Seeing 3D Through 2D Lenses: 3D Few-Shot Class-Incremental Learning via Cross-Modal Geometric Rectification

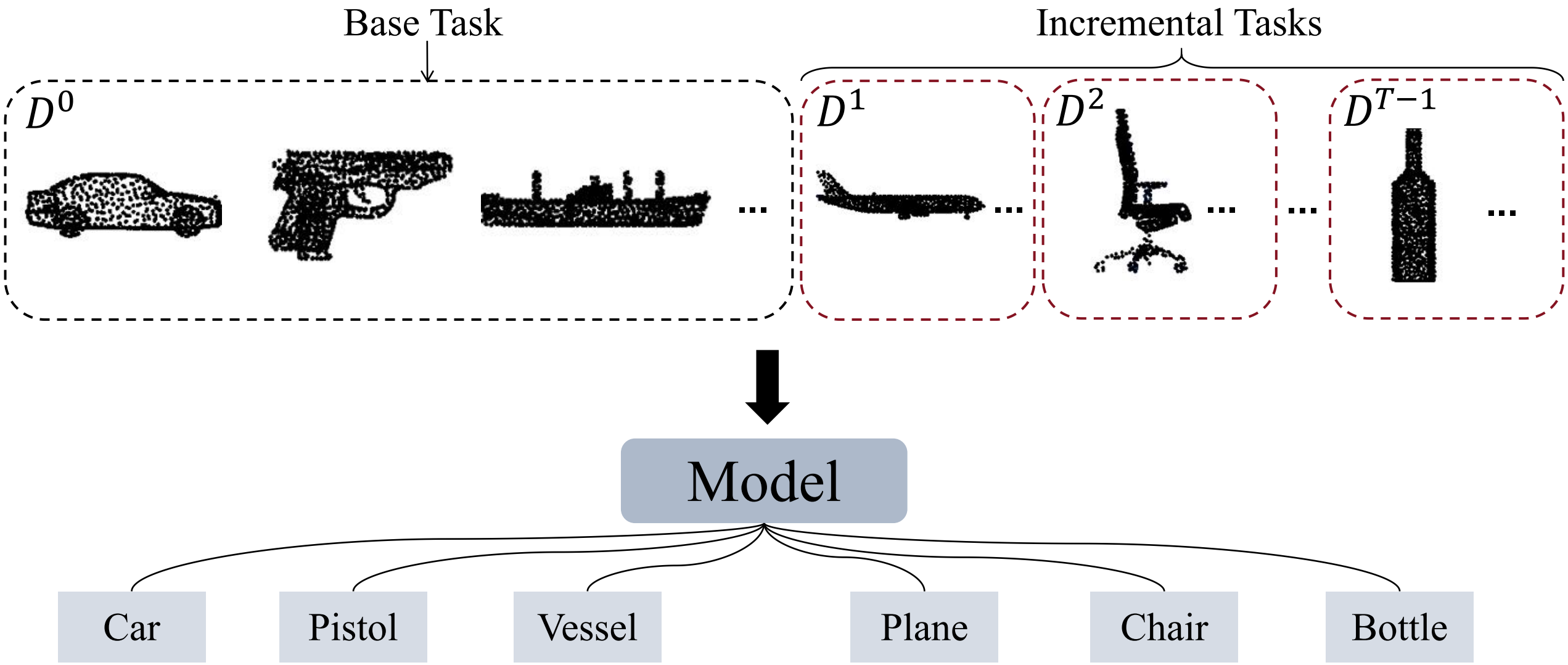
Tuo Xiang[1] Xuemiao Xu[1,2,3,4] Bangzhen Liu[1*] Jinyi Li[1] Yong Li[1*] Shengfeng He[5]

[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science

[3]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information

[4]Ministry of Education Key Laboratory of Big Data and Intelligent Robot [5]Singapore Management University

