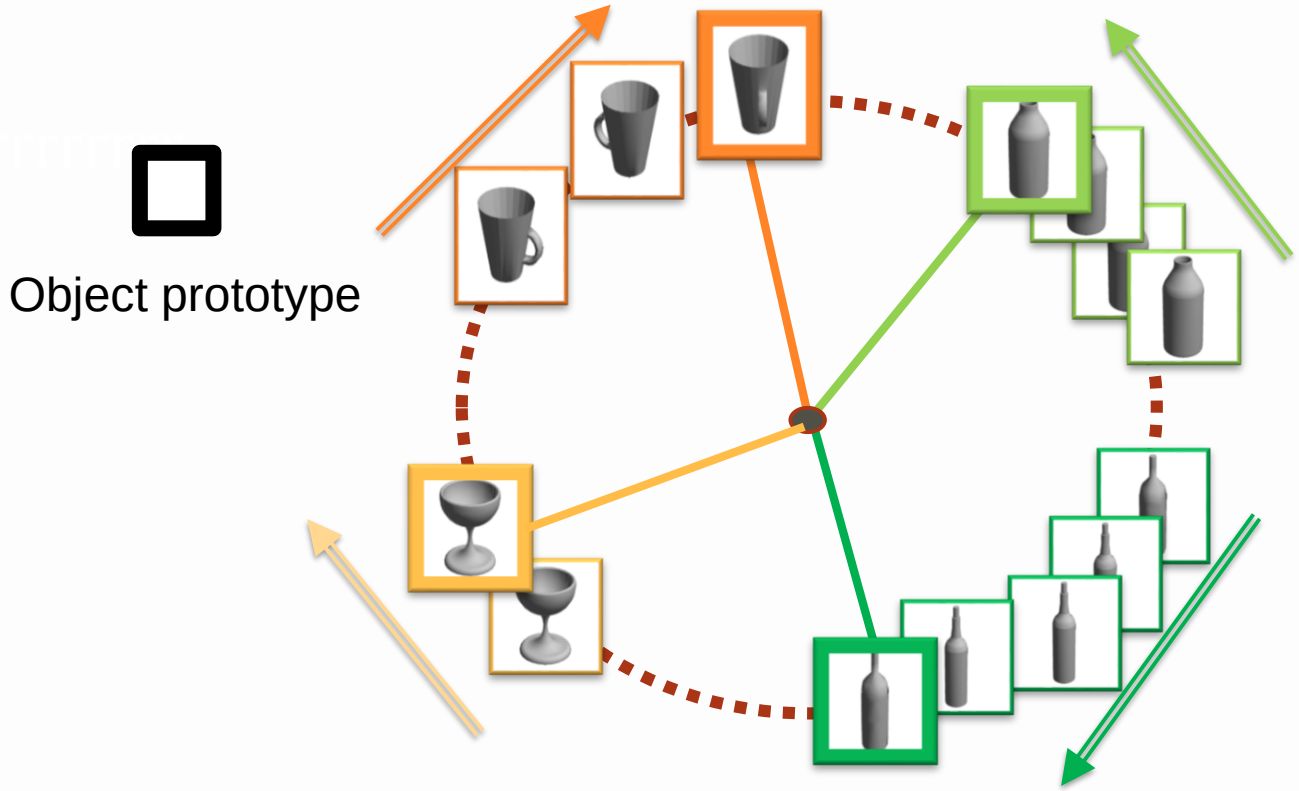
- These limitations prevent many applications of interest.
- For example, a household robot of limited memory is tasked with picking scattered objects and returning them to their locations.
- The model needs to generalize to unseen objects and views during inference.



Pretrained stage **Inference stage**

Methods

- A baseline is to train a softmax classifier on seen objects and encourage views of an object clustering around its prototype.
- However, the softmax classifier is successful only on seen objects and unable to generalize to unseen objects.



$$P_{Y|X}(i|x) = \frac{\exp(W_i^T f_\theta(x))}{\sum_{k=1}^{N} \exp(W_k^T f_\theta(x))},$$

