View Inter-Prediction GAN: Unsupervised Representation Learning for 3D Shapes by Learning Global Shape Memories to Support Local View Predictions




Zhizhong Han[1,2] **Mingyang Shang**[1] **Yu-Shen Liu**[1] **Matthias Zwicker**[2]

[1]School of Software, Tsinghua University, Beijing, China

[2]Department of Computer Science, University of Maryland, College Park, USA

Content

- Background
- Motivation
- Current solutions
- The key idea of VIP-GAN
- Problem statement
- Technical details
- Results
- Contributions

Background

- Feature learning for 3D shapes is crucial for 3D shape analysis:
  - Classification
  - Retrieval
  - Segmentation
- Supervised 3D feature learning has produced remarkable results:
  - Under large scale 3D benchmarks
  - Train deep neural networks
  - With supervised information, such as class labels and point correspondences

Motivation

- However, obtaining supervised information requires intense manual labeling effort.
- Therefore, unsupervised 3D feature learning with deep neural networks is an important research challenge.

Current solutions

- Several studies have addressed this challenge.
  - Train deep learning models using “supervised” information mined from the unsupervised scenario.
- Different strategies for the prediction of a shape :
  - From itself by minimizing reconstruction error or embedded energy.
  - From its context given by views or local shape features.
  - From views and itself together.
- Use *all views* to provide a holistic context of 3D shapes.

The key idea of VIP-GAN

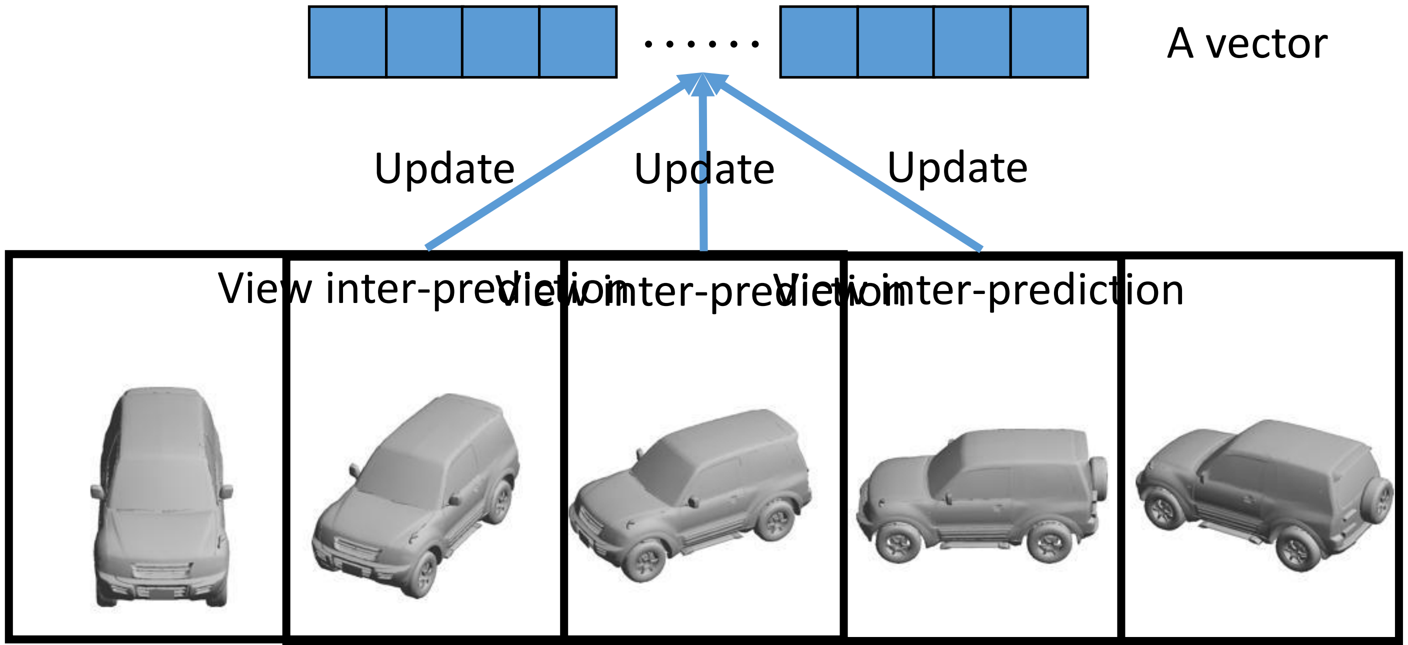
- In contrast, our approach called *View Inter-Prediction GAN (VIP-GAN)* learns to
  - make *multiple local view inter-predictions* among neighboring views.
- The view inter-prediction task mimics human perception of view-dependent patterns:
  - Based on changes between neighbor views, easily imagine the center view.
  - Reversely, based on the center, easily imagine the neighbor views.

The key idea of VIP-GAN

- As a key idea, VIP-GAN implements the 3D shape representation as a *shape-specific global memory*.
- Its contents are learned to support all local view inter-prediction tasks for each shape.