Introduction



Introduction

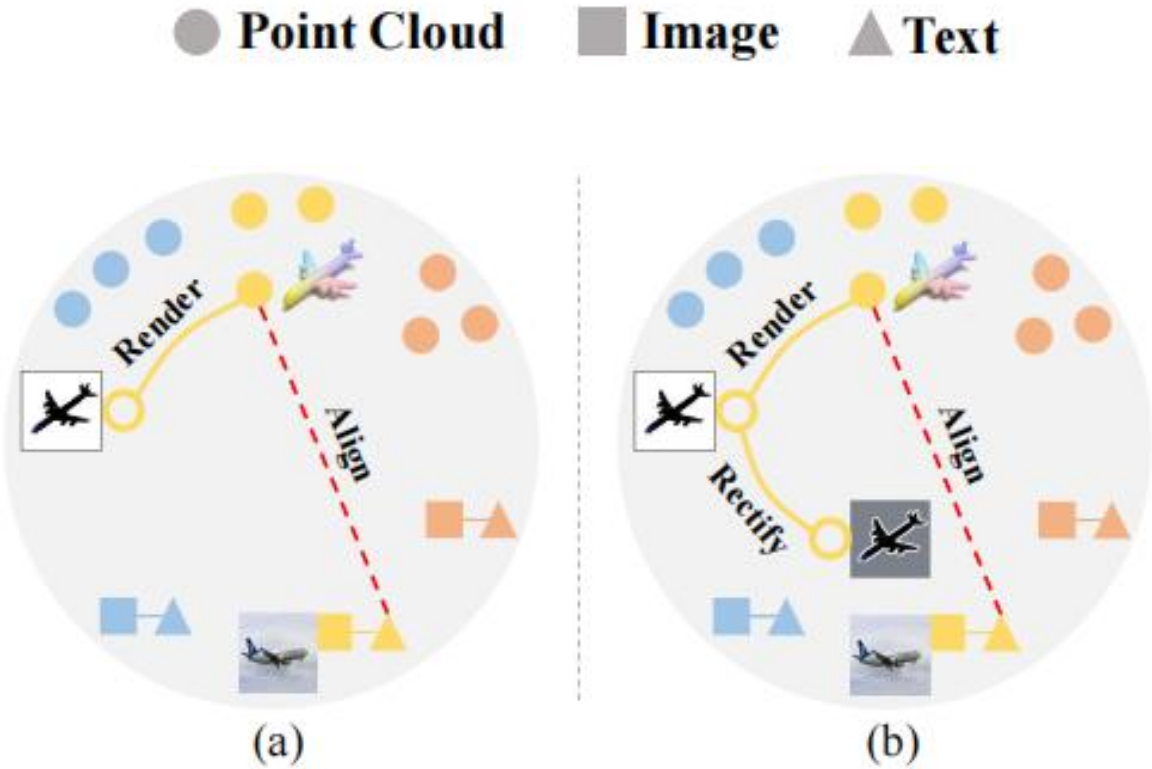
