Let Images Give You More: Point Cloud Cross-Modal Training for Shape Analysis

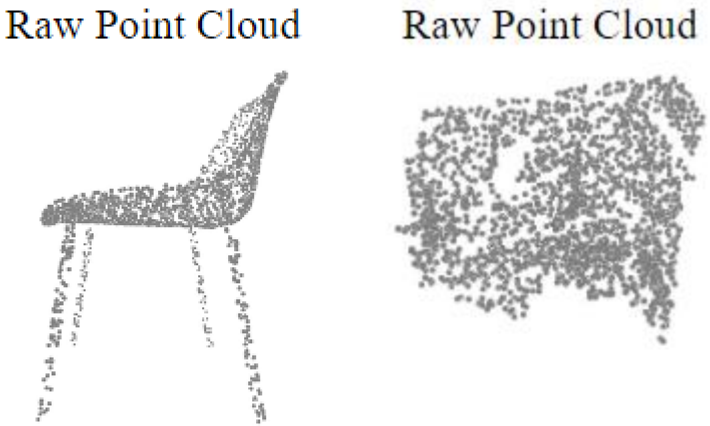
Xu Yan[1,2†], Heshen Zhan[1,2†], Chaoda Zheng[1,2], Jiantao Gao[4],
Ruimao Zhang[3], Shuguang Cui[2,1,5], Zhen Li[2,1]*

[1]FNii, CUHK-Shenzhen, [2]SSE, CUHK-Shenzhen,
[3]SDS, CUHK-Shenzhen, [4]USV, Shanghai University, [5]Pengcheng Lab

Motivation

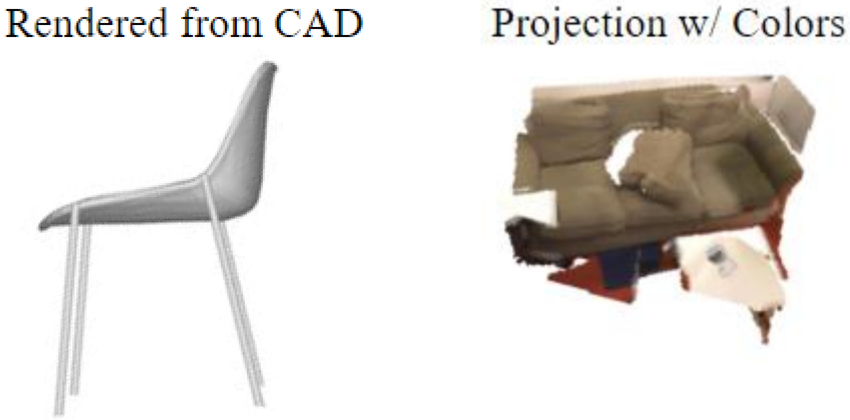
3D point cloud:
Partial and geometric information.
Only sparse and textureless features.

2D image:
Rich color and fine-grained texture.
Ambiguous in depth and shape sensing.



Two problems

Or one?



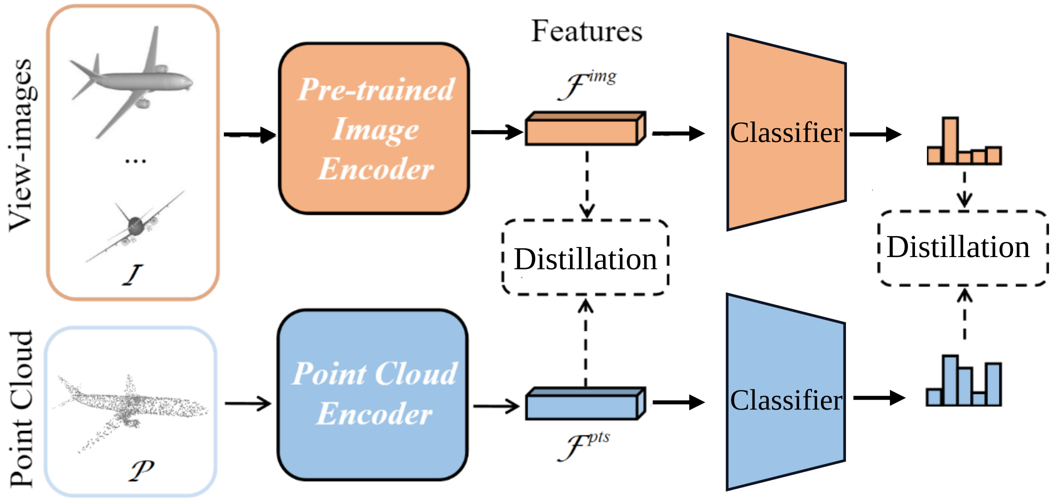
Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