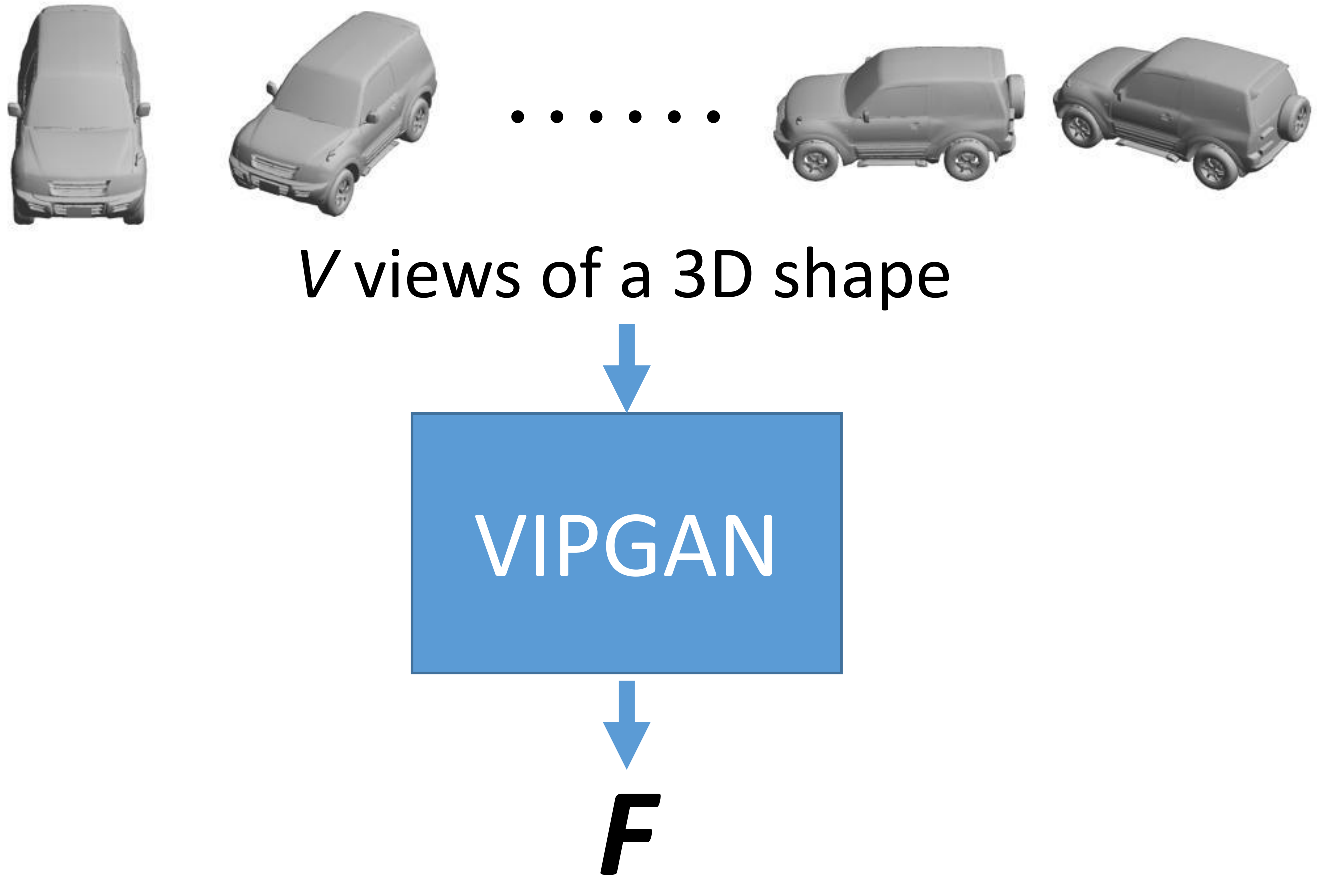
The global feature of a 3D shape:

The sequential views of the 3D shape:



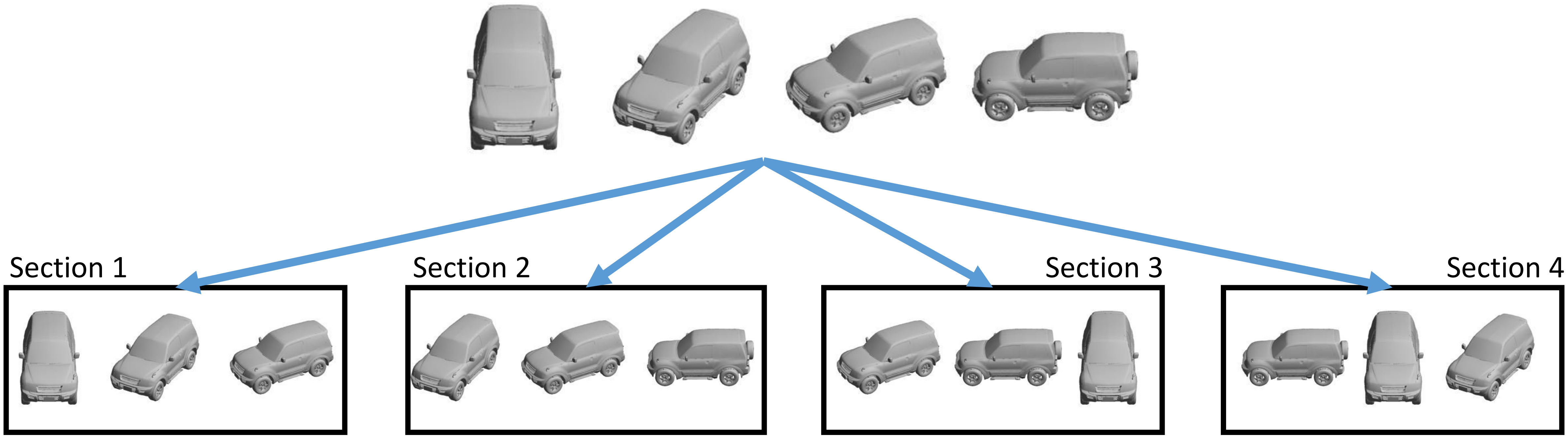
Problem statement

- Learn a global feature ***F*** of a 3D shape m from its *V* views *without any supervised information*.