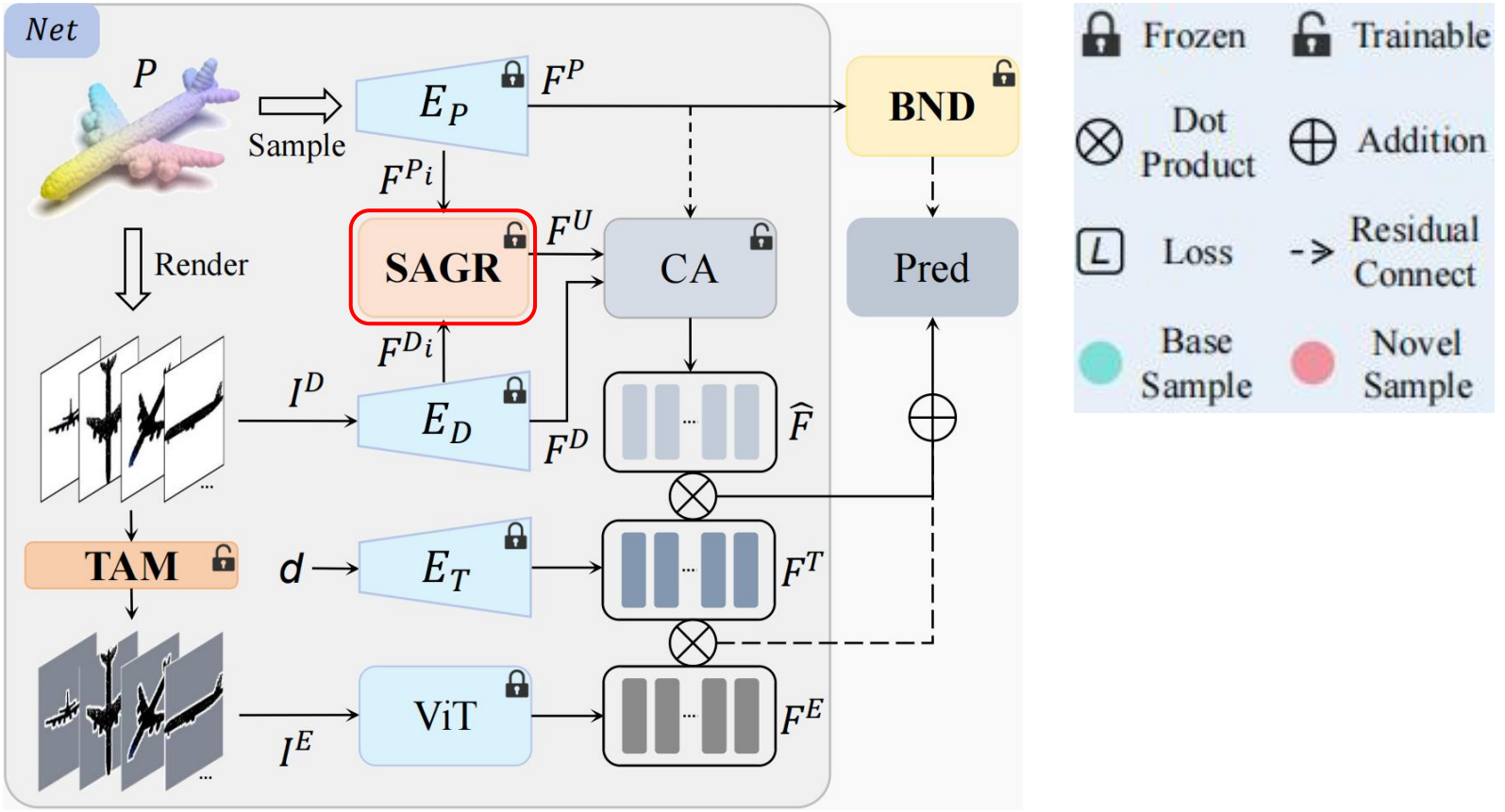
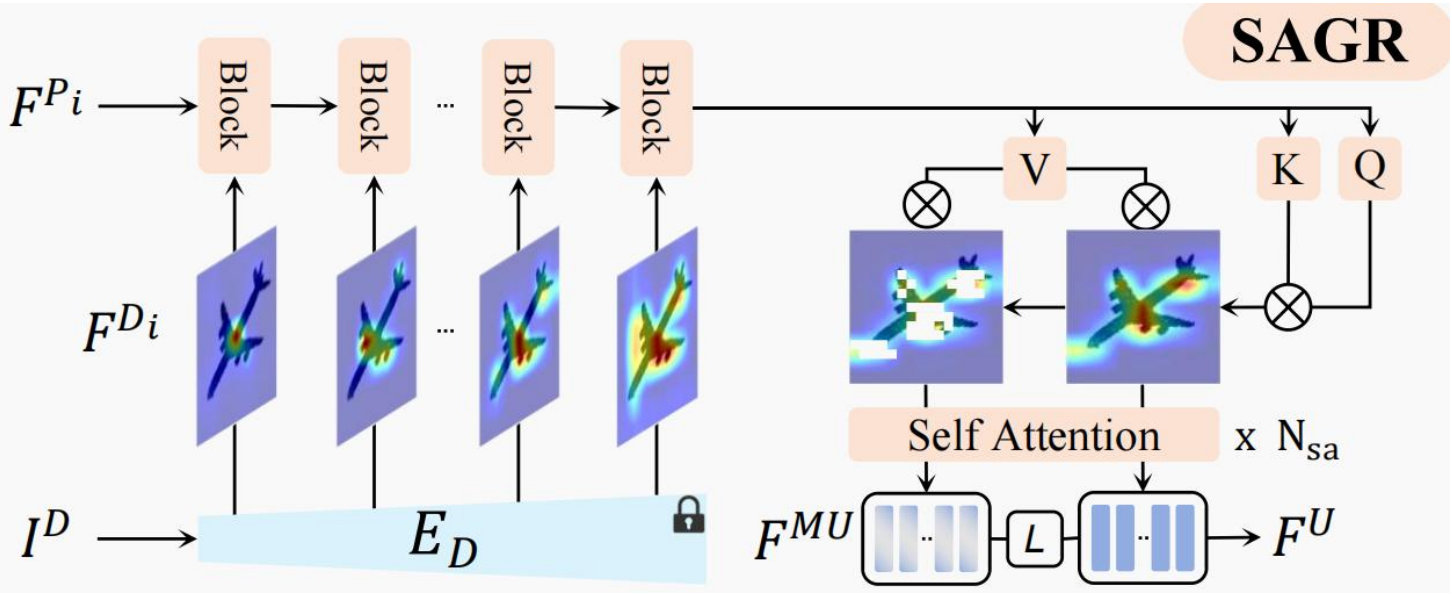


