Cross-Modal Prior-Guided Training with Visual Foundation Models for Unsupervised LiDAR Point Cloud Registration

Kezheng Xiong[1*], Shiyun Xu[1*], Sheng Ao[1], Mingming Wang[2,3], Siqi Shen[1], Chenglu Wen[1†], Cheng Wang[1]

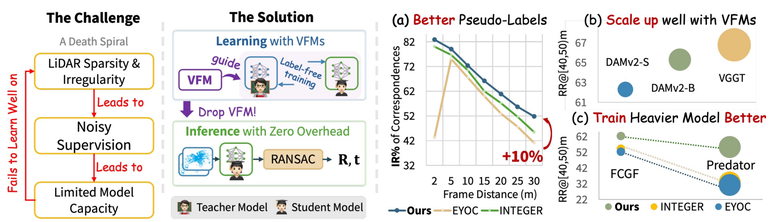
[1] Xiamen University, China [2] Tsinghua University, China [3] Guangzhou Automobile Group Co., Ltd., China

* Equal contribution † Corresponding author Contact: xiongkezheng@stu.xmu.edu.cn

Introduction & Motivation

Breaking the unsupervised "death spiral"

Sparse, irregular LiDAR geometry produces noisy pseudo-labels. The resulting supervision restricts the trainability of larger registration models.



PROMOTER uses visual priors to improve supervision and scale to stronger VFMs and registration models.

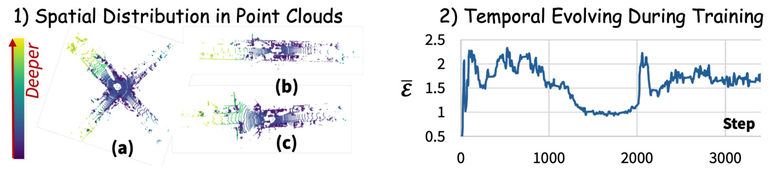
Challenges in transferring visual priors

Partial coverage. The 32-beam nuScenes LiDAR covers ≤1% of image pixels, leaving few point–pixel assignments.

Ambiguous projection. Many-to-one mappings, occlusions and calibration errors complicate visual–geometric alignment.

Modality and task gaps. Appearance and irrelevant semantics can distract from the geometry needed for registration.

Visual guidance during training. LiDAR-only registration at inference.



AdaPI adapts expert routing across points and training stages (Fig. 5).

Efficiency on KITTI (Table 2)

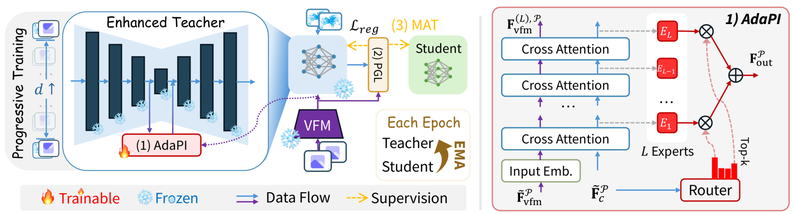
Methods	mRR%	RR@$d \in [40, 50)$	Training (h)	Testing (s/pair)
EYOC	83.2	52.3	**44.9**	
INTEGER*	84.0	54.2	76.0	∼0.16
PROMOTER	**86.5**	**63.3**	60.6	
Δ% vs. ⋆	+2.5%	+9.1%	-20%	–

≈20% less training time vs. INTEGER

Same inference cost

Method

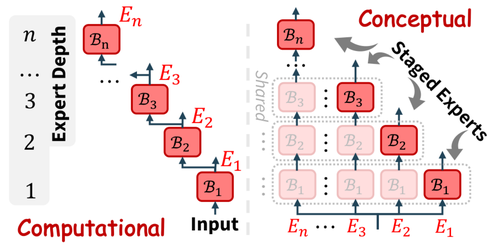
PROMOTER teacher–student framework



AdaPI enhances the teacher with frozen VFM priors. PGL produces correspondence pseudo-labels. MAT transfers the enhanced teacher's matchability to the student.

1. Adaptive Prior Integrator (AdaPI)

Geometry-aware embeddings fuse local 3D structure with VFM features. Cross-attention propagates visual priors to the full cloud. A top-k router mixes staged experts per point.



Deeper experts reuse shallower computations.

A zero-initialized residual projection introduces visual priors gradually.

2. Perceptual-Geometric Labeler (PGL)

Generate pose candidates from spatial compatibility. Rank them using visual consistency × geometric inlier count. Extract correspondences with the selected pose.

