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

Kezheng Xiong[ab*], **Shiyun Xu**[ab*], **Sheng Ao**[ab], **Mingming Wang**[cd], **Siqi Shen**[ab], **Chenglu Wen**[ab†], **Cheng Wang**[ab]
[a]Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University, Xiamen, China.
[b]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China.
[c]School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China.
[d]Guangzhou Automobile Group Co., Ltd. Guangzhou 511434, China.
{xiongkezheng,shiyunxu}@stu.xmu.edu.cn
wmm24@mails.tsinghua.edu.cn,
{aosh,siqishen,clwen,cwang}@xmu.edu.cn

Abstract

Visual Foundation Models (VFMs) have demonstrated strong capabilities in visual-geometric tasks such as image matching and 3D reconstruction. Recent work has revealed the potential of VFMs in label-free ***indoor RGB-D*** registration, where the visual modality is intrinsic to the task. A natural question follows: *can this success extend to more challenging **outdoor** scenarios characterized by sparse, irregular LiDAR data with only partial visual coverage?* We answer this with PROMOTER, a novel teacher–student framework for unsupervised LiDAR registration that bridges visual priors and sparse LiDAR geometry. To leverage pre-trained geometric priors for 3D registration, we must disentangle them from confounding signals introduced by modality and task gaps under partial visual coverage. To this end, we introduce a lightweight Adaptive Prior Integrator that extracts and integrates these priors into an enhanced teacher, together with a Perceptual-Geometric Labeler that produces high-quality pseudo-labels. For student training, we propose Matchability-Anchored Training to improve robustness under noisy supervision. Experiments on KITTI and nuScenes demonstrate state-of-the-art performance and improved scalability across registration models and VFMs, while incurring no additional parameters at inference. Code will be released.

1 Introduction

Point cloud registration is a fundamental task in robotics and autonomous driving [1, 2]. Supervised methods [3, 4] have advanced the field but require large datasets with accurate poses, which are costly and often impractical in complex outdoor scenarios due to inertial drift and GPS-denied conditions. Unsupervised methods [5, 6] thus emerge as a promising alternative, exploiting self-supervision from the intrinsic geometric cues in point clouds. However, real-world LiDAR data (*e.g.*, 32-beam low-resolution nuScenes [7]) are extremely sparse and irregular, which can induce the "death spiral" shown in Fig. 1 (Left): weak geometric heuristics produce noisy supervision [5], which in turn forces

*Equal contribution
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

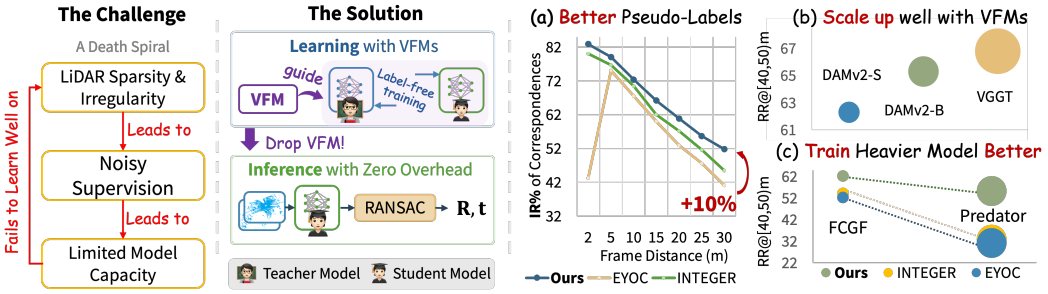


Figure 1: **(1) The Challenge.** Existing unsupervised methods can enter a "death spiral" that degrades performance. **(2) Our Solution.** PROMOTER addresses the challenge with even less training time (Sec. 5.3), introducing no additional parameters or data modalities at inference. **(3) Results.** PROMOTER leverages VFMs in unsupervised training, enabling: (a) state-of-the-art pseudo-label quality; (b) effective scaling with stronger VFMs in terms of parameters (DAMv2-S→DAMv2-B) and training recipes (DAMv2-B→VGGT); and (c) improved scalability across registration models.

reliance on low-capacity backbones (*e.g.*, FCGF) that are less prone to overfitting such noise [8]. This dependence on simple models hinders scaling to modern, high-capacity registration architectures and limits the ability to learn effectively from sparse LiDAR.

Can stronger priors break this spiral? Visual Foundation Models (VFMs) have recently transformed computer vision, including visual-geometric tasks such as image matching and 3D reconstruction [9–12]. Trained on vast 2D data, they acquire strong geometric priors with impressive generalization [13]. Motivated by this, recent work [14, 15] leverages VFMs in unsupervised *indoor RGB-D registration*, demonstrating the strong geometric reasoning potential of VFMs. This indoor success, however, benefits from the inherent visual modality in RGB-D data, whereas outdoor LiDAR provides sparse geometry with only partial camera overlap. Extending this success to outdoor LiDAR is therefore *non-trivial*. The restricted LiDAR-camera overlap yields only limited point-pixel assignments and causes information loss, with the 32-beam LiDAR in nuScenes covering $\leq$1% of image pixels. Moreover, 3D-2D projection introduces many-to-one ambiguity that can be amplified by calibration and synchronization errors. Finally, existing VFM-based methods still require VFMs and images *at inference time*, which is costly for resource-constrained deployments (*e.g.*, autonomous vehicles).

To address these challenges, we propose PROMOTER, a teacher–student framework to **PRO**be and ta**M**e ge**O**me**T**ric priors from VFMs for unsup**E**rvised point cloud **R**egistration. It achieves strong performance and scales with stronger VFMs, notably improving the trainability of heavier registration models (Fig. 1 (Right)). The geometric priors provided by VFMs can, however, become entangled with confounding factors (*e.g.*, dynamic objects, lighting, or task-irrelevant semantics) introduced by domain and task gaps, making naive use ineffective. We therefore introduce a lightweight **Ada**ptive **P**rior **I**ntegrator (AdaPI) that extracts effective priors from the VFM feature space and adaptively integrates them to enhance the teacher. Next, **P**erceptual-**G**eometric **L**abeler (PGL) produces high-fidelity correspondence pseudo-labels using VFMs as an external validation signal [9]. For student learning, we formulate **M**atchability-**A**nchored **T**raining (MAT) to stabilize training under noisy supervision guided by the VFM-enhanced teacher. Crucially, unlike prior work [14–16] that requires VFMs at test time, PROMOTER introduces *no additional data or parameters at inference*, preserving the deployment efficiency of the base registration model.

Extensive experiments on KITTI and nuScenes demonstrate state-of-the-art performance, with particularly notable gains on the challenging nuScenes benchmark. Our method scales across VFMs and improves the trainability of high-capacity registration models. Our contributions are as follows:

- To our knowledge, we introduce PROMOTER, the *first* approach to extend the power of VFMs to unsupervised outdoor LiDAR registration *without* sacrificing inference efficiency.
- We propose a train-only AdaPI that extracts effective priors from VFMs and adaptively integrates them into a teacher, along with a PGL that produces reliable pseudo-labels with VFM assistance.
- We formulate MAT for student training, enabling robust learning under noisy supervision.