Figure 1. Direct alignment of point clouds with CLIP's text features proves challenging due to the inherent domain gap. (a) While projecting depth maps to leverage 2D priors can bring point cloud and text features closer, discrepancies between the depth map and CLIP's input space impede effective alignment. (b) In contrast, through 2D-3D rectification, point clouds are enriched with hierarchical 2D information derived from CLIP, facilitating a more precise alignment with text features.

Introduction



Method



$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

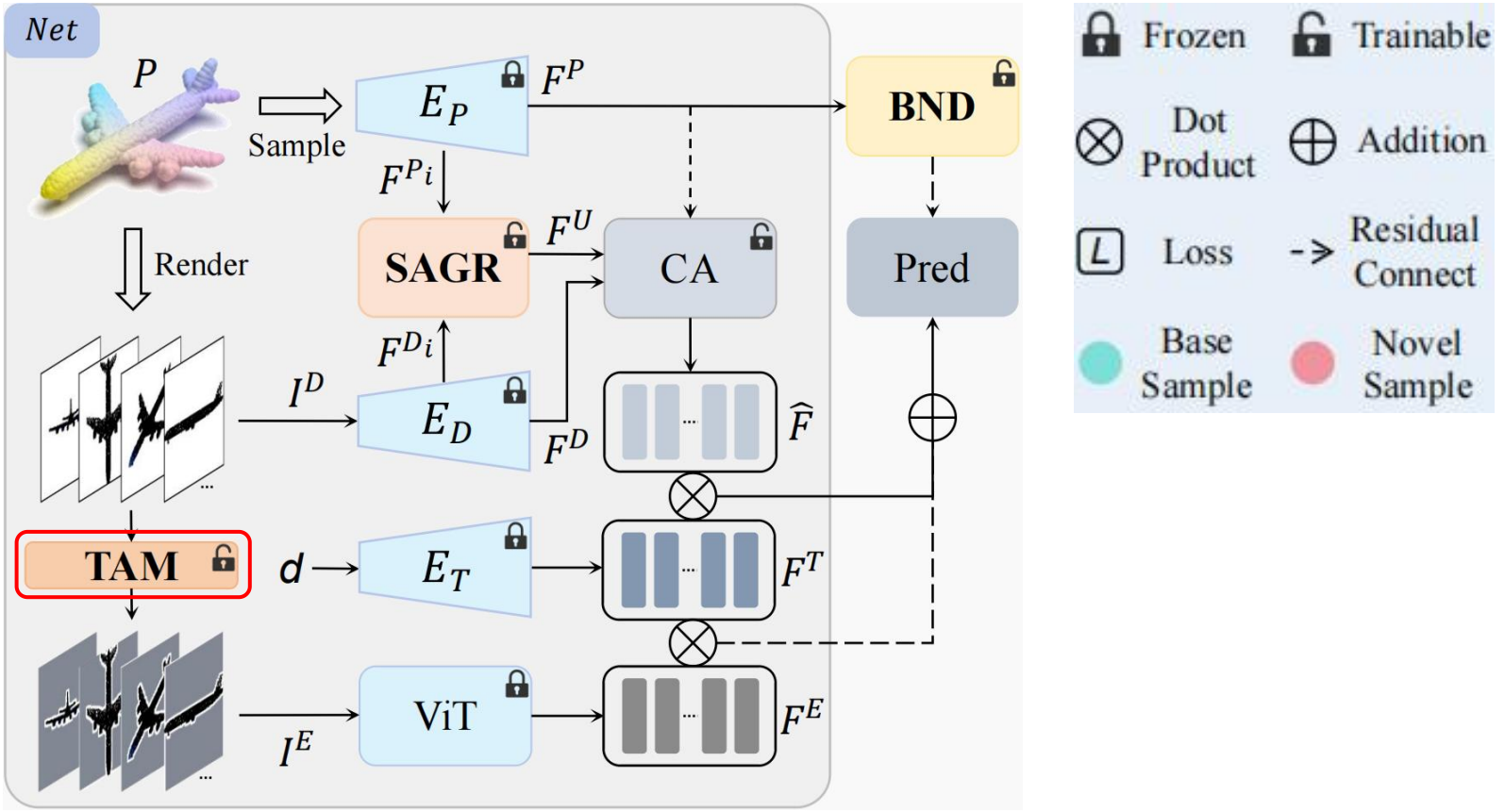
$$F^{P_{i+1}} = \begin{cases} \text{Attn}(F^{P_i}, F^{D_i}, F^{D_i}) & i \in L_r \\ \text{Attn}(F^{P_i}, F^{P_i}, F^{P_i}) & \text{otherwise} \end{cases},$$

$$R^M[q,k] = \begin{cases} R[q,k] & R[q,k] < \text{Top}(M_R) \\ 0 & \text{otherwise} \end{cases}.$$

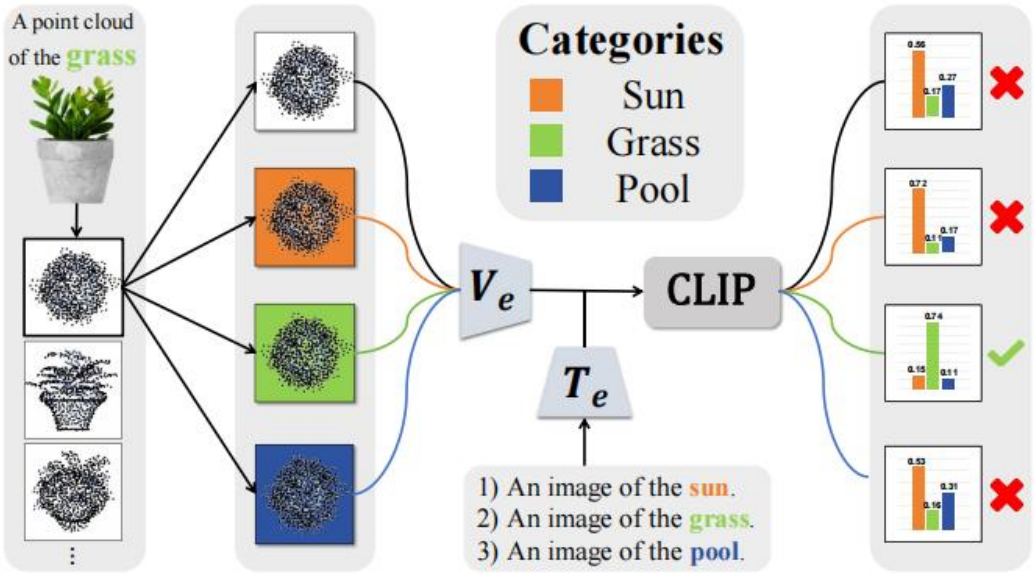
$$\mathcal{L}_{mc} = \frac{1}{B^2}\sum_{i,j=1}^{B} ||sim(F_i^U, F_j^U) - sim(F_i^{MU}, F_j^{MU})||_2^2$$