3. Matchability-Anchored Training (MAT)

Bidirectional cross-model contrastive learning transfers point relations. A Gram-matrix regularizer aligns global second-order feature structure.

$$\mathcal{L}_{\text{MAT}} = \left(\mathcal{L}_{\text{reg}}(\mathbf{F}_s^{\mathcal{P}}, \text{sg}(\mathbf{F}_t^{\mathcal{Q}}); \hat{\mathcal{C}}) + \mathcal{L}_{\text{reg}}(\text{sg}(\mathbf{F}_t^{\mathcal{P}}), \mathbf{F}_s^{\mathcal{Q}}; \hat{\mathcal{C}})\right) + \lambda_{\mathcal{R}} \mathcal{R}_{struct} \quad (5)$$

$$\mathcal{R}_{struct} = (\left\|\mathbf{G}_s^{\mathcal{P}} - \text{sg}(\mathbf{G}_t^{\mathcal{P}})\right\|_{\text{F}}^2 + \left\|\mathbf{G}_s^{\mathcal{Q}} - \text{sg}(\mathbf{G}_t^{\mathcal{Q}})\right\|_{\text{F}}^2)/2C^2 \quad (6)$$

Student objective: registration loss + weighted MAT loss

Experiments

KITTI and nuScenes. RR: RRE < 5° and RTE < 2 m.
mRR averages recall over five distance intervals.
U marks unsupervised methods (Table 1).

Dataset	Method	U	mRR	RR@$d \in$ $[5,10)$	$[10,20)$	$[20,30)$	$[30,40)$	$[40,50)$
KITTI	FCGF	–	77.4	98.4	95.3	86.8	69.7	36.9
	FCGF+C	–	84.6	100.0	97.5	90.1	79.1	56.3
	Predator	–	87.9	100.0	98.6	97.1	80.6	63.1
	SpinNet	–	39.1	99.1	82.5	13.7	0.0	0.0
	D3Feat	–	66.4	99.8	98.2	90.7	38.6	4.5
	CoFiNet	–	82.1	99.9	99.1	94.1	78.6	38.7
	GeoTrans.	–	42.2	100.0	93.9	16.6	0.7	0.0
	EYOC	✓	83.2	**99.5**	96.6	89.1	78.6	52.3
	RIENet	✓	50.7	96.3	72.1	38.2	24.4	22.6
	INTEGER	✓	84.0	**99.5**	97.1	89.6	79.6	54.2
	PROMOTER	✓	**86.5**	**99.5**	**97.5**	**90.6**	**81.6**	**63.3**
nuScenes	FCGF	–	39.5	87.9	63.9	23.6	11.8	10.2
	FCGF+C	–	59.3	96.2	85.1	59.6	35.8	20.0
	Predator	–	51.0	99.7	72.2	52.8	16.2	14.3
	EYOC	✓	61.7	96.7	85.6	61.8	37.5	26.9
	RIENet	✓	47.1	96.5	57.9	36.6	25.8	18.9
	INTEGER	✓	63.2	97.1	86.9	62.9	39.6	29.4
	PROMOTER	✓	**72.0**	**98.9**	**88.4**	**72.5**	**54.6**	**45.5**
KITTI ↓ nuScenes	EYOC	✓	55.3	96.2	75.6	58.7	26.6	19.7
	RIENet	✓	46.2	83.3	73.2	43.5	19.8	11.1
	INTEGER	✓	62.5	97.5	84.6	62.6	37.8	30.2
	PROMOTER	✓	**66.1**	**99.5**	**87.6**	**68.2**	**40.5**	**34.5**

nuScenes vs. INTEGER: +8.8 pp mRR and +16.1 pp RR at 40–50 m

Scaling with stronger priors and models

VFMs (Table 3)

VFM	#Params	mRR	$d \in [40,50)$ RR	RRE	RTE
► *VFMs for Visual-Geometric Tasks*					
a) DAMv2-S	24.8M	86.5	63.3	1.03	0.56
b) DAMv2-B	97.5M	87.2	65.8	0.98	0.55
c) VGGT	0.94B	**87.5**	**67.3**	**0.91**	**0.52**
► *Other Visual Model Baselines*					
d) DINOv2-B*	86.1M	86.2	62.8	1.17	0.55
e) EfficientNet-B7*	66.0M	*Failed to Converge*			

Registration models (Table 4)

Method	Reg. Model	#Params	mRR	$d \in [40,50)$ RR	RRE	RTE
PROMOTER	FCGF	8.76M	86.5	63.3	1.03	**0.56**
	D3Feat	14.79M	**87.3**	**66.3**	**0.95**	**0.56**
EYOC				61.1	31.2	1.81	0.65
INTEGER	Predator	22.77M	61.6	32.7	1.38	**0.59**
PROMOTER				**84.3**	**55.8**	**1.36**	0.61

Stronger VFMs improve recall. PROMOTER also improves Predator's mRR from 61.6 to 84.3 vs. INTEGER.

Consistent student features on KITTI (Fig. 4)