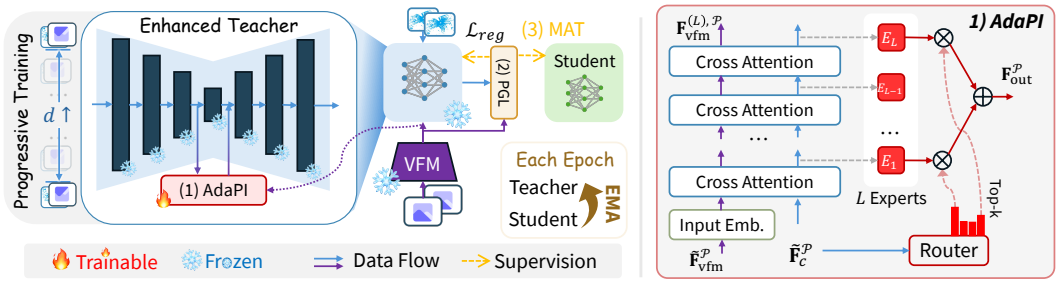


Figure 2: **The Overall Pipeline of PROMOTER.** Our (1) AdaPI first extracts reliable priors from VFMs using 3D geometry and then adaptively integrates them into the teacher. The (2) PGL then provides robust pseudo-labels for student training using the proposed (3) MAT objective.

2 Related Works

Unsupervised Registration. The high cost of acquiring large-scale datasets with precise pose labels has motivated unsupervised registration. Early methods were confined to indoor RGB-D data, using visual modalities [17] or differentiable rendering [18, 19]. Early point-cloud-only approaches employ Gaussian Mixture Models [20, 21], neighborhood consensus [22], or ICP as a reward proxy [23]; low-overlap scenarios and the irregularity of LiDAR point clouds, however, violate their design assumptions, leading to failures. For outdoor LiDAR data, teacher–student frameworks [24] have been proposed, enhanced with spatial filtering [5], and later improved with contextual information [6]. These methods nonetheless exhibit limited scalability to more complex registration architectures (*i.e.*, largely restricted to FCGF [25]), a key challenge that we address.

Registration with VFMs. VFMs [26, 27, 11, 9] have advanced various downstream 3D tasks [28, 29, 15, 14]. Chen and Qu [16] explored VFMs for *supervised* registration. Recent methods achieve label-free indoor RGB-D registration by augmenting handcrafted features [15] or detecting matchable regions [14] using VFMs. However, their success relies on dense point-to-pixel alignment and the availability of images and VFMs at test time, which is less compatible with the sparse LiDAR and partial camera overlap characteristic of outdoor settings.

3 Preliminaries

Problem Formulation. Given two point clouds $\mathcal{P} \in \mathbb{R}^{m\times 3}$ and $\mathcal{Q} \in \mathbb{R}^{n\times 3}$, point cloud registration aims to find the rigid transformation $\mathbf{T}=\{\mathbf{R},\mathbf{t}\}$ that optimally aligns $\mathcal{P}$ to $\mathcal{Q}$ with rotation $\mathbf{R}\in \mathrm{SO}(3)$ and translation $\mathbf{t}\in\mathbb{R}^3$. Accurate registration requires discriminative features for matching, which demands reliable supervision.

3D-2D Projection. VFMs for visual-geometric tasks (*e.g.*, VGGT [9], DepthAnything V2 [11]) produce pixel-wise features and dense geometric outputs (*e.g.*, depth maps, point maps) using a ViT [30] with DPT [31] heads. To assign pixel-space outputs to a 3D point $\mathbf{p}_i$, we project $\mathbf{p}_i$ onto image-plane coordinates (u_i, v_i) using camera intrinsics $\mathbf{K}$ and extrinsics $\mathbf{T}_{\text{cam}}$, and then apply bilinear interpolation to retrieve information from the surrounding pixels. Due to the sparsity of LiDAR data, only subsets $\mathcal{V}^{\mathcal{P}}\subset\mathcal{P}, \mathcal{V}^{\mathcal{Q}}\subset\mathcal{Q}$ of points fall within the image frustum, posing a challenge for effectively propagating and utilizing VFM-derived information across the full point cloud.

4 Methods

Overall Pipeline of PROMOTER. As shown in Fig. 2, for a point cloud pair $(\mathcal{P},\mathcal{Q})$ with corresponding images $(\mathcal{I}^{\mathcal{P}},\mathcal{I}^{\mathcal{Q}})$, AdaPI (Sec. 4.1) enhances the teacher by effectively integrating priors from VFMs. PGL (Sec. 4.2) then mines accurate pseudo-labels $\mathcal{C}$ with the VFM as an external validation signal. The student is trained using the MAT objective, and the teacher is updated via the exponential moving average (EMA) of the student, except during the initial warm-up phase (Sec. 4.3). We adopt *progressive training* [5, 6] to gradually adapt the student to distant point cloud pairs. During training, the frame interval $I\in\mathbb{N}^+, I\in[1, I_{\max}]$ is stochastically determined for each pair, with $I_{\max}$

increasing linearly from 1 to a predefined maximum I_{end} over the course of training. $I=1$ represents easy adjacent frames, while larger I corresponds to low-overlap (*e.g.*, $\leq$40%) and challenging pairs.

4.1 Adaptive Prior Integrator

AdaPI is a lightweight, *train-only* module that enhances the teacher for reliable pseudo-labeling. It is *discarded at inference*, introducing zero overhead. AdaPI's design addresses three key challenges that arise when enhancing the teacher with VFMs: **(C1)** Due to the limited FoV overlap between LiDAR and cameras, the number of visible points $|\mathcal{V}^{\mathcal{P}}|$ can be **small**, making it difficult to propagate visual priors effectively across the entire point cloud $\mathcal{P}$; **(C2)** VFM-derived features exhibit **varying reliability across points** owing to occlusions, domain gaps, and VFM-inherent uncertainties, hindering naïve integration; and **(C3)** The learning difficulty and matching ambiguity vary significantly across points, necessitating an **adaptive, per-point enhancement strategy**.

As illustrated in Fig. 2 (Right), AdaPI tackles **(C1)** by first extracting geometry-aware priors $\mathbf{F}_{\text{prior}}^{\mathcal{V}^{\mathcal{P}}}$ via *visual input embedding* ("Input Emb."). These priors then attentively interact with the teacher's features, propagating information to the rest of the point cloud. To address **(C2–C3)**, we introduce a *mixture of staged experts* that adaptively injects the extracted priors into the teacher through data-driven computation allocation. Across points, it dynamically routes computation so that straightforward regions rely on shallow experts while ambiguous points are allocated deeper, more expressive processing; this attenuates the influence of unreliable visual priors, reducing negative transfer. During training, as progressive training [6, 5] increases task difficulty, AdaPI helps stabilize optimization by gradually shifting from shallow (fine-grained) to deeper (more abstract) processing [32, 33] on a per-point basis (see Sec. 5.3 for analysis). Coupled with MAT, this yields a data-driven curriculum that progressively adjusts the knowledge transferred to the student.