Technical details

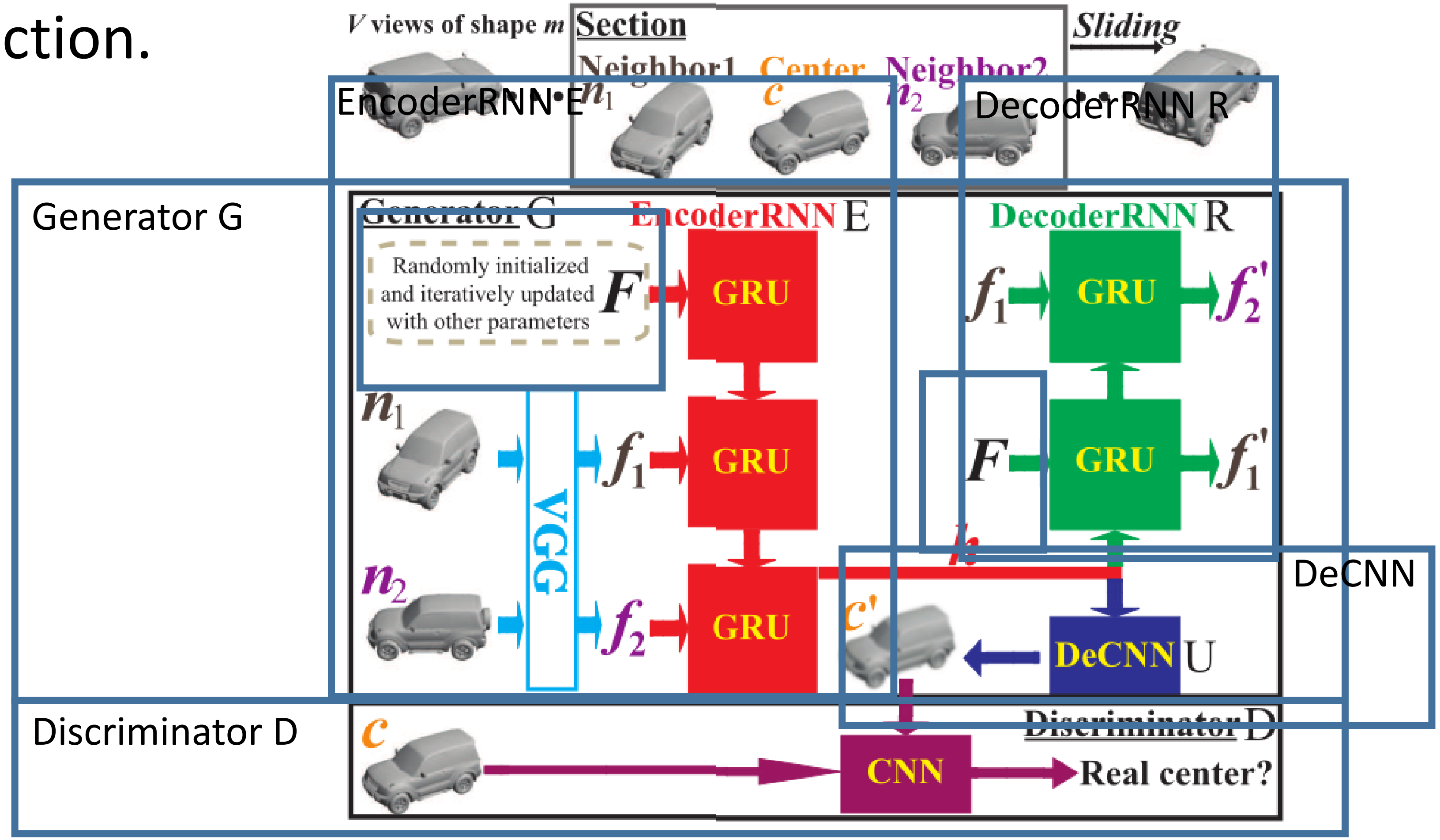
- First, we split the set of *V* views into *V* sections of equal length.



Each one of the *V* views is the center in each section.

Technical details

- VIP-GAN learns ***F*** for the 3D shape by inter-view prediction in each section.



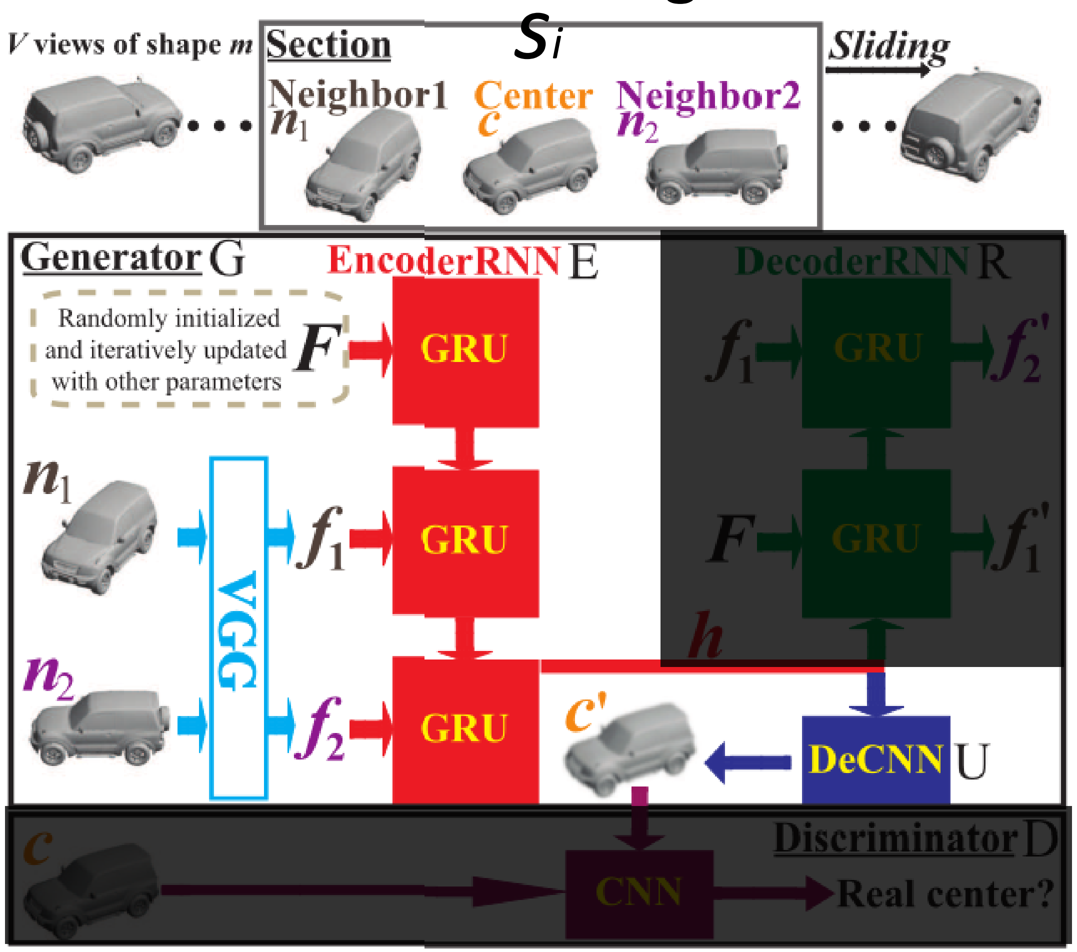
Technical details

- For each section S_i, the prediction of the center c from its neighbors

Center view prediction loss:

$$L_{\mathrm{U}} = \|\mathrm{U}(\boldsymbol{s}_i) - c\|_2^2$$

This center view prediction is conducted in the image space.