Could we use the rich information hidden in 2D images to boost 3D point cloud shape analysis?

Motivation

Knowledge distillation:
Takes extra image inputs **only in training phases**.
Not computation-intensive during inference.
Don't need paired-images during inference.

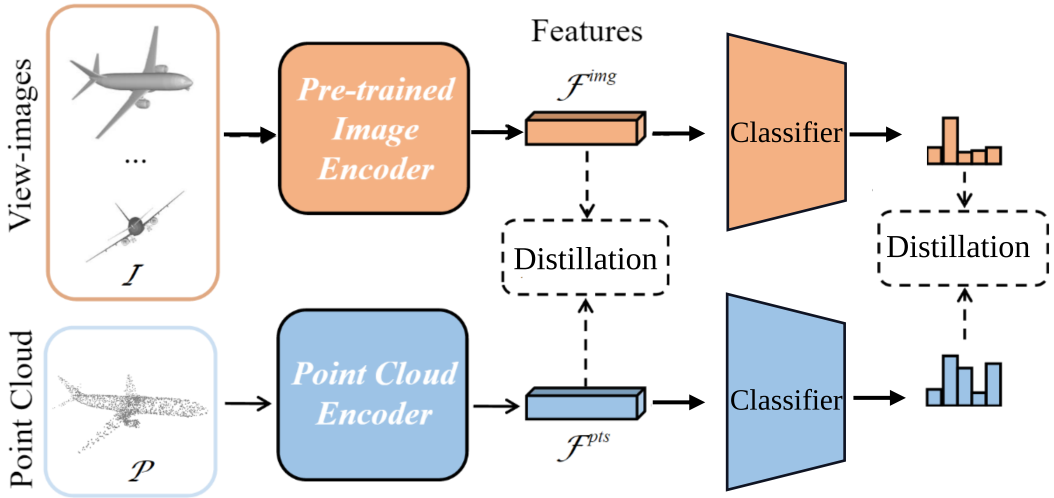


Motivation

Our cross-modality setting:

3D and 2D data contain different information.

Encoders are quite different.