Plugging into the Teacher Model. AdaPI is highly versatile and can be plugged into modern feature-pyramid-based registration models [34, 35]. As shown in Fig. 2 (Left), it is inserted at the bottleneck to enhance the coarsest latent features $(\mathbf{F}_c^{\mathcal{P}}, \mathbf{F}_c^{\mathcal{Q}})$ before standard decoding. This allows refined information to propagate to finer, higher-resolution levels, improving the selectivity of integration while leveraging the teacher's inherent robustness to reduce the risk of feature collapse.

4.1.1 Visual Input Embedding

Extracting effective visual embeddings from VFM features for AdaPI is non-trivial. VFM features originate from dense, textured 2D images, whereas registration operates on sparse, irregular, geometry-only 3D point clouds. This domain gap can introduce confounding signals such as dynamic objects, lighting variations, or task-irrelevant semantics, and the one-to-many ambiguity from 3D-2D projection can further hinder naïve feature utilization.

To address this, we first extract VFM features $\mathbf{F}_{\text{VFM}}^{\mathcal{P}}$ for the visible points $\mathcal{V}^{\mathcal{P}}$ via 3D-2D projection. Originating from the visual domain, these features lack the explicit local geometric information crucial for registration [36, 37], necessitating geometric refinement to compute the final embeddings:

$$\mathbf{F}_{\text{prior}}^{\mathcal{V}^{\mathcal{P}}} = h\left(f(\mathcal{V}^{\mathcal{P}}) \odot g(\mathbf{F}_{\text{VFM}}^{\mathcal{P}})\right) \tag{1}$$

where $f(\cdot)$ is a lightweight graph convolution block [38] that efficiently aggregates local geometry, and $g(\cdot)$ projects $\mathbf{F}_{\text{VFM}}^{\mathcal{P}}$ into the same dimensionality. These are fused via Hadamard product $\odot$, and the mapping $h(\cdot)$ (an MLP) produces the final embedding. This fusion is expressive yet efficient because the Hadamard product implicitly maps inputs to a high-dimensional space [39], modeling complex, non-additive interactions without explicitly doubling the feature dimensionality [39]. The same process is applied to $\mathcal{Q}$. The resulting embeddings retain rich VFM priors while encoding 3D geometry. Through dense interactions between $\mathbf{F}_c^{\mathcal{P}}$ and $\mathbf{F}_{\text{prior}}^{\mathcal{V}^{\mathcal{P}}}$, the subsequent network can suppress visual distractors and amplify relevant geometric signals for registration [37, 40, 36].

4.1.2 Mixture of Staged Experts

The first m of AdaPI's L_e cross-attention blocks [30, 41] naturally form the m-th staged expert E_m, whose output $\mathbf{F}_e^{(m)}$ represents the teacher features enhanced by m levels of processing. As shown in Fig. 3, unlike conventional parallel [42, 43] or recursive [44] experts, routing among staged experts encourages *depthwise specialization*: deeper experts build upon shallower ones, capturing

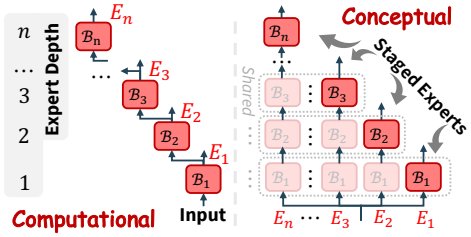


Figure 3: **Computational and Conceptual View of Staged Experts.** Our staged design enables deeper experts to inherit computations from shallower ones. This implicit weight-sharing ensures efficiency while yielding a clear abstraction hierarchy.

Algorithm 1 Perceptual-Geometric Labeler (PGL)

Input: $\mathcal{P}$, $\mathcal{Q}$; Initial matching $\tilde{\mathcal{C}}$; VFM features.
Output: Purified pseudo-labels $\mathcal{C}$.
Generate K candidates $\{\mathbf{T}_k\}_{k=1}^{K}$ utilizing spatial compatibility [46–48].
for each $\mathbf{T}_k$ **do**
  Warp $\mathbf{p}' = \mathbf{T}_k\mathbf{p}$ and retrieve their VFM features $\mathbf{f}_{\text{VFM}}^{\mathbf{p}'_i,k}$.
  Compute s_k^{p} on mutually visible points.
  Rank $\mathbf{T}_k$ by $s_k^{\text{p}} \cdot \text{IC}(\mathbf{T}_k)$.
end for
Select $\mathbf{T}^* = \arg\max_{\mathbf{T}_k} s_k^{\text{p}} \cdot \text{IC}(\mathbf{T}_k)$.
Extract final pseudo-labels $\mathcal{C}$ via NN search.

increasingly complex relationships and extracting more abstract features [32, 33]. This data-driven flexibility is critical for adaptive integration. It is also efficient by construction, since intermediate computations from early blocks are inherently reused by deeper experts.

The lightweight MLP router $\mathcal{R}$ then predicts routing weights $\mathbf{W}$ using $\mathbf{F}_c^{\mathcal{P}}$, retaining only the top-k_e experts per point. We then aggregate expert outputs to derive enhanced teacher features $\hat{\mathbf{F}}_c^{\mathcal{P}}$:

$$\mathbf{F}_{\text{aggr}} = \sum_{i=1}^{L_e} \left(\mathbf{W}_{:,i} \odot \mathbf{F}_e^{(i)} \right), \quad \hat{\mathbf{F}}_c^{\mathcal{P}} = \mathbf{F}_c^{\mathcal{P}} + \alpha \mathbf{W}_0 \mathbf{F}_{\text{aggr}}, \tag{2}$$

where $\odot$ denotes element-wise product. The zero-initialized projection $\mathbf{W}_0$ provides an initial identity mapping to gradually introduce information from VFMs without destabilizing training, and the scaling factor α [45] controls the total enhancing strength. The resultant features are subsequently passed to the teacher's decoder for standard multi-scale processing (Fig. 2 (Left)). The same is applied to $\mathcal{Q}$, and the updating of AdaPI is detailed in Sec. 4.3. In addition, we define $\mathcal{E}$ to measure the mean processing depth allocated to a point using $\mathbf{W}$. This metric provides insight into the learning dynamics (see Sec. 5.3):

$$\mathcal{E} = \mathbf{W}\,[1, \ldots, L_e]^\top, \quad \bar{\mathcal{E}} = \left(\textstyle\sum_{p_i \in \mathcal{P}} \mathcal{E}_i\right)/|\mathcal{P}| \tag{3}$$