$$\widehat{F} = f\left((f'(\text{concat}(F^P, F^U)) + w \cdot F^D) \odot \lambda\right),$$

Introduction



Method

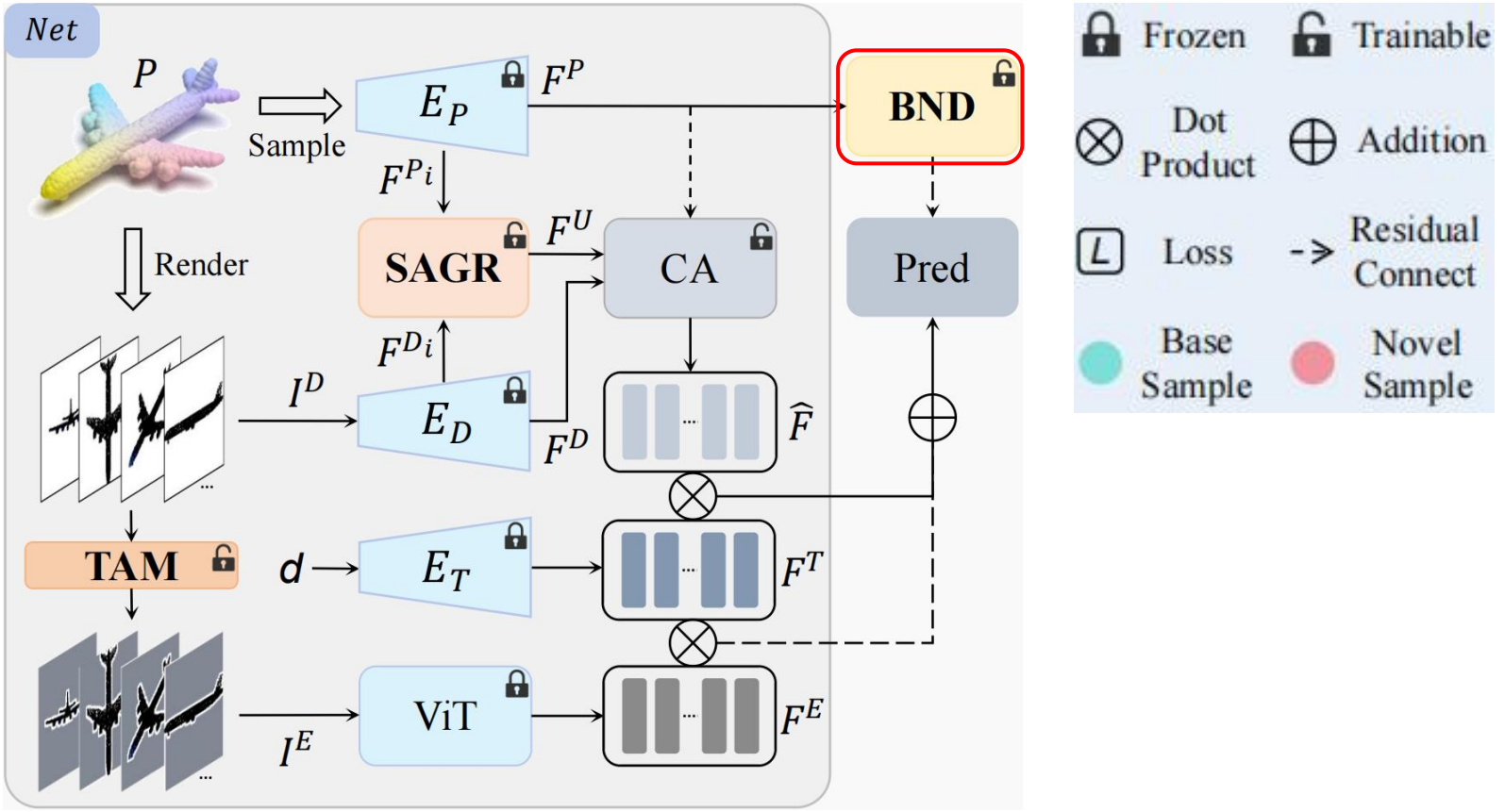


$$\text{logits} = \mathbf{F^T} \cdot \widehat{\mathbf{F}} + \text{CLIP}(I^E, d),$$