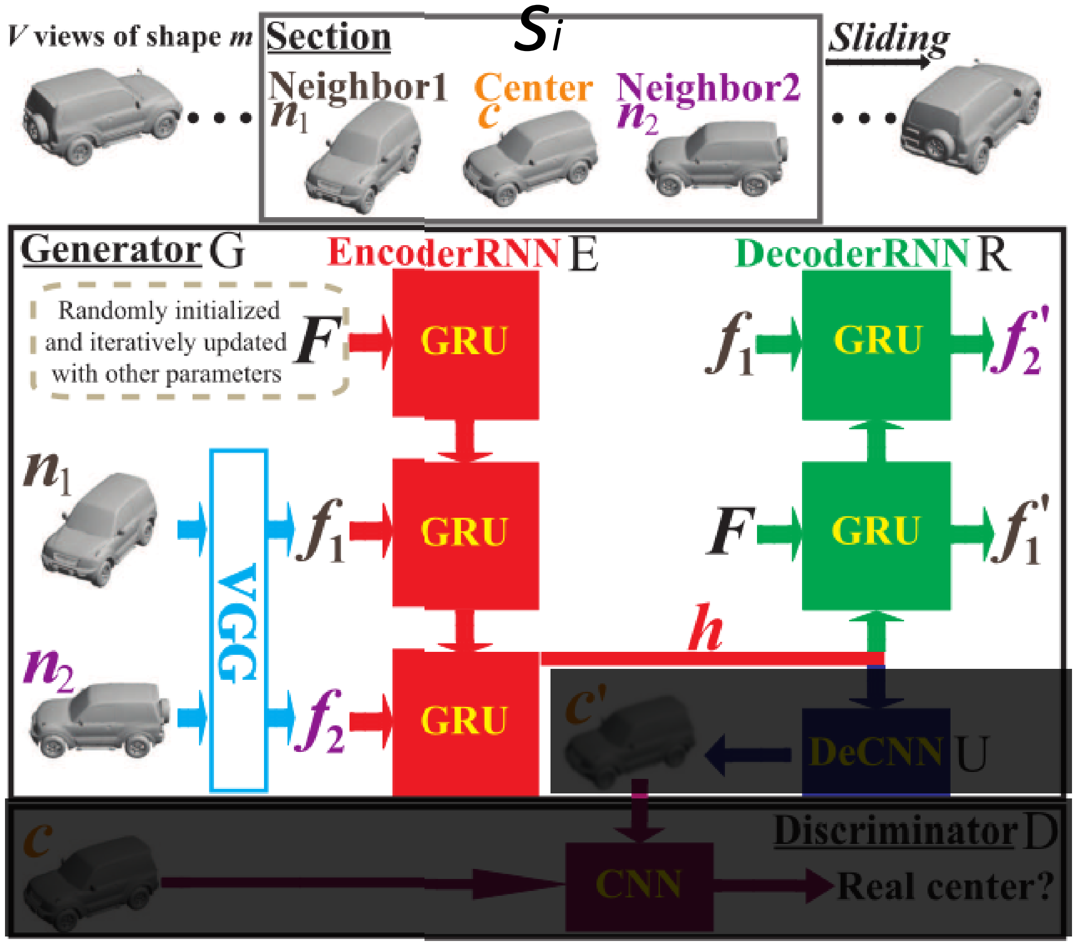
Technical details

- For each section s_i, the prediction of the neighbors from its center.

Neighbor views prediction loss:

$$L_{\mathrm{R}} = \frac{1}{N}\sum_{j=1}^{N} \|\mathrm{R}(\boldsymbol{s}_i)_j - \boldsymbol{f}_j\|_2^2$$

To enable VIP-GAN to more fully understand the 3D shape, the neighbor view prediction is conducted in the feature space, which is different from the center view prediction.