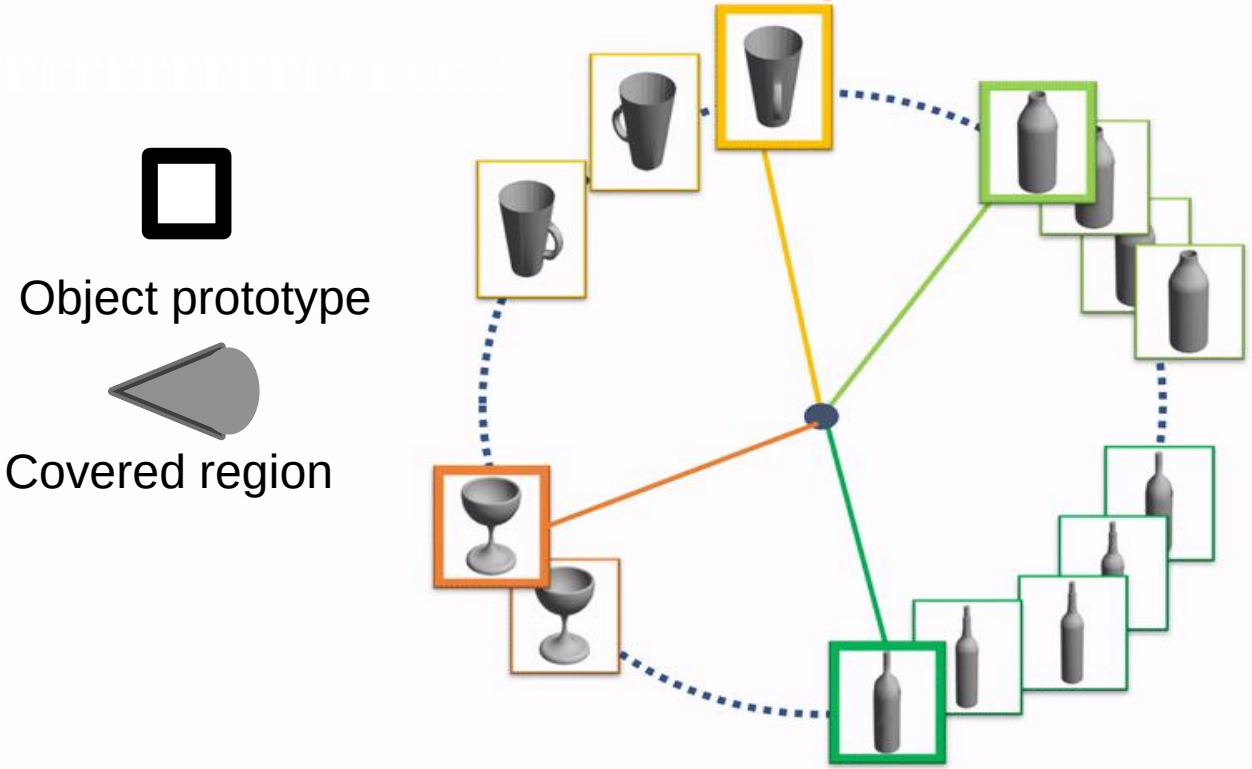
w_i is the parameter vector of object i

Methods

- We propose to replace the classifier weight with the normalized embedding of a randomly chosen object view for better generalization.
- This is referred as **multiview stochastic prototype embedding (MVSPE)**.

$$P^{S}_{Y|X}(i|x) = \frac{exp(f_\theta(x_i^{v_i})^T f_\theta(x)/\tau)}{\sum_{k=1}^{N} exp(f_\theta(x_k^{v_k})^T f_\theta(x)/\tau)},$$

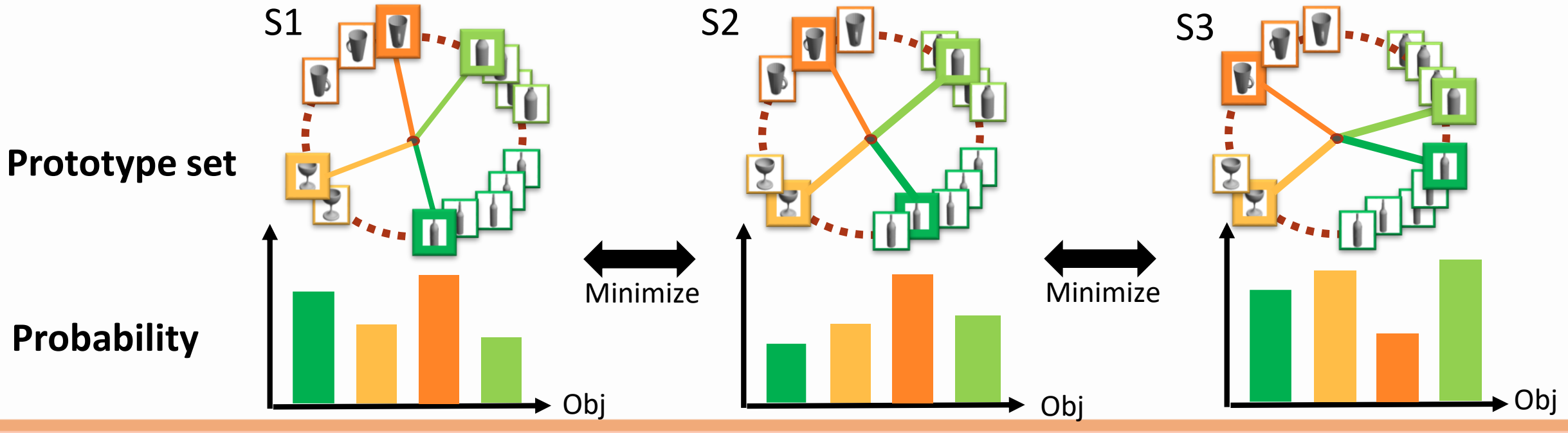
where $f_\theta(x_i^{v_i})$ is the normalized embedding of a randomly chosen view of object i, τ is the temperature and v_i is the view selector.