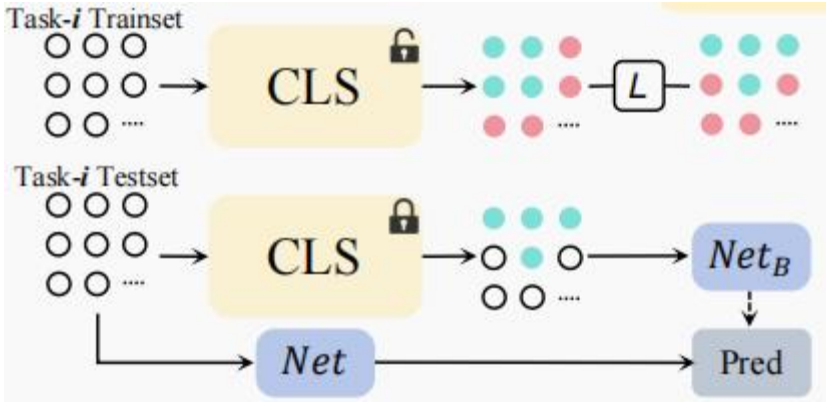
$$\mathbf{c} = \frac{\tanh\left(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{F^P} + \mathbf{b}_1) + \mathbf{b}_2\right) + 1}{2},$$

$$\mathcal{L}_c = \frac{1}{V}\left(\sum_{i=1}^{V} \frac{1 - \cos(\mathbf{F}_i^E, \mathbf{F^T})}{2}\right)$$

Introduction



Method



$$\mathcal{L}_{\text{BND}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{1}{|D_{\text{train}}^{t}|}\sum_{i=1}^{|D_{\text{train}}^{t}|}[-y_i\log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)]$$

Experiment

Table 1. Quantitative comparison under cross-domain settings. The best and second-best results are **bolded** and underlined, respectively.

Method \ Task	ShapeNet → CO3D													ModelNet → ScanObjectNN						ShapeNet → ScanObjectNN					
	39	44	49	54	59	64	69	74	79	84	89	$\Delta_A \downarrow$	$AA \uparrow$	26	30	34	37	$\Delta_A \downarrow$	$AA \uparrow$	44	49	54	59	$\Delta_A \downarrow$	$AA \uparrow$
FT	81.0	20.2	2.3	1.7	0.8	1.0	1.0	1.3	0.9	0.5	1.6	59.3	10.2	88.4	6.4	6.0	1.9	55.8	25.7	81.4	38.7	4.0	0.9	73.2	31.3
Joint	81.0	79.5	78.3	75.2	75.1	74.8	72.3	71.3	70.0	68.8	67.3	1.8	74.0	88.4	79.7	74.0	71.2	6.9	78.3	81.4	82.5	79.8	78.7	2.0	80.6
LwF [19]	81.0	57.4	19.3	2.3	1.0	0.9	0.8	1.3	1.1	0.8	1.9	50.4	15.3	88.4	35.8	5.8	2.5	66.7	33.1	81.4	47.9	14.0	5.9	56.6	37.3
IL2M [3]	81.0	45.6	36.8	35.1	31.8	33.3	34.0	31.5	30.6	32.3	30.0	10.7	38.4	88.4	58.2	52.9	52.0	15.0	62.9	81.4	53.2	43.9	45.8	18.8	56.1
ScaIL [4]	81.0	50.1	45.7	39.1	39.0	37.9	38.0	36.0	33.7	33.0	35.2	8.5	42.6	88.4	56.5	55.9	52.9	14.2	63.4	81.4	49.0	46.7	40.0	19.6	54.3
EEIL [5]	81.0	75.2	69.3	63.2	60.5	57.9	53.0	51.9	51.3	47.8	47.6	5.1	59.9	88.4	70.2	61.0	56.8	13.5	69.1	81.4	74.5	69.8	63.4	8.0	72.3
FACT [44]	81.4	76.0	70.3	68.1	65.8	63.5	63.0	60.1	58.2	57.5	55.9	3.7	65.4	89.1	72.5	68.3	63.5	10.5	73.4	82.3	74.6	69.9	66.8	6.7	73.4
Sem-aware [8]	80.6	69.5	66.5	62.9	63.2	63.0	61.2	58.3	58.1	57.2	55.2	3.7	63.2	88.5	73.9	67.7	64.2	10.0	73.6	81.3	70.6	65.2	62.9	8.1	70.0
Microshape [10]	82.6	77.9	73.9	72.7	67.7	66.2	65.4	63.4	60.6	58.1	57.1	3.6	67.8	89.3	73.2	68.4	65.1	9.8	74.0	82.5	74.8	71.2	67.1	6.6	73.9
FILP-3D [36]	91.4	80.7	80.6	76.2	75.7	68.2	66.8	62.9	59.1	60.2	57.1	4.9	70.8	**93.6**	85.0	78.1	74.1	7.5	82.7	92.3	87.3	83.8	82.4	3.7	86.5
C3PR [7]	83.6	80.0	77.8	75.4	72.8	72.3	70.3	67.9	64.9	64.1	63.2	**2.8**	72.0	88.3	75.7	70.6	67.8	8.3	75.6	84.5	77.8	75.5	71.9	5.2	77.4
Ours	**91.4**	**89.8**	**86.8**	**84.3**	**80.3**	**78.4**	**74.9**	**71.7**	**71.3**	**69.0**	**66.3**	3.2	**78.6**	93.3	**88.8**	**81.3**	**74.9**	**7.0**	**84.6**	**92.4**	**90.0**	**87.0**	**84.0**	**3.1**	**88.4**