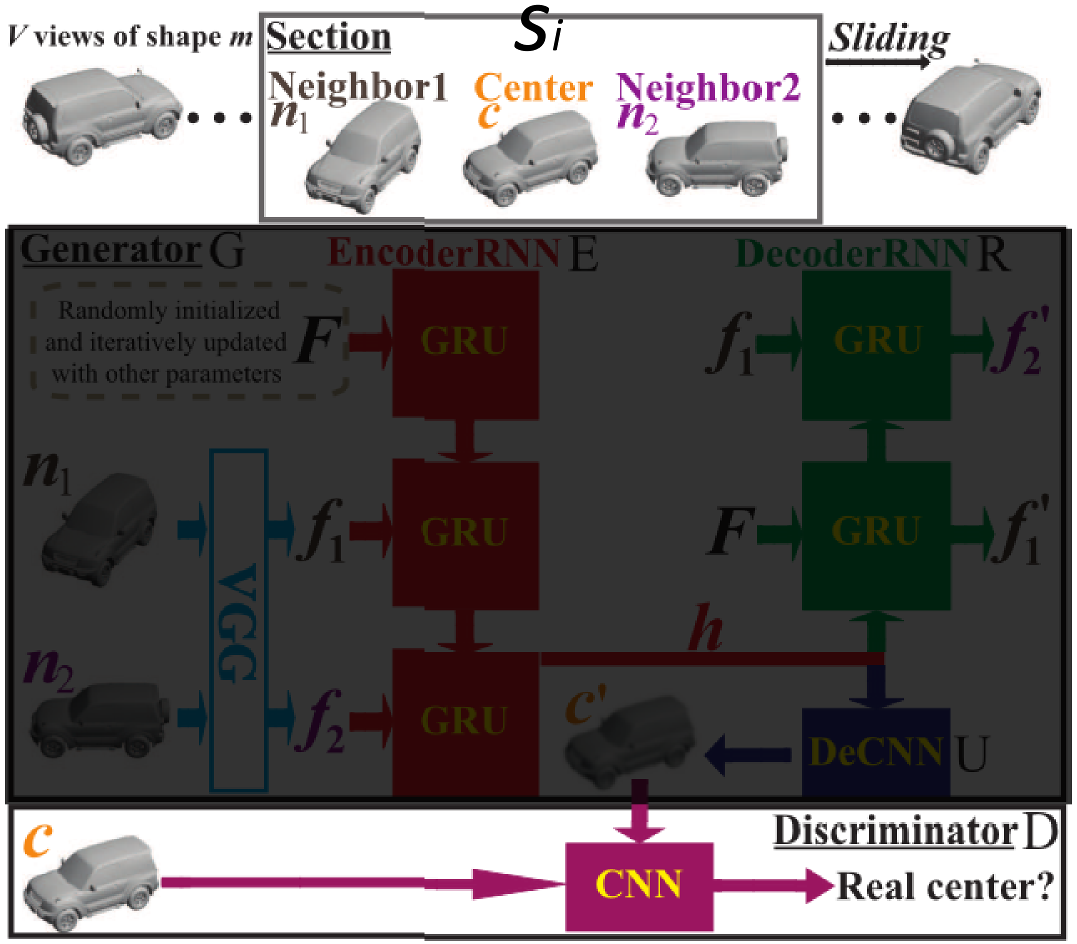


Technical details

- For each section S_i, making the predicted center view more real.

Adversarial loss on predicted center view:

$$L_{\mathrm{D2U}} = \log(1 - \mathrm{D}(\mathrm{U}(\boldsymbol{s}_i)))$$



Technical details

- For each section S_i ,

The loss of VIP-GAN:

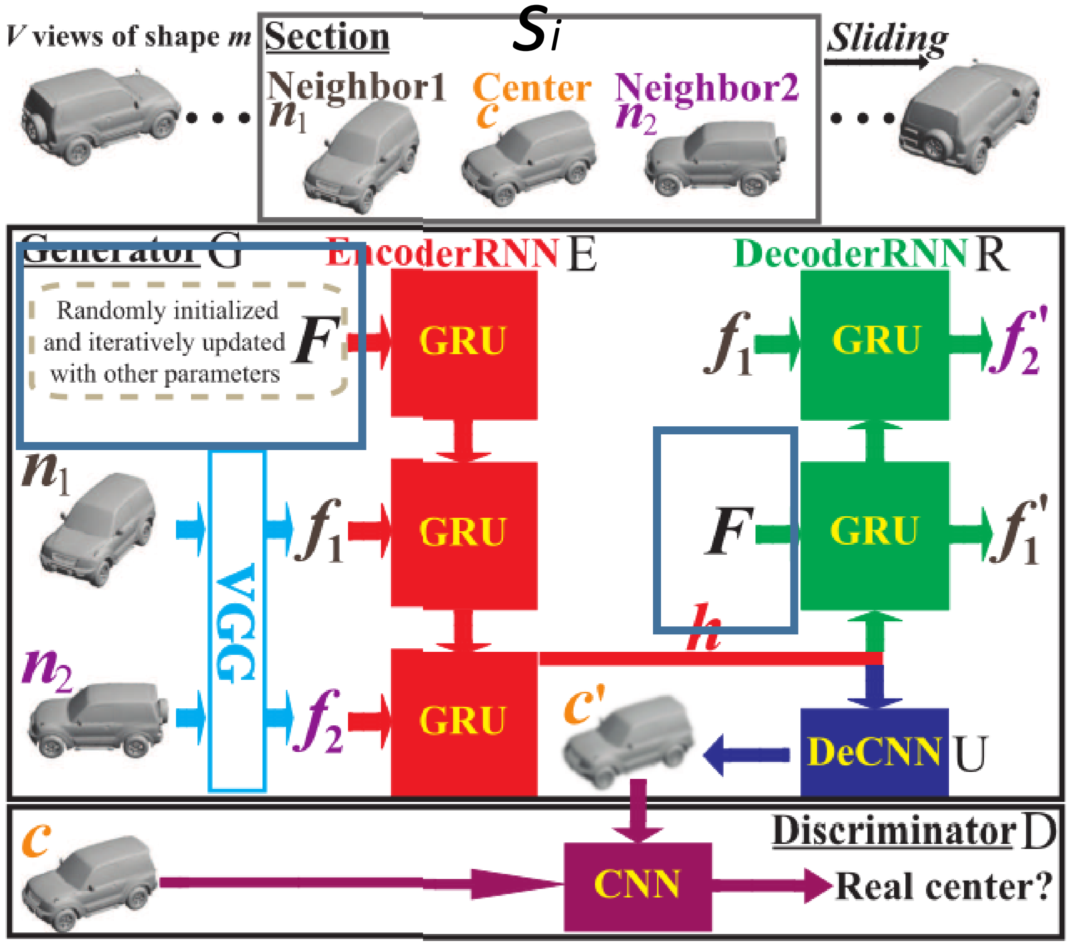
$$L = L_{\mathrm{U}} + \alpha L_{\mathrm{R}} + \beta L_{\mathrm{D2U}}$$

α and β control the balance among the aforementioned losses.

F is learned by being iteratively updating:

$$\boldsymbol{F} \leftarrow \boldsymbol{F} - \varepsilon \times \partial L / \partial \boldsymbol{F}$$

ε is the learning rate.