Motivation

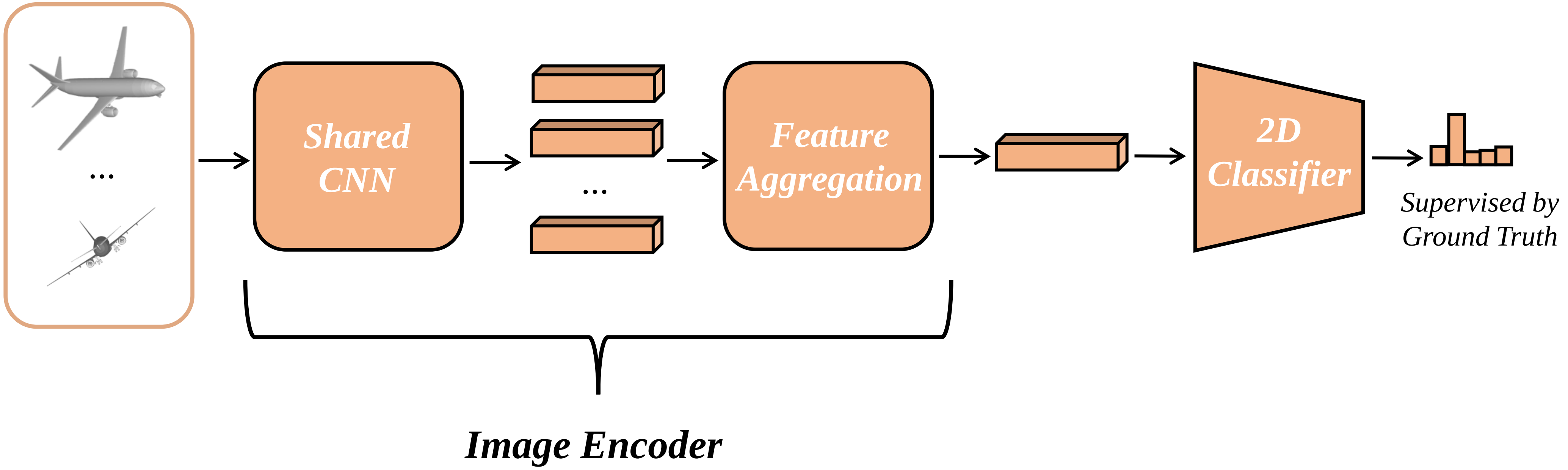
NEURAL INFORMATION PROCESSING SYSTEMS

New cross-modal knowledge distillation methods are needed!

PointCMT

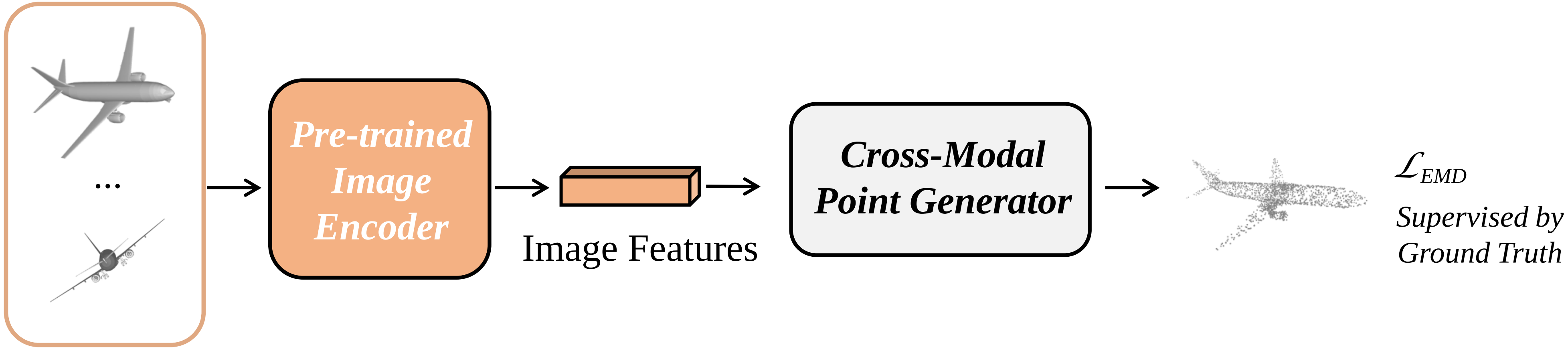
Obtaining image encoder:

$$\mathcal{F}^{img} = \mathcal{A}\{\text{CNN}(\mathcal{I}_v)\}_{v=1}^{V}.$$



PointCMT

Training cross-modal point generator (CMPG):