4.2 Perceptual-Geometric Labeler

Existing unsupervised methods rely exclusively on current teacher features and simple geometric heuristics for pseudo-labeling [24, 6]. This self-referential cycle often suffers from confirmation bias, severely restricting the trainability of high-capacity models. By contrast, PGL aims to break this cycle by leveraging VFMs as an independent external validation signal to produce reliable pseudo-labels, as summarized in Alg. 1. Following standard practices [5, 6], we first apply a spatial filter [5] on the initial matches to obtain a reliable subset $\tilde{\mathcal{C}}$. Because VFMs possess strong cross-view consistency [49], true geometric matches should also exhibit high *perceptual consistency*. That is, if $\mathcal{P}$ and $\mathcal{Q}$ are correctly aligned, the visual features retrieved at the projected 2D locations of the transformed points should be strongly correlated. To quantify this for each transformation candidate, we define a simple yet effective *perceptual consistency score* s_k^p evaluated over the set $\mathcal{V}_k = \{p_i \in \mathcal{V}^P \mid \mathbf{T}_k p_i \text{ is visible in view } \mathcal{Q}\}$:

$$s_k^{\text{p}} = \tfrac{1}{|\mathcal{V}_k|} \textstyle\sum_{\mathbf{p}_i \in \mathcal{V}_k} \left(1 + \cos\left(\mathbf{F}_{\text{VFM}}^{\mathbf{p}_i}, \mathbf{F}_{\text{VFM}}^{\mathbf{p}'_i,k}\right)\right)/2 \tag{4}$$

This visual metric s_k^p provides a complementary signal to the traditional Inlier Count (IC) criterion commonly used in outlier rejection [46–48], which relies solely on geometry. By ranking transformation candidates using the joint metric $s_k^{\text{p}} \cdot \text{IC}(\mathbf{T}_k)$, we select $\mathbf{T}^*$ based on both geometric and visual evidence, exploiting the complementary strengths of the two modalities. Finally, following Liu et al. [5], we extract the final purified pseudo-labels $\mathcal{C}$ through a nearest-neighbor (NN) search between the transformed point cloud $\mathcal{P}' = \mathbf{T}^*\mathcal{P}$ and $\mathcal{Q}$.

4.3 Initialization and Training Objectives

Teacher Initialization and Updating. Teacher–student training requires the teacher (including AdaPI) to provide sufficiently reliable supervision from the start, making initialization critical. Prior work initializes the teacher using synthetic data [6] or identity transformations as ground-truth approximations [5]; these can introduce domain gaps and may encourage early memorization of noise. In contrast, we find that at the first ($I = 1$) stage (*i.e.*, consecutive LiDAR frames) of progressive training, PGL can already mine high-fidelity pseudo-labels by effectively leveraging VFM priors, even when the teacher is untrained. We thus warm up by optimizing the teacher and AdaPI without updating the student during this stage. After warm-up, the teacher is updated only via the student's EMA, while AdaPI remains trainable to adapt to the evolving training process.

During training, we update AdaPI with the *native* registration task loss of the underlying registration model, denoted $\mathcal{L}_{\text{reg}}$. This preserves PROMOTER's versatility across diverse registration models. $\mathcal{L}_{\text{reg}}$ of recent registration models is typically a contrastive loss defined on $\mathcal{C}$ (*e.g.*, Hardest Contrastive Loss [25], Circle Loss [34, 4]). In addition, we apply a load-balancing loss [42] $\mathcal{L}_{\text{load}}$ to encourage utilization of staged experts. The overall teacher objective is thus $\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{load}}\mathcal{L}_{\text{load}}$

Matchability-Anchored Training for Student. Existing unsupervised methods either overlook feature-space knowledge transfer [5] or rely on a limited set of anchors [6]. Such designs discard the rich structure in the VFM-enhanced teacher's feature space and are brittle under noisy pseudo-labels. To address this, we introduce MAT to directly guide the student in the feature space. Intuitively, registration depends on point *relations* (*i.e.*, matchability), and the enhanced teacher provides a reliable reference due to its improved matching capability (Fig. 1). Park et al. [50] pioneered the transfer of *relational* knowledge with dense pairwise similarities, but the $\mathcal{O}(N^2)$ cost is intractable for large-scale outdoor point clouds. To overcome this, we introduce an efficient surrogate that replaces dense relational alignment with a sparse objective plus a global structural regularizer $\mathcal{R}_{struct}$, transferring inter-point relations while maintaining computational efficiency:

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

Here, $\hat{\mathcal{C}} \subset \mathcal{C}$ is a high-confidence subset obtained by mutual NN filtering in the enhanced teacher's feature space on the pseudo-labels $\mathcal{C}$ from PGL, and $\text{sg}(\cdot)$ denotes stop-gradient. We formulate this bidirectionally by reusing the model's native registration loss $\mathcal{L}_{\text{reg}}$ with cross-model contrastive targets. This design preserves architectural flexibility across diverse registration models (see Sec. 5.3). By forcing the student's features $\mathbf{F}_s^{\mathcal{P}}$ to match the enhanced teacher's features $\text{sg}(\mathbf{F}_t^{\mathcal{Q}})$ (and vice versa), the first term transfers point matchability.

While highly efficient, this term confines relations only to *corresponding* points, neglecting the broader topology of the teacher's feature space and leaving the student blind to relations among non-corresponding points. To compensate, $\mathcal{R}_{struct}$ treats the enhanced teacher as a reliable *structural anchor*, aligning the latent space structure using second-order statistics via Gram matrices [51]:

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

where $\mathbf{G}_s = (\mathbf{F}_s^{\mathcal{P}})^\top \mathbf{F}_s^{\mathcal{P}}/N \in \mathbb{R}^{C\times C}$ is the normalized student Gram matrix computed from $\mathbf{F}_s^{\mathcal{P}} \in \mathbb{R}^{N\times C}$; other terms are computed similarly. In theory, minimizing Eq. (6) is equivalent to minimizing the Maximum Mean Discrepancy with a second-order polynomial kernel [52], which is effective for aligning the global second-order statistical distributions of the two latent spaces. With $\lambda_{\mathcal{R}}$ as the weight for $\mathcal{R}_{struct}$ and λ_{MAT} as the balance factor, the complete student training objective is therefore:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{MAT}}\mathcal{L}_{\text{MAT}}. \tag{7}$$

5 Experiments

We evaluate PROMOTER on two challenging datasets, KITTI [53] and nuScenes [7], following their official splits. The evaluation protocol follows the standard settings of existing unsupervised work for outdoor LiDAR scenarios [5, 6]. Implementation details, hyperparameters, and experimental setups are provided in App. Sec. C.

Image Usage. KITTI provides only front-facing cameras, while nuScenes has multiple camera views. For fair comparison and robustness validation, we use images from the left rectified color camera

Table 1: **Comparisons with State-of-the-Art Methods.** "✓" in the column "U" denotes the methods are *Unsupervised*. Otherwise, they are supervised. The best *unsupervised* results are highlighted in bold. "KITTI→nuScenes" denotes generalizability results from KITTI to nuScenes.

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**

for KITTI and the front camera for nuScenes. Additional cameras could improve performance with richer priors but introduce computational overhead, which we leave to future work (see App. Sec. B).

Metrics. Following [35, 4, 34], we report *Relative Rotation Error* (RRE) and *Relative Translation Error* (RTE) on successful pairs [34, 35, 4]. The criteria for *Registration Recall* (RR) are $\mathrm{RRE} < 5^\circ \wedge \mathrm{RTE} < 2\mathrm{m}$. For challenging distant scenarios, following Liu et al. [5], Xiong et al. [6], we report $\mathrm{RR@}\,[d_1, d_2)$ for pairs with distance $d \in [d_1, d_2)$ and *mean RR* (mRR), averaged across all $[d_1, d_2)$ intervals. Following Liu et al. [5], Xiong et al. [6], pseudo-label quality is measured by the teacher's *Inlier Ratio* (IR) at the first teacher–student training epoch (tIR@1st Epoch).