Results

- Experimental evaluation in
  - 3D shape classification
  - 3D shape retrieval
- Dataset
  - ModelNet10
  - ModelNet40
  - ShapeNet

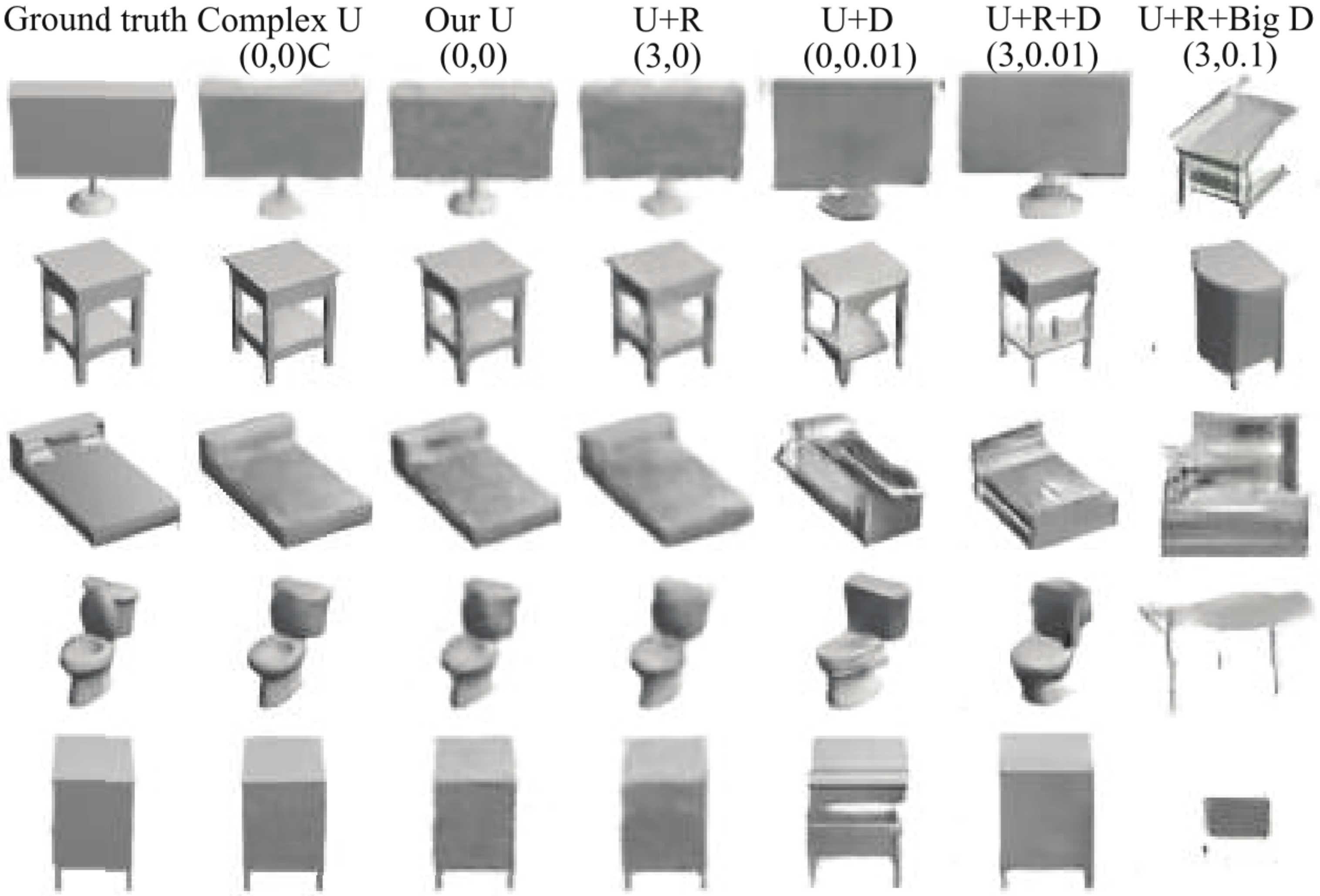
Results

- Parameter setup and ablation studies

	Only U	Only U-C	Only R	Only D	U+3*R	U+3*R+0.05*D	CGAN
Instance	84.80	75.77	90.53	47.80	92.51	**94.05**	89.10
Class	83.96	74.78	89.88	44.49	92.08	**93.71**	88.34