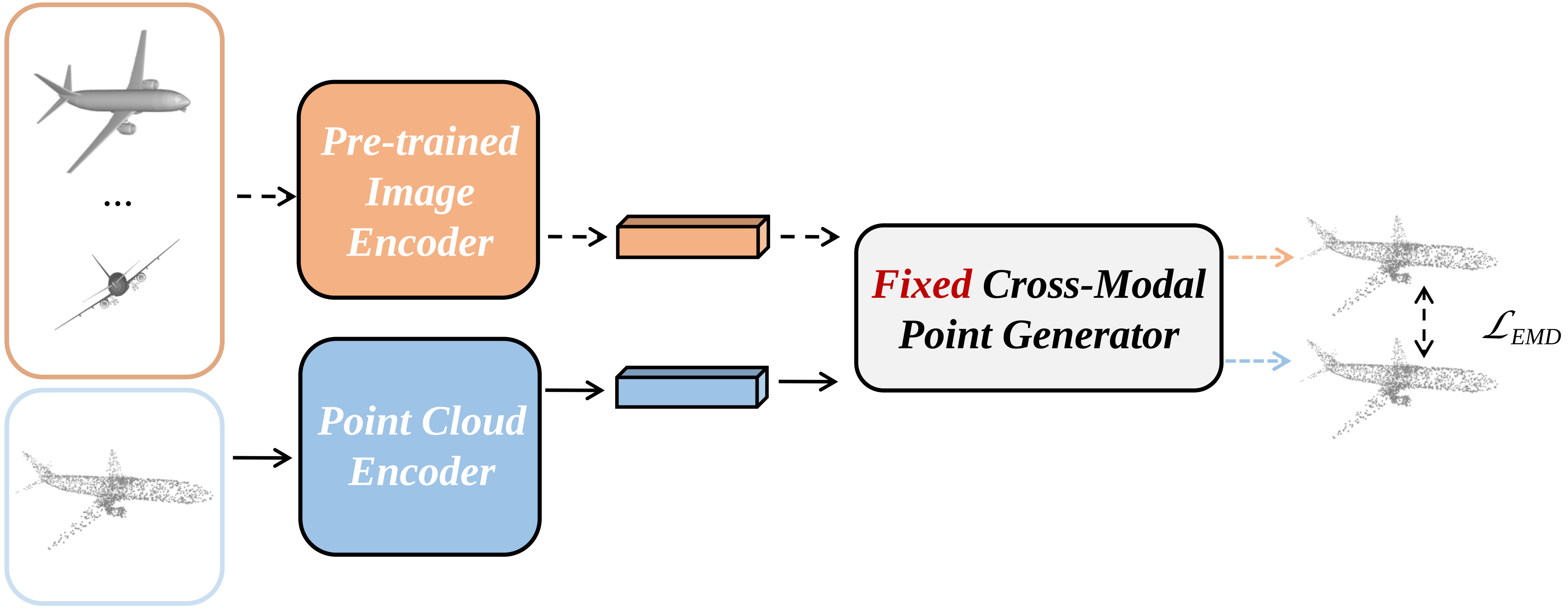
$$\mathcal{L}_{\text{EMD}}(\mathcal{P}, \hat{\mathcal{P}}^{img}) = \min_{\phi} \sum_{p \in \mathcal{P}} ||p - \phi(p)||$$

Compared with traditional L_2 loss, the EMD distance is natural for solving an assignment problem for permutation-invariant point sets!

PointCMT

Feature Enhancement Loss:

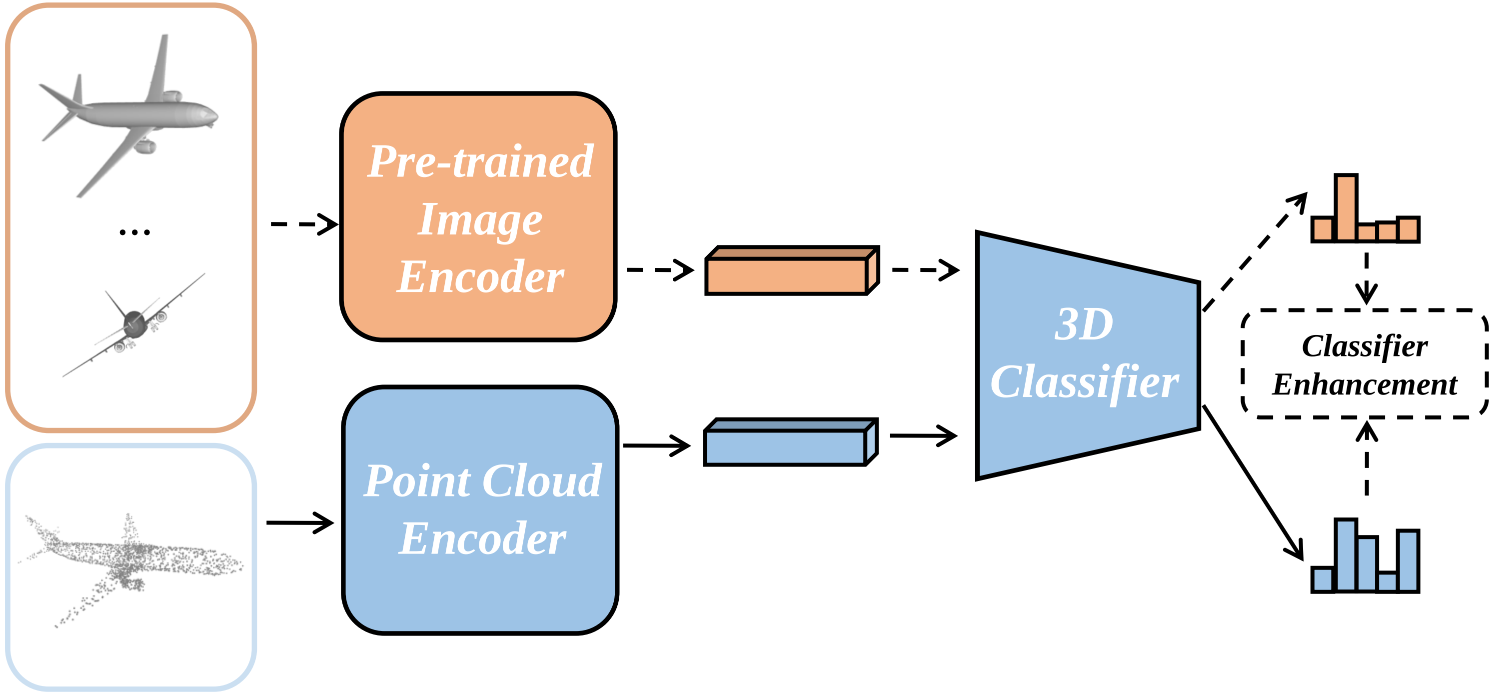
$$\mathcal{L}_{\text{Feature}} = \mathcal{L}_{\text{EMD}}(\hat{\mathcal{P}}^{pts}, \hat{\mathcal{P}}^{img}) = \min_{\phi} \sum_{p \in \hat{\mathcal{P}}^{pts}} ||p - \phi(p)||$$