Baselines. Following existing outdoor unsupervised methods [5, 6], we compare with supervised methods FCGF [25], Predator [34], SpinNet [54], D3Feat [55], CoFiNet [56], and Geometric Transformer (GeoTrans.) [35]. We also report FCGF+C, an FCGF trained with progressive training as in Liu et al. [5]. For unsupervised methods, we compare with RIENet [22], EYOC [5], and INTEGER [6]. We use the FCGF model and RANSAC for pose estimation unless otherwise specified, enabling fair comparison.

5.1 Main Results

Overall Performance. Quantitative results are shown in Tab. 1. PROMOTER achieves state-of-the-art (SoTA) performance on both datasets. Improvements are particularly notable in challenging distant scenarios on the nuScenes dataset, demonstrating PROMOTER's ability to handle sparse, irregular, low-resolution (32-beam only) LiDAR data. We achieve this with a relatively *small* VFM (DAMv2-S) and registration model (FCGF), showing high efficiency. In Fig. 4, we provide qualitative results on KITTI. Point colors are derived from t-SNE [57] of extracted features. PROMOTER's effective use of VFMs yields student features with clear cross-frame consistency, leading to accurate registration. Additional qualitative results are provided in App. Sec. H.

Table 2: **Wall-Clock Efficiency Comparison.** Training/testing times on KITTI. We show the difference ($\Delta\%$) with recent state-of-the-art INTEGER (marked '$\star$') [6].

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	$\sim$0.16
PROMOTER	**86.5**	**63.3**	60.6	
$\Delta\%$ vs. $\star$	+2.5%	+9.1%	-20%	–

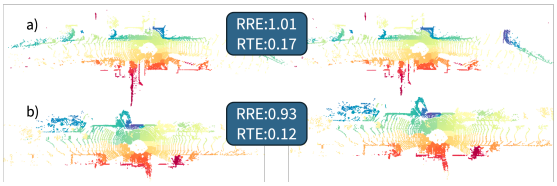


Figure 4: **Qualitative Results on KITTI.** We provide qualitative results on KITTI. Following Choy et al. [25], colors are derived from t-SNE [57] of extracted features.

Table 3: **Scaling with VFMs.** All results use FCGF. "*" means that feature maps are interpolated to match the input image resolutions.

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*			

Table 4: **Adaptability Across Registration (Reg.) Models.** PROMOTER improves trainability of high-capacity Predator model, where existing unsupervised methods struggle.

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	Predator	22.77M	61.1	31.2	1.81	0.65
INTEGER			61.6	32.7	1.38	**0.59**
PROMOTER			**84.3**	**55.8**	**1.36**	0.61

Generalization Across Datasets. Cross-dataset generalization is a critical challenge for learning-based registration [4, 5, 34, 54, 58, 59]. When trained on KITTI and tested on nuScenes, PROMOTER outperforms existing methods by clear margins, especially in distant scenarios (*e.g.*, +4.3% in RR@$[40, 50)$). This demonstrates improved robustness to domain gaps arising from different LiDAR resolutions (KITTI's 64 beams vs. nuScenes's 32 beams) and scene characteristics, indicating that PROMOTER successfully transfers generalizable knowledge from VFMs [26] to the student.

Improved Computational Efficiency. Despite leveraging VFMs, PROMOTER *reduces training time* relative to the prior SoTA INTEGER [6], which relies on time-consuming iterative label mining and per-batch full-teacher adaptation. Their learning process is less stable than ours and requires more training epochs. Our inference efficiency matches the original FCGF [25], since VFMs and AdaPI are used only during training and discarded at inference.

5.2 Performance Scaling and Adaptability

Scaling with VFMs. In Tab. 3, PROMOTER scales effectively with increasingly powerful VFMs. From DAMv2-S (a) to the larger DAMv2-B (b) and the more extensively trained VGGT (c), performance improves consistently. VGGT's advantage over DAMv2-B likely stems from its extensive multi-task training recipe, which includes a point tracking task that encourages cross-frame feature consistency. We also test DINOv2 (d), a *general-purpose* VFM, and EfficientNet-B7 (e), a conventional CNN *with comparable parameters*. Using EfficientNet leads to convergence failure, whereas DINOv2 achieves reasonable performance, slightly below that of visual-geometric VFMs. DINOv2's capability for matching and depth estimation via linear probing [26] suggests it partially captures geometric priors despite not being specifically designed for visual-geometric tasks. EfficientNet likely lacks such priors due to its limited training scale, causing convergence failure. See App. Sec. G for further discussion.

Adaptability Across Registration Models. In Tab. 4, PROMOTER adapts well to various registration architectures, surpassing existing unsupervised methods. We evaluate the fully-convolutional FCGF, the joint detection-description D3Feat, and the attention-based Predator. Performance improves with the larger D3Feat model. Training the sophisticated, high-capacity Predator (2.6$\times$ FCGF's parameters) is particularly challenging for unsupervised methods. While prior methods struggle, PROMOTER demonstrates *improved* trainability on heavy models, with moderate performance drops

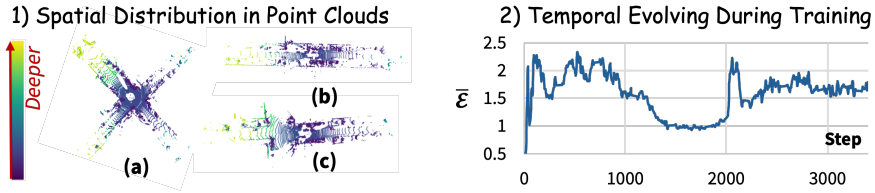


Figure 5: **Expert Routing Behaviors.** *Left*: spatial distribution of routing depths on KITTI; *Right*: AdaPI's mean effective depths $\bar{\mathcal{E}}$ per step in the early training stages of progressive training.

relative to FCGF, while substantially closing the gap to the supervised Predator baseline (Tab. 1). This opens avenues for future work to close the remaining gap.

5.3 Analysis of AdaPI's Expert Routing Behavior

Architectural dynamism is a key feature of AdaPI. Fig. 5 (Right) shows AdaPI's mean effective depth $\bar{\mathcal{E}}$ evolving per step (Eq. (3)) during the first two stages of progressive training with $k_e = 2$. Across stages, the router shifts toward deeper depths (more interactions) when supervision becomes harder (*i.e.*, larger d) and then gradually returns to shallower processing (retaining more original details) as the teacher stabilizes, reflecting a training-time curriculum that regulates prior injection for stable learning. Fig. 5 (Left) shows the spatial distribution of routing depths across representative KITTI scenes. Distinctive structures (*e.g.*, buildings, trees) route to shallow experts, indicating that salient regions need minimal refinement, whereas sparse or planar regions route to deeper experts to resolve ambiguities. We also observe deeper routing patterns for camera-visible points with degraded cues (occlusion, motion blur, lighting) and for points outside the camera FoV, where multiple interactions are needed for information propagation. These patterns jointly support our design of adaptive, data-driven per-point integration. As a result, dynamic routing outperforms static, hand-crafted depth schedules (Tab. 5) and stabilizes learning dynamics by effectively ensuring a discriminative feature space in early training. We further visualize this effect in App. Sec. H.1.