Table 2. Quantitative comparison under within-domain settings. The best and second best results are **bolded** and underlined, respectively.

Method \ Task	ShapeNet → ShapeNet									ModelNet → ModelNet						
	25	30	35	40	45	50	55	$\Delta_A \downarrow$	$AA \uparrow$	20	25	30	35	40	$\Delta_A \downarrow$	$AA \uparrow$
FT	87.0	25.7	6.8	1.3	0.9	0.6	0.4	53.7	17.5	89.8	9.7	4.3	3.3	3.0	44.3	22.0
Joint	87.0	85.2	84.3	83.0	82.5	82.2	81.3	1.1	83.6	89.8	88.2	87.0	83.5	80.5	2.7	85.8
LwF [19]	87.0	60.8	33.5	15.9	3.8	3.1	1.8	44.0	29.4	89.8	36.0	9.1	3.6	3.1	52.2	28.3
IL2M [3]	87.0	58.6	45.7	40.7	50.1	49.4	49.3	15.0	54.4	89.8	65.5	58.4	52.3	53.6	12.7	63.9
ScaIL [4]	87.0	56.6	51.8	44.3	50.3	46.3	45.4	13.6	54.5	89.8	66.8	64.5	58.7	56.5	10.4	67.3
EEIL [5]	87.0	77.7	73.2	69.3	66.4	65.9	65.8	4.5	72.2	89.8	75.4	67.2	60.1	55.6	11.2	69.6
FACT [44]	87.5	75.3	71.4	69.9	67.5	65.7	62.5	5.4	71.4	90.4	81.3	77.1	73.5	65.0	7.9	77.5
Sem-aware [8]	87.2	74.9	68.1	69.0	68.1	66.9	63.8	5.4	71.1	91.3	82.2	74.3	70.0	64.7	8.2	76.5
Microshape [10]	87.6	83.2	81.5	79.0	76.8	73.5	72.6	3.1	79.2	93.6	83.1	78.2	**75.8**	67.1	7.9	79.6
C3PR [7]	88.0	81.6	77.8	76.7	76.9	76.2	74.7	2.7	78.8	91.6	82.3	75.8	72.2	**70.9**	**6.2**	78.6
Ours	**91.6**	**88.1**	**87.3**	**86.3**	**87.0**	**86.5**	**86.4**	**1.2**	**87.6**	**95.0**	**84.8**	**81.0**	72.2	65.9	8.7	**79.8**

Experiment

Table 3. Ablation studies on each module of our proposed framework on ShapeNet→CO3D.

Variants	SAGR	TAM	BND	39	44	49	54	59	64	69	74	79	84	89	$AA\uparrow$	PD(%)	$\Delta_A\downarrow$
Baseline				91.2	88.1	85.5	81.3	76.2	72.8	65.6	62.8	62.1	58.2	51.9	72.3	6.3	5.4
V1	✓			91.3	88.8	84.8	82.0	78.1	74.2	69.5	65.1	64.6	61.2	57.7	74.3	4.3	4.5
V2		✓		91.3	88.8	85.1	81.7	76.1	72.3	69.1	63.9	62.7	57.0	53.8	72.9	5.7	5.1
V3			✓	91.4	88.8	85.8	81.5	77.6	74.6	68.6	65.9	64.7	59.9	58.2	74.3	4.3	4.4
V4		✓	✓	91.4	88.7	86.3	82.1	78.1	74.9	69.9	66.8	65.9	62.1	59.8	75.1	3.5	4.1
V5	✓		✓	91.3	87.2	86.4	81.5	78.8	75.0	72.5	67.5	69.4	66.5	64.0	76.4	2.2	4.0
V6	✓	✓		91.2	88.9	84.7	81.2	77.5	74.7	70.8	67.9	67.2	63.1	60.1	75.3	3.3	4.1
Full	✓	✓	✓	**91.4**	**89.8**	**86.8**	**84.3**	**80.3**	**78.4**	**74.9**	**71.7**	**71.3**	**69.0**	**66.3**	**78.6**	-	**3.2**