Results

- Generated center views



Results

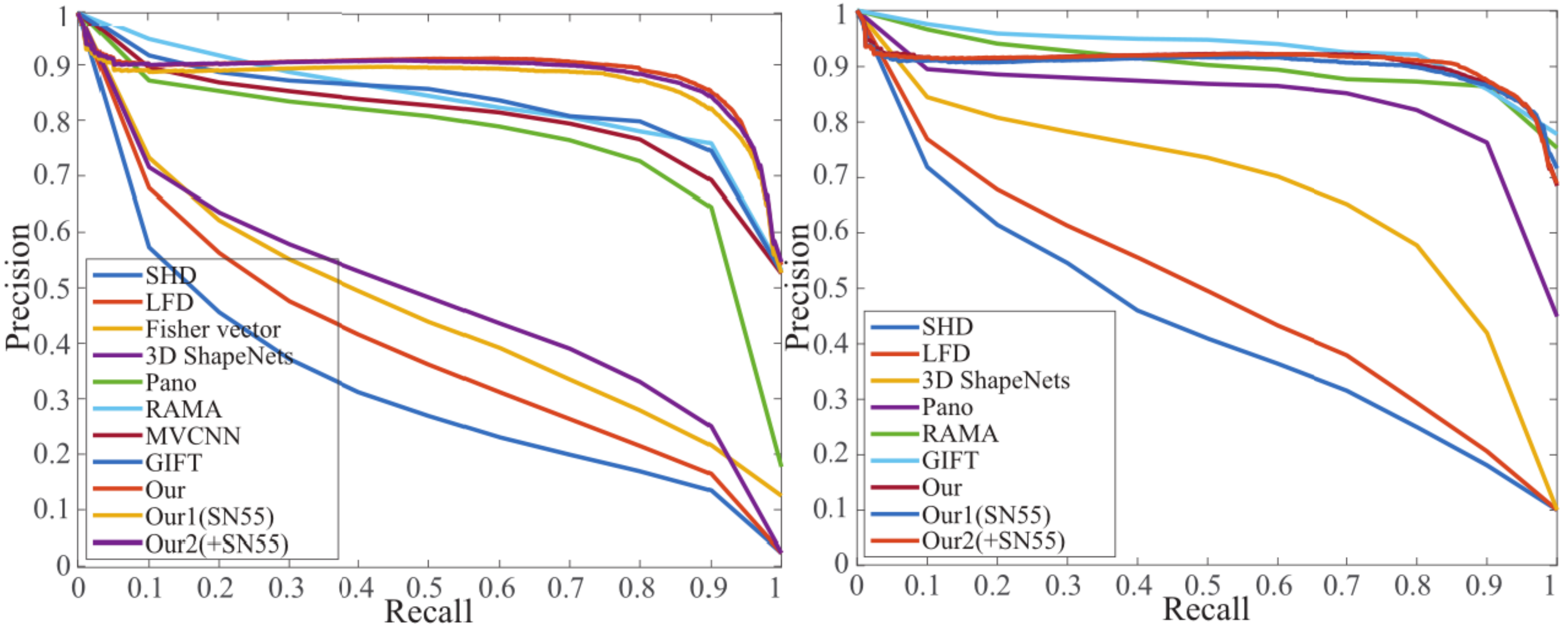
- Classification

Methods	Supervised	MN40	MN10
MVCNN	Yes	90.10	-
MVCNN-Multi	Yes	91.40	-
ORION	Yes	-	93.80
3DDescriptorNet	Yes	-	92.40
Pairwise	Yes	90.70	92.80
GIFT	Yes	89.50	91.50
PANORAMA	Yes	90.70	91.12
VoxNet	Yes	-	92.00
VRN	Yes	91.33	93.80
RotationNet	Yes	90.65	93.84
PointNet++	Yes	91.90	-
T-L	No	74.40	-
LFD	No	75.47	79.90
Vconv-DAE	No	75.50	80.50
3DGAN	No	83.30	91.00
LGAN	No	85.70	95.30
LGAN(MN40)	No	87.27	92.18
FNet	No	88.40	94.40
FNet(MN40)	No	84.36	91.85
Our	No	**91.98**	**94.05**
Our1(SN55)	No	90.19	92.18
Our2(+SN55)	No	91.25	92.84

Activate Windows
Go to Settings to activate

Results

- Retrieval



(a) ModelNet40 (b) ModelNet10

Methods	MN40	MN10
GeoImage	51.30	74.90
Pano	76.81	84.18
MVCNN	79.50	-
GIFT	81.94	91.12
RAMA	83.45	87.39
Trip	88.00	-
Our	**89.23**	**90.69**
Our1(SN55)	**87.66**	**90.09**
Our2(+SN55)	**88.87**	**90.75**

Results

- Retrieval under ShapeNet55

	Micro				
Methods	P	R	F1	mAP	NDCG
Kanezaki	81.0	80.1	**79.8**	77.2	86.5
Zhou	78.6	77.3	76.7	72.2	82.7
Tatsuma	76.5	80.3	77.2	74.9	82.8
Furuya	**81.8**	68.9	71.2	66.3	76.2
Thermos	74.3	67.7	69.2	62.2	73.2
Deng	41.8	71.7	47.9	54.0	65.4
Li	53.5	25.6	28.2	19.9	33.0
Mk	79.3	21.1	25.3	19.2	27.7
Su	77.0	77.0	76.4	73.5	81.5
Bai	70.6	69.5	68.9	64.0	76.5
Taco	70.1	71.1	69.9	67.6	75.6
Our	60.0	**80.3**	61.2	**83.5**	**89.4**
Our+	60.0	**80.3**	61.2	**83.6**	**89.5**
Our accuracy		**82.97**			
Our+ accuracy		**82.51**			

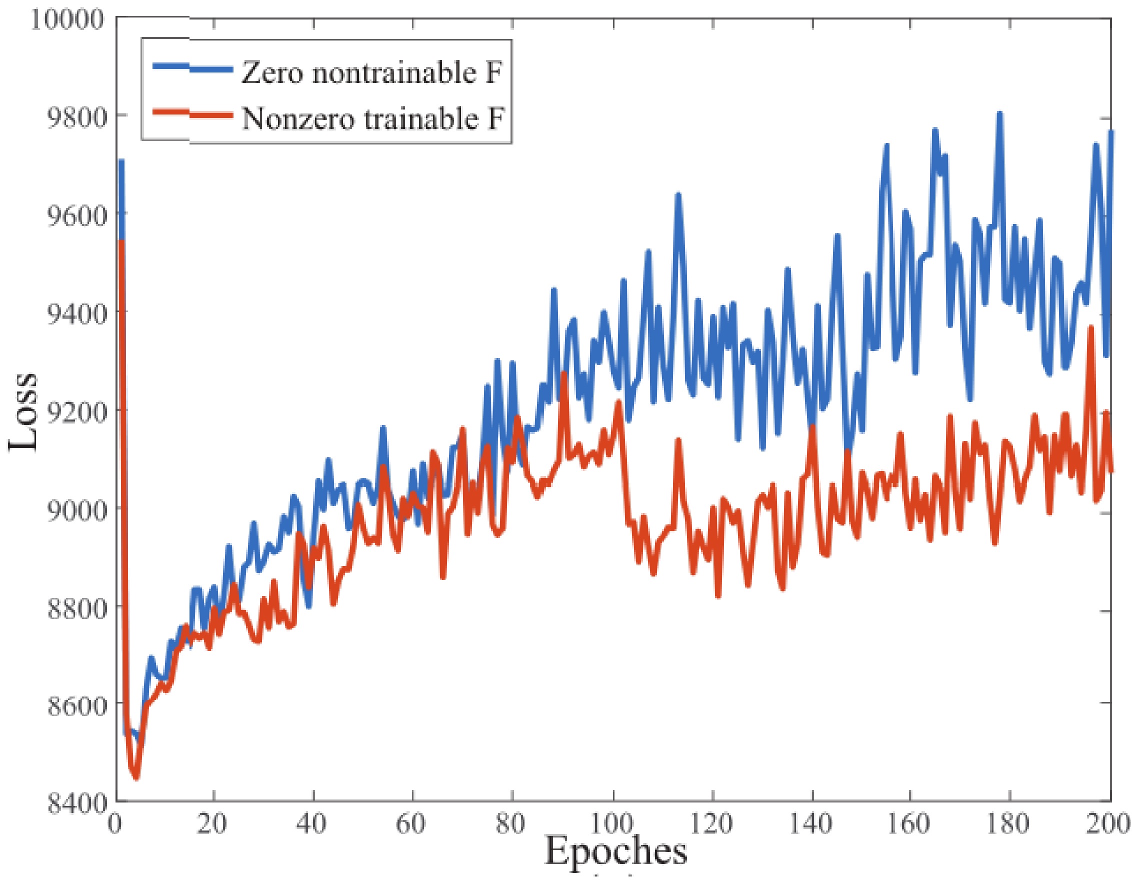
	Macro				
Methods	P	R	F1	mAP	NDCG
Kanezaki	60.2	63.9	**59.0**	58.3	65.6
Zhou	59.2	65.4	58.1	57.5	65.7
Tatsuma	51.8	60.1	51.9	49.6	55.9
Furuya	**61.8**	53.3	50.5	47.7	56.3
Thermos	52.3	49.4	48.4	41.8	50.2
Deng	12.2	66.7	16.6	33.9	40.4
Li	21.9	40.9	19.7	25.5	37.7
Mk	59.8	28.3	25.8	23.2	33.7
Su	57.1	62.5	57.5	56.6	64.0
Bai	44.4	53.1	45.4	44.7	54.8
Our	18.9	**81.2**	24.0	**69.2**	**83.7**
Our+	18.8	**81.3**	24.0	**69.9**	**84.0**