PointCMT

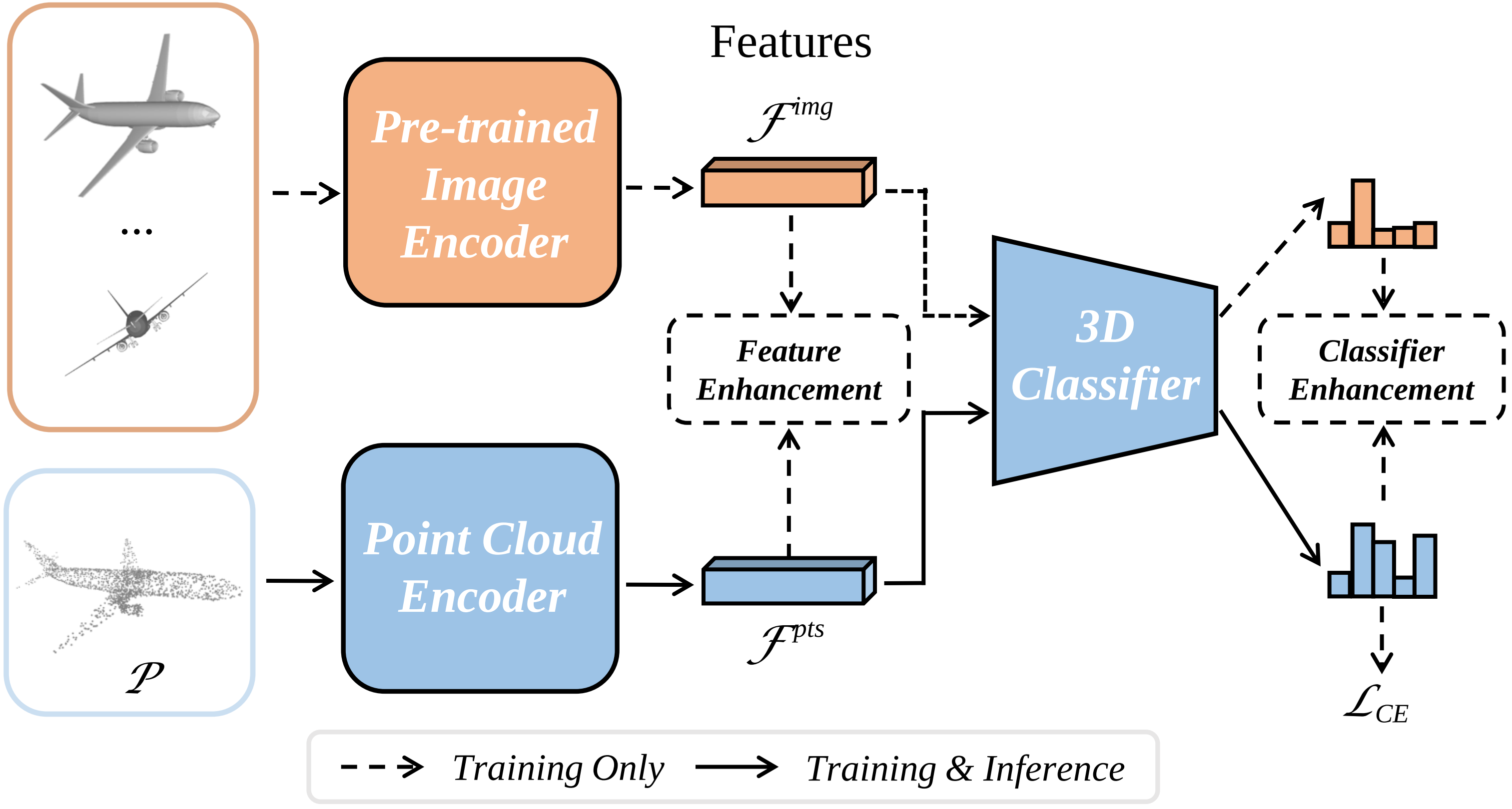
Classifier Enhancement Loss:

$$\mathcal{L}_{\text{Classifier}} = \mathcal{D}_{KL}(\text{Cls}^{pts}(\mathcal{F}^{img})||\text{Cls}^{pts}(\mathcal{F}^{pts}))$$



PointCMT

The whole framework of PointCMT:



Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

Classification results on ModelNet40 dataset

Method	Input	#Points	mAcc(%)	OA(%)	Speed	Param.
PointNet [33]	pnt	1k	86.0	89.2	-	3.47M
PointNet++ [34]	pnt, nor	5k	-	91.9	-	1.47M
PointCNN [25]	pnt	1k	88.0	92.5	-	-
PointConv [47]	pnt, nor	1k	-	92.5	$80^{\dagger}$	18.6M
KPConv [39]	pnt	7k	-	92.9	$10^{\dagger}$	15.2M
PointASNL [54]	pnt, nor	1k	-	93.2	-	-
PosPool [29]	pnt	5k	-	93.2	-	-
Point Transformer [60]	pnt	1k	90.6	93.7	-	-
GBNet [36]	pnt	1k	91.0	93.8	$112^{\dagger}$	8.4M
GDANet [51]	pnt	1k	-	93.8	$14^{\dagger}$	0.9M
SimpleView [12]	pnt	1k	-	93.9	2208	1.64M
CurveNet [49]	pnt	1k	-	94.2	$15^{\dagger}$	2.0M
PointMLP [31]	pnt	1k	**91.4**	**94.5**	139	12.6M
DGCNN [43] (baseline)	pnt	1k	90.2	92.9	518	1.68M
RS-CNN [27] (baseline)	pnt	1k	89.3	92.9	2174	1.17M
PointNet++ [34] (baseline)	pnt	1k	90.1	93.4*	300	1.62M
DGCNN **w/** PointCMT	pnt	1k	90.8 (+0.6)	93.5 (+0.6)	518	1.68M
RS-CNN **w/** PointCMT	pnt	1k	90.1 (+0.8)	93.8 (+0.9)	2174	1.17M
PointNet++ **w/** PointCMT	pnt	1k	<u>91.2</u> (+1.1)	<u>94.4</u> (+1.0)	300	1.62M