Experiment

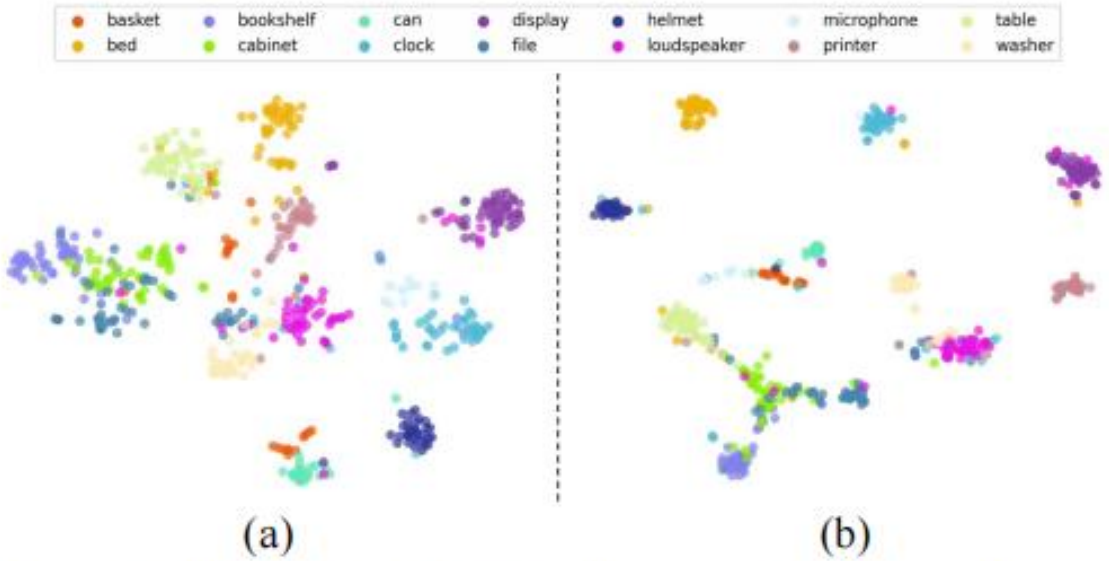


Figure 5. T-SNE visualizations before and after employing CMGR. (a) Relying solely on unrectified 3D features leads to sparse intra-class distributions and unclear inter-class boundaries, while (b) incorporating geometric rectification significantly enhances the compactness and discriminability of the feature space.

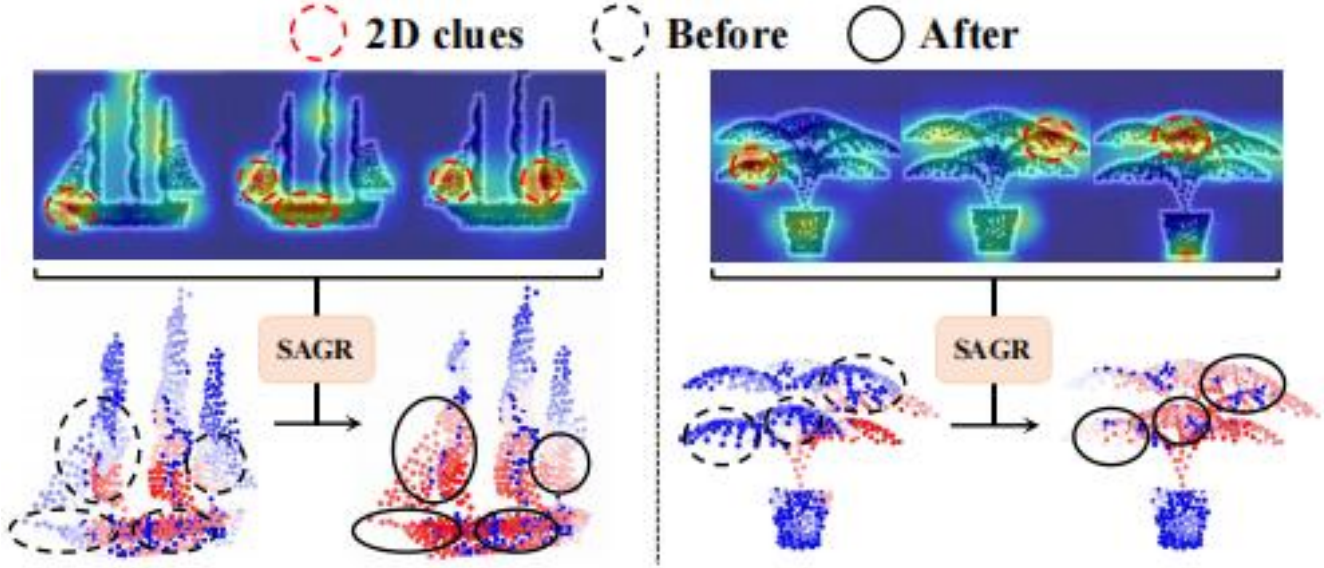


Figure 6. Point cloud attentions before and after geometric rectification, the red regions have higher attention weights compared to the blue regions.

Thank You !