5.4 Ablation Study

We conduct extensive ablation studies on KITTI. In Tab. 5, replacing the *visual input embedding* with a direct MLP or using alternative fusion mechanisms degrades performance, indicating the necessity of local geometric aggregation. AdaPI's dynamic routing is crucial: removing it and routing to maximum depth in a static manner causes a large performance drop, while naïve hard-coded depth schedules underperform. In particular, the linearly-falling strategy, which contradicts the observation in Fig. 5, fails by a large margin, underscoring the importance of data-driven adaptability. PGL is essential for pseudo-label quality, substantially improving the teacher's initial pseudo-labels. The MAT objective is crucial for robust learning; all its variants degrade distant-range results.

Table 5: **Ablation Study on KITTI.**

Methods	tIR@1st Epoch	mRR	$d \in [40, 50)$ RR	RRE	RTE
w/o input emb. in AdaPI	80.6	85.4	60.8	1.08	0.59
Eq. (1):$\odot \rightarrow$addition	80.3	85.6	60.3	1.13	0.57
Eq. (1):$\odot \rightarrow$concat.	81.5	85.7	60.8	1.17	0.55
w/o routing in AdaPI	76.9	84.5	58.8	1.05	0.56
linearly-rising routing	81.1	85.5	61.3	1.03	0.61
linearly-falling routing	80.3	83.4	57.9	1.32	0.59
w/o MAT	82.8	83.9	58.3	1.16	0.56
MAT w/o $\mathcal{R}$	82.8	85.1	60.8	1.07	0.72
MAT w/ $\mathcal{R}$ only	82.8	85.3	60.3	1.03	0.61
w/o PGL	79.5	84.9	59.8	1.12	0.58
Full PROMOTER	82.8	86.5	63.3	1.03	0.56

6 Conclusion

In this paper, we presented PROMOTER, which extracts geometric priors from VFMs and tames them through a concise AdaPI to enhance the teacher. PGL generates high-quality pseudo-labels, and the student is trained robustly via the MAT objective. Experiments show state-of-the-art performance, improved scalability to heavier registration models and stronger VFMs, and high efficiency with no inference-time overhead. With its improved scaling and adaptability, PROMOTER highlights new

opportunities in the challenging unsupervised LiDAR registration. To our knowledge, it is the first unsupervised point cloud registration method to leverage VFMs for outdoor LiDAR data without sacrificing inference efficiency.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.42571514). This work was partially supported by the National Natural Science Foundation of China (No. 62501502)

References

[1] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the CVPR Conference*, pages 12786–12796, 2022.

[2] Haobo Jiang, Mathieu Salzmann, Zheng Dang, Jin Xie, and Jian Yang. Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems*, 36:21285–21297, 2023.

[3] Kezheng Xiong, Maoji Zheng, Qingshan Xu, Chenglu Wen, Siqi Shen, and Cheng Wang. Speal: Skeletal prior embedded attention learning for cross-source point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6279–6287, 2024.

[4] Zhenxuan Zeng, Qiao Wu, Xiyu Zhang, Lin Yuanbo Wu, Pei An, Jiaqi Yang, Ji Wang, and Peng Wang. Unlocking generalization power in lidar point cloud registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22244–22253, 2025.

[5] Quan Liu, Hongzi Zhu, Zhenxi Wang, Yunsong Zhou, Shan Chang, and Minyi Guo. Extend your own correspondences: Unsupervised distant point cloud registration by progressive distance extension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11366–11374, 2024.

[6] Kezheng Xiong, Haoen Xiang, Qingshan Xu, Chenglu Wen, Siqi Shen, Jonathan Li, and Cheng Wang. Mining and transferring feature-geometry coherence for unsupervised point cloud registration. *Advances in Neural Information Processing Systems*, 37:35468–35491, 2024.

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[8] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[10] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF CVPR Conference*, pages 20697–20709, 2024.

[11] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[13] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[14] Weijie Wang, Guofeng Mei, Bin Ren, Xiaoshui Huang, Fabio Poiesi, Luc Van Gool, Nicu Sebe, and Bruno Lepri. Zero-shot point cloud registration. *arXiv preprint arXiv:2312.03032*, 2023.

[15] Haobo Jiang, Jin Xie, Jian Yang, Liang Yu, and Jianmin Zheng. Zero-shot rgb-d point cloud registration with pre-trained large vision model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16943–16952, 2025.

[16] Congjia Chen and Yufu Qu. Dinoreg: Strong point cloud registration with vision foundation model. *arXiv preprint arXiv:2509.24370*, 2025.

[17] Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6433–6442, 2021.

[18] Mohamed El Banani, Luya Gao, and Justin Johnson. Unsupervisedr&r: Unsupervised point cloud registration via differentiable rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7129–7139, 2021.

[19] Mingzhi Yuan, Kexue Fu, Zhihao Li, Yucong Meng, and Manning Wang. Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17694–17705, 2023.

[20] Guofeng Mei, Hao Tang, Xiaoshui Huang, Weijie Wang, Juan Liu, Jian Zhang, Luc Van Gool, and Qiang Wu. Unsupervised deep probabilistic approach for partial point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13611–13620, 2023.

[21] Xiaoshui Huang, Sheng Li, Yifan Zuo, Yuming Fang, Jian Zhang, and Xiaowei Zhao. Unsupervised point cloud registration by learning unified gaussian mixture models. *IEEE Robotics and Automation Letters*, 7(3):7028–7035, 2022.

[22] Yaqi Shen, Le Hui, Haobo Jiang, Jin Xie, and Jian Yang. Reliable inlier evaluation for unsupervised point cloud registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2198–2206, 2022.

[23] Haobo Jiang, Yaqi Shen, Jin Xie, Jun Li, Jianjun Qian, and Jian Yang. Sampling network guided cross-entropy method for unsupervised point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6128–6137, 2021.

[24] Heng Yang, Wei Dong, Luca Carlone, and Vladlen Koltun. Self-supervised geometric perception. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021.

[25] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8958–8966, 2019.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[27] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[28] Pufan Zou, Shijia Zhao, Weijie Huang, Qiming Xia, Chenglu Wen, Wei Li, and Cheng Wang. Adaco: Overcoming visual foundation model noise in 3d semantic segmentation via adaptive label correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11086–11094, 2025.

[29] Shijia Zhao, Qiming Xia, Xusheng Guo, Pufan Zou, Maoji Zheng, Hai Wu, Chenglu Wen, and Cheng Wang. Sp3d: Boosting sparsely-supervised 3d object detection via accurate cross-modal semantic prompts. In *Proceedings of the CVPR Conference*, pages 29374–29384, 2025.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[32] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

[33] Teresa Dorszewski, Lenka Tětková, Robert Jenssen, Lars Kai Hansen, and Kristoffer Knutsen Wickstrøm. From colors to classes: Emergence of concepts in vision transformers. In *World Conference on Explainable Artificial Intelligence*, pages 28–47. Springer, 2025.