Experiment

Classification results on ScanObjectNN dataset:

Method	OBJ_ONLY mAcc(%)	OBJ_ONLY OA(%)	PB_T50_RS mAcc(%)	PB_T50_RS OA(%)
3DmFV [3]	-	73.8	58.1	63.0
PointNet [33]	-	79.2	63.4	68.2
SpiderCNN [52]	-	79.5	69.8	73.7
PointNet++ [34]	-	84.3	75.4	77.9
DGCNN [43]	-	86.2	73.6	78.1
PointCNN [25]	-	85.5	75.1	78.5
DRNet [35]	-	-	78.0	80.3
GBNet [36]	-	-	77.8	80.5
SimpleView [12]	86.2	89.0	-	80.8
PRANet [4]	-	-	79.1	82.1
MVTN [15]	-	-	-	82.8
PointNet++ [34] (baseline)	85.4±0.2	87.4±0.1	75.5±0.3	79.2±0.2
PointMLP [31] (baseline)	89.1±0.3	92.2±0.3	83.9±0.5	85.4±0.3
PointNet++ w/ PointCMT	89.0±0.3 (+3.7)	91.6±0.2 (+4.3)	79.9±0.3 (+4.4)	83.1±0.2 (+3.9)
PointMLP w/ PointCMT	**91.8±0.2** (+2.6)	**93.2±0.3** (+1.0)	**84.4±0.4** (+0.4)	**86.4±0.3** (+1.0)

Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

Ablation study on ModelNet40 and ScanObjetNN dataset:

Model	FE	CE	ModelNet40	OBJ_ONLY	PB_T50_RS
PointNet++	✗	✗	93.4	87.5	79.4
	✓	✗	93.8 (+0.4)	89.2 (+1.7)	82.5 (+3.1)
	✗	✓	94.0 (+0.6)	91.3 (+3.8)	82.3 (+2.9)
	✓	✓	94.4 (+1.0)	91.8 (+4.3)	83.3 (+3.9)

Comparison with Knowledge Distillation Methods:

Method	ModelNet40	PB_T50_RS
Baseline	93.4	79.4
Hinton *et al.* [17]	93.1 (-0.3)	81.8 (+2.4)
Huang *et al.* [21]	93.6 (+0.2)	82.0 (+2.6)
Yang *et al.* [55]	93.9 (+0.5)	81.1 (+1.7)
PointCMT (ours)	**94.4** (+1.0)	**83.3** (+3.9)

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

NEURAL INFORMATION PROCESSING SYSTEMS

Let Images Give You More: Point Cloud Cross-Modal Training for Shape Analysis

Thanks for watching!

Xu Yan[1,2†], *Heshen Zhan*[1,2†], *Chaoda Zheng*[1,2], *Jiantao Gao*[4],
Ruimao Zhang[3], *Shuguang Cui*[2,1,5], *Zhen Li*[2,1*]

[1]FNii, CUHK-Shenzhen, [2]SSE, CUHK-Shenzhen,
[3]SDS, CUHK-Shenzhen, [4]USV, Shanghai University, [5]Pengcheng Lab