Activ

Results

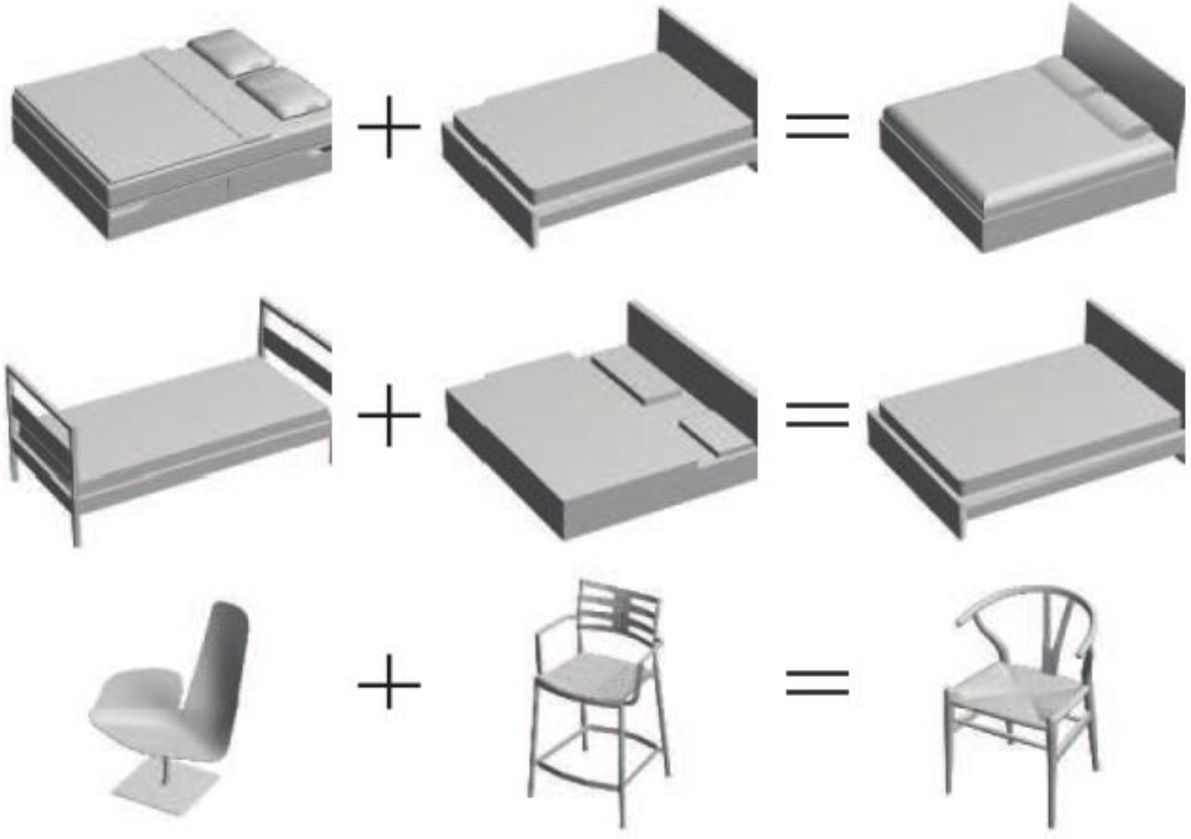
- Effectiveness of our implicit view aggregation

ACC	Non-trainable $\boldsymbol{F}$		Trainable $\boldsymbol{F}$		
	MaxP	MeanP	MaxP	MeanP	Our
Ins	84.58	87.22	81.72	82.49	**94.05**
Cla	83.95	87.38	80.60	81.73	**93.71**



Results

- Effectiveness of our implicit view aggregation



Contributions

- We propose VIP-GAN as a novel deep learning model to perform unsupervised 3D global feature learning through view inter-prediction with adversarial training, which leads to state-of-the-art performance in shape classification and retrieval.
- VIP-GAN makes it possible to mine fine-grained “supervised” information within the multi-view context of 3D shapes by imitating human perception of view-dependent patterns, which facilitates effective unsupervised 3D global feature learning.
- We introduce a novel implicit aggregation technique for 3D global feature learning based on RNN, which enables VIP-GAN to aggregate knowledge learned from each view prediction across a view sequence effectively.

Thank you!