[34] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4267–4276, 2021.

[35] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022.

[36] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15859–15869, 2021.

[37] Yepeng Liu, Wenpeng Lai, Zhou Zhao, Yuxuan Xiong, Jinchi Zhu, Jun Cheng, and Yongchao Xu. Liftfeat: 3d geometry-aware local feature matching. *arXiv preprint arXiv:2505.03422*, 2025.

[38] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[39] Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu. Rewrite the stars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5694–5703, 2024.

[40] Xinjiang Wang, Zeyu Liu, Yu Hu, Wei Xi, Wenxian Yu, and Danping Zou. Featurebooster: Boosting feature descriptors with a lightweight neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7630–7639, 2023.

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[42] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23 (120):1–39, 2022.

[43] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[44] Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, et al. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025.

[45] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

[46] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13221–13231, 2022.

[47] Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2023.

[48] Yifei Zhang, Hao Zhao, Hongyang Li, and Siheng Chen. Fastmac: Stochastic spectral sampling of correspondence graph. *arXiv preprint arXiv:2403.08770*, 2024.

[49] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025.

[50] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[51] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[52] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[53] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[54] Sheng Ao, Qingyong Hu, Bo Yang, Andrew Markham, and Yulan Guo. Spinnet: Learning a general surface descriptor for 3d point cloud registration. In *CVPR*, pages 11753–11762, 2021.

[55] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6359–6367, 2020.

[56] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021.

[57] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

[58] Sheng Ao, Yulan Guo, Qingyong Hu, Bo Yang, Andrew Markham, and Zengping Chen. You only train once: Learning general and distinctive 3d local descriptors. *TPAMI*, 45(3):3949–3967, 2022.

[59] Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *CVPR*, pages 1255–1264, 2023.

[60] Hao Xu, Shuaicheng Liu, Guangfu Wang, Guanghui Liu, and Bing Zeng. Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In *Proceedings of the IEEE/CVF ICCV*, pages 3132–3141, 2021.

[61] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7163–7172, 2019.

[62] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11366–11374, 2020.

[63] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 602–618, 2018.

[64] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5545–5554, 2019.

[65] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF CVPR*, pages 6677–6686, 2022.

[66] Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17702–17711, 2023.

[67] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2514–2523, 2020.

[68] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15994–16003, 2021.

[69] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[70] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, and Rongqi Gu. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16014–16023, 2021.

[71] Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Marc Pollefeys, and Hesheng Wang. Regformer: an efficient projection-aware transformer network for large-scale point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8451–8460, 2023.

[72] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[73] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.

[74] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[75] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024.



Figure 6: **Rationale for Single-View Image Usage on KITTI.** KITTI [53] provides only two front-facing cameras with substantial overlap in their field of view (FoV). We therefore use only the left camera's image per scan to avoid redundancy, since using both doubles the computational overhead while providing minimal additional information. Accordingly, we also use only one image for nuScenes [7] to enable fair comparison, quantifying the benefits of VFM-derived information on denser (64 beams in KITTI) and sparser (32 beams in nuScenes) point clouds.

The Appendix is organized as follows: Section A provides additional related works; Section B details the datasets and justifies our camera usage; Sections C and D provide implementation and experimental details; Section E offers deeper insights into PROMOTER's components; Section F describes the registration models we use; Section G analyzes why conventional CNNs fail (supporting our VFM hypothesis); Section H presents additional experiments; and Section I discusses limitations and future work.

A Additional Related Works

Learning-Based Registration. Learning-based methods typically follow two main paradigms. Direct regression methods [60–62] predict transformations from global features; they are efficient but often struggle with large-scale, low-overlap scenarios. Correspondence-based methods [25, 63, 64, 35, 65, 66, 3] first establish point correspondences before computing transformations, and have demonstrated strong performance.

Robust Pose Estimators. Robustly estimating transformations from noisy initial correspondences is another critical direction, with learning-based estimators [36, 67, 68] often surpassing traditional approaches [69]. Recently, non-parametric methods have achieved efficient inlier selection using spatial compatibility [46] and graph structures [47, 48].

B Dataset

We evaluate PROMOTER on two standard public[3] outdoor benchmarks widely adopted for outdoor LiDAR point cloud registration [70, 71, 4], especially those focusing on the unsupervised registration of outdoor LiDAR data [5, 6].

[3]The licence of KITTI Odometry is *Creative Commons Attribution-NonCommercial-ShareAlike 3.0*, and the licence of nuScenes is *Creative Commons Attribution-NonCommercial-ShareAlike 4.0*.

Table 6: **Field of View (FoV) Coverage Analysis on KITTI.** The percentage of LiDAR points covered by both cameras and by a single camera. We additionally report the maximum difference ($\Delta_{\text{max}}\%$) between using either camera only and using both cameras, which in theory represents the maximum potential information gain from using both cameras.

Seq.	Both (%)	Left Only (%)	Right Only (%)	$\Delta_{\text{max}}\%$
00	16.59	16.11	16.18	0.48
01	16.22	15.97	15.89	0.33
02	16.65	16.09	15.89	0.76
03	16.57	16.12	16.23	0.45
04	16.52	16.07	16.25	0.45
05	16.56	16.08	16.14	0.48
Avg.	16.52	16.07	16.10	0.49

- **KITTI Odometry** [53] provides 11 sequences with synchronized 64-beam LiDAR scans (Velodyne HDL-64E, 100 m range) and stereo cameras, captured in German urban/highway environments. We follow standard splits [70, 35, 4]: sequences 00–05 for training, 06–07 for validation, and 08–10 for testing.

- **nuScenes** [7] contains 1000 diverse driving scenes from Boston and Singapore (850 train/val, 150 test [70, 34, 4]), featuring 32-beam LiDAR (70 m range) and 6 surround-view cameras. Compared to KITTI, nuScenes is *more challenging*: sparser scans (32 vs. 64 beams), shorter range (70 m vs. 100 m), and greater scene diversity (complex intersections, varied lighting)—all stress-testing PROMOTER's ability to cope with scarcity of geometric information.

B.1 Rationale for Single-Camera Setup

We use a single forward-facing camera to extract information from VFMs: the left rectified camera ('image_2') for KITTI and the front-facing camera ('CAM_FRONT') for nuScenes. While multi-view images could provide richer information, we deliberately adopt a single-camera setup for three key reasons:

(i) **Minimal Redundancy on KITTI.** As Fig. 6 and Table 6 show, KITTI's stereo cameras exhibit massive overlap: $> 95\%$ of visible LiDAR points appear in *both* views, yielding mostly $< 0.6\%$ additional coverage ($\Delta_{\text{max}}\%$) from using both cameras, defined as:

$$\Delta_{\text{max}}\% = \text{Both} - \min(\text{Left Only}, \text{Right Only}) \tag{8}$$

This marginal gain does not justify doubling the computational overhead.