Methods

- The probability shift of choosing different view (prototype) set is minimized with KL divergence and leads to our final **view invariant stochastic prototype embedding (VISPE)**.

$$L_{KL} = \sum_{k=1}^{m} P_{Y|X}^{s_1}(i|x) log \frac{P_{Y|X}^{s_1}(i|x)}{P_{Y|X}^{s_2}(i|x)}$$

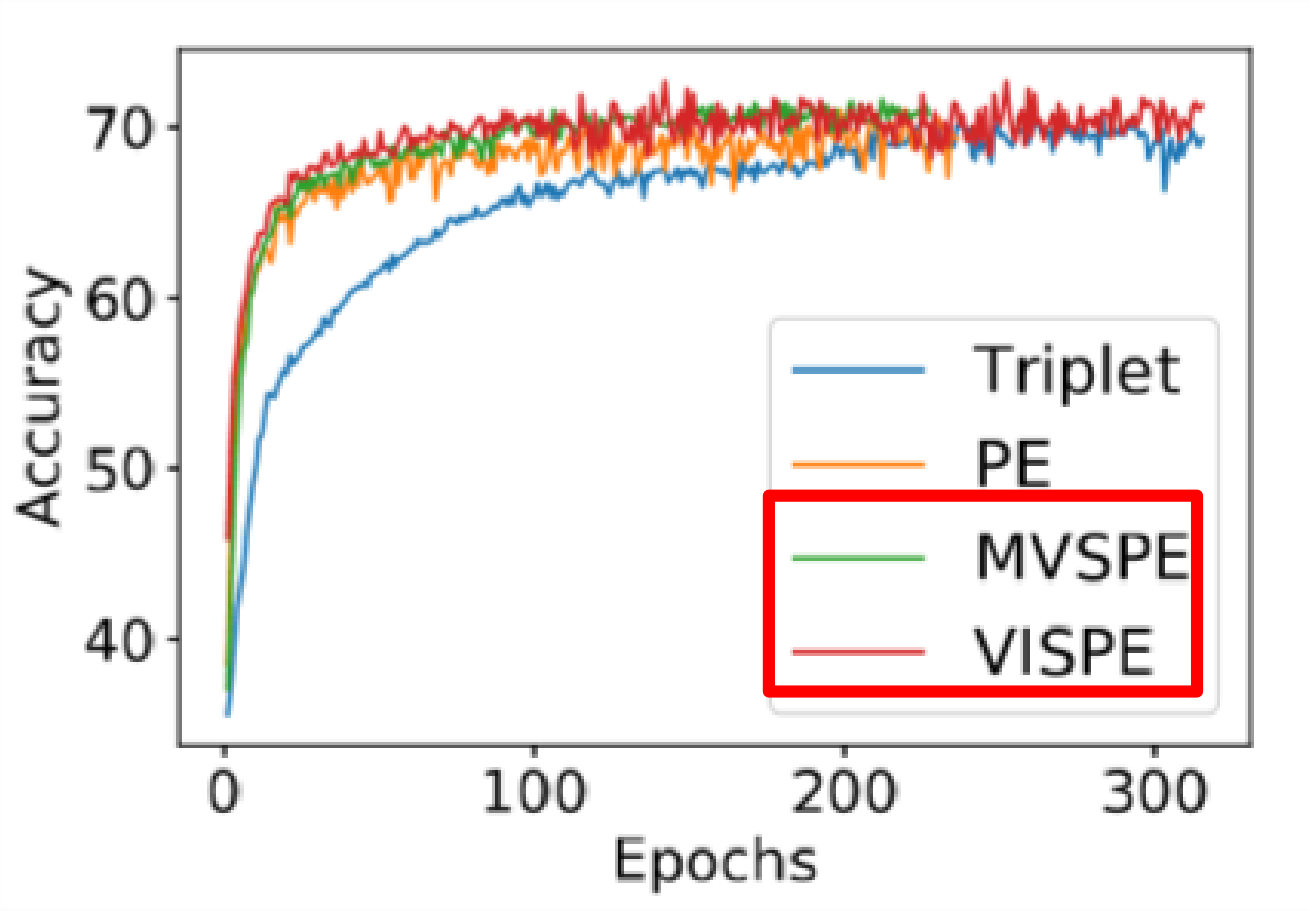


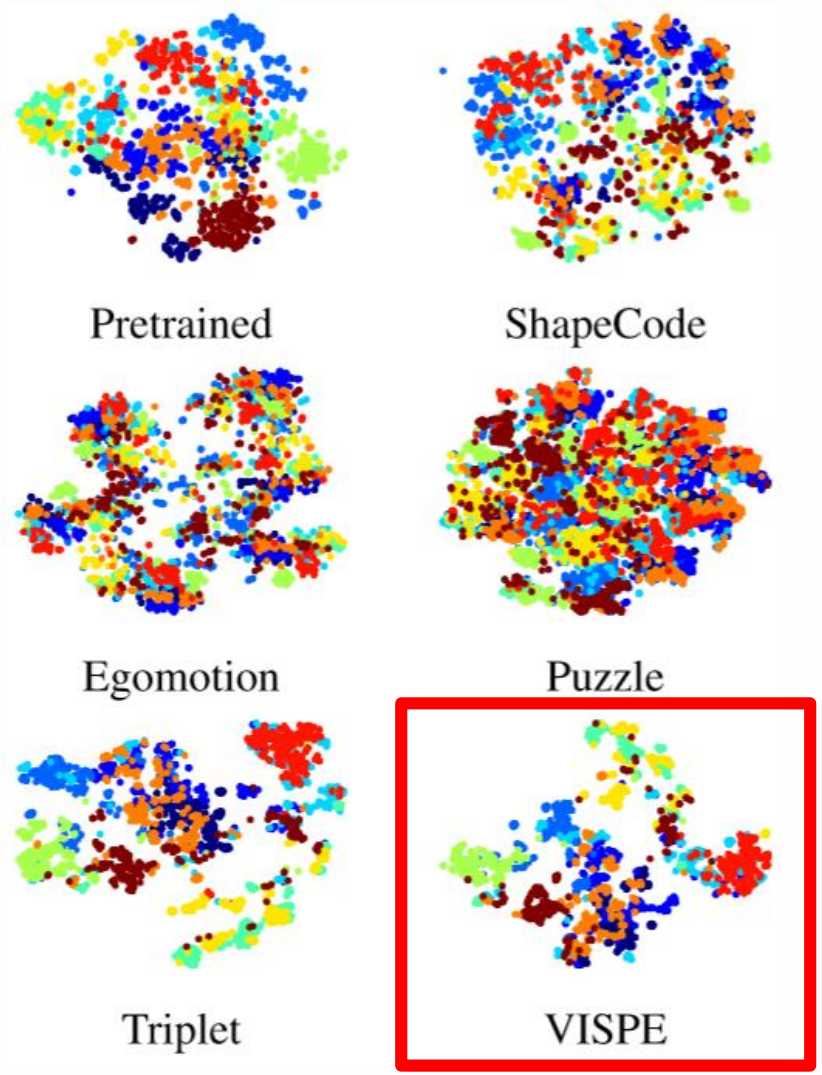
Experiment

Datasets	Surrogate	ModelNet		ShapeNet		ModelNet-S	
Methods / Classes	Task	seen	unseen	seen	unseen	seen	unseen
Chance	N/A	3.3	10.0	3.3	4.0	3.3	10.0
Pretrained [1]	N/A	62.7	52.7	63.9	58.1	58.2	55.2
Autoencoder [2]	Context	31.8	37.2	29.8	26.3	34.7	38.8
Egomotion [3]	Motion	32.4	34.7	72.6	47.1	33.0	35.2
Puzzle [4]	Sequence	34.4	41.5	67.8	48.6	34.8	42.4
UEL [5]	Data Aug.	47.9	46.5	68.7	53.4	46.4	48.2
ShapeCode [6]	View	39.4	46.5	67.1	42.3	38.8	47.2
MVCNN [7]	N/A	39.6	48.1	30.3	32.4	36.7	44.8
Triplet [8]	N/A	70.1	62.4	81.2	61.2	64.7	62.1
Instance classifier	N/A	57.7	58.9	69.3	60.4	52.3	54.6
PE	Object	69.7	61.7	81.6	63.8	62.1	60.4
MVSPE	Object	70.3	63.2	82.4	64.6	64.6	62.1
VISPE	Object	**71.2**	**64.4**	**82.9**	**65.5**	**66.2**	**64.3**

Outperform baselines for object recognition on seen and unseen classes

Experiment





Faster convergence

Better embedding structure

Thank you for listening

Code available at
https://github.com/chihhuiho/VISPE