(ii) **Fair Cross-Dataset Comparison.** Using one camera on *both* datasets isolates the impact of LiDAR characteristics (64-beam KITTI vs. 32-beam nuScenes) from confounding effects of different camera coverage. This enables cleaner attribution of performance differences to PROMOTER's effectiveness on varying point densities, consistent with the view that improvements on nuScenes stem from better handling of sparse data.

(iii) **Challenging Stress-Test Condition.** A single-camera setup represents the *minimal* scenario for VFM-derived information. If PROMOTER succeeds here, it is likely to improve with richer input. This also mirrors practical deployments where cost constraints limit sensor configurations.

Importantly, PROMOTER is *compatible* with multi-view inputs; Section H demonstrates performance gains when incorporating additional cameras on nuScenes. PROMOTER addresses partial LiDAR-camera coverage through its architectural design. During training, AdaPI extracts information only for camera-visible points $\mathcal{V}^{\mathcal{P}} \subset \mathcal{P}$ via 3D-2D projection. It then propagates this sparse information to *all* points through feature interactions (Fig. 2), allowing uncovered regions to benefit from geometric priors via spatial context. This helps address the challenges described in Section 4 of the main paper. At inference, the student operates solely on point clouds, requiring *no* camera input and thus avoiding FoV limitations entirely.

C Implementation Details

We implement PROMOTER in PyTorch [72] and train on a *single* NVIDIA RTX 3090 GPU (24GB). Hardware: Intel® Xeon® 2186 CPU, 512GB RAM, 4× RTX 3090 GPUs. We use MinkowskiEngine [73] as in the official FCGF implementation to ensure fair comparison with prior unsupervised work [5, 6]. Despite utilizing large VFMs during training, the total GPU memory footprint (VFM + registration model + auxiliary modules) remains under 22GB on KITTI, making PROMOTER accessible with standard research hardware. All VFM-related modules are training-only and incur no inference overhead, matching the original registration model's deployment efficiency.

D Experimental Setup

D.1 Training and Hyperparameters

The optimization strategy differs between the proposed components and the student registration model due to inherent differences in their architectures and training objectives. To update the parameters of AdaPI, we use the AdamW [74] optimizer with an initial learning rate of 1×10^{-5} and weight decay of 1×10^{-3}, a standard choice for attention-based architectures such as AdaPI. Learning rate warm-up is applied for the first 5 epochs, after which cosine learning rate decay is applied for each progressive training stage between 1×10^{-5} and 1×10^{-6}. The student model's optimization, voxel size and data augmentation settings remain consistent with the original papers of each registration network, ensuring fair comparison. To ensure stability, we employ gradient accumulation and gradient clipping (max norm of 2.0). During teacher–student training, the teacher model is updated via Exponential Moving Average (EMA) with a decay rate of 0.4 per epoch, following existing practices [5, 6] to ensure fair comparison. We empirically set $\lambda_{\mathcal{R}} = 0.2$ and $\lambda_{\text{MAT}} = 0.005$ for the losses, respectively. The load-balancing loss weight is set following prior works [42]. We train for 210 epochs with a batch size of 4 for KITTI and 6 for nuScenes.

D.2 Evaluation Protocols

We adopt standard metrics from outdoor LiDAR registration efforts [35, 4, 34]: *Relative Rotation Error* (RRE) and *Relative Translation Error* (RTE) measure accuracy on successful pairs, while *Registration Recall* (RR) measures the success rate using thresholds $\mathrm{RRE} < 5^\circ \wedge \mathrm{RTE} < 2\mathrm{m}$.

For challenging distant scenarios [5, 6] where overlap is minimal, we follow recent unsupervised work [5, 6] and report distance-binned metrics: $\mathrm{RR@}\,[d_1, d_2)$ denotes RR for pairs with inter-frame distance $d \in [d_1, d_2)$, and *mean RR* (mRR) averages across the intervals $\{[5, 10), [10, 20), [20, 30), [30, 40), [40, 50)\}$. This distance-based analysis is critical for outdoor scenarios: registration difficulty escalates dramatically with distance due to reduced overlap, increased sparsity, and geometric ambiguity.

Pseudo-label quality is measured by the teacher's *Inlier Ratio* (IR) at the first teacher-student epoch (tIR@1$^{\text{st}}$ Epoch) [5, 6], directly reflecting supervision quality and correlating with final performance.

D.3 Metrics Definitions

Relative Rotation Error (RRE). RRE measures the geodesic distance in degrees between the ground-truth rotation matrix $\mathbf{R}_{gt}$ and the predicted rotation matrix $\mathbf{R}_p$:

$$\mathrm{RRE}(\mathbf{R}_p, \mathbf{R}_{gt}) = \arccos\left(\frac{\mathrm{trace}(\mathbf{R}_p^T \mathbf{R}_{gt}) - 1}{2}\right) \cdot \frac{180}{\pi}. \tag{9}$$

This provides a measure of angular deviation regardless of rotation axis.

Relative Translation Error (RTE). RTE measures the Euclidean distance in meters between the ground-truth translation vector $\mathbf{t}_{gt}$ and the predicted vector $\mathbf{t}_p$:

$$\mathrm{RTE}(\mathbf{t}_p, \mathbf{t}_{gt}) = \|\mathbf{t}_p - \mathbf{t}_{gt}\|_2. \tag{10}$$

Registration Recall (RR). RR measures the percentage of point cloud pairs where both RRE and RTE are simultaneously below specified thresholds (RRE $< 5^\circ$ and RTE $< 2\,\mathrm{m}$). For a set of M

point cloud pairs, this is computed as:

$$\text{RR} = \frac{1}{M}\sum_{i=1}^{M}[\![\text{RRE}_i < \tau_r \wedge \text{RTE}_i < \tau_t]\!], \tag{11}$$

where $\tau_r = 5^\circ$ and $\tau_t = 2\text{m}$ are the rotation and translation thresholds, respectively, and $[\![\cdot]\!]$ is the Iverson bracket (1 if true, 0 if false). Following common practice [34, 25], the mean RRE and RTE are computed only over successfully registered pairs.

Mean Registration Recall (mRR). To evaluate performance across different difficulty levels, we report recall over specific distance intervals, denoted as $\text{RR}@[d_1, d_2)$. The mRR is the average recall across the set of intervals $\mathcal{D} = \{[5, 10), [10, 20), [20, 30), [30, 40), [40, 50)\}$:

$$\text{mRR} = \frac{1}{|\mathcal{D}|}\sum_{[d_1,d_2)\in\mathcal{D}} \text{RR}@[d_1, d_2). \tag{12}$$

This metric provides a comprehensive assessment of registration performance across varying difficulty levels, from easy nearby pairs to challenging distant pairs with minimal overlap.

Inlier Ratio (IR). To measure the quality of generated pseudo-labels, we report the Inlier Ratio of the teacher model's correspondences at the beginning of teacher-student training, denoted as "tIR@1$^{\text{st}}$ Epoch". This is computed as the percentage of correspondences that satisfy the spatial compatibility criterion (*i.e.*, the ground-truth transformation aligns the correspondence within a distance threshold; we set the threshold as 0.3 for KITTI following Xiong et al. [6], Liu et al. [5]). Higher IR values indicate more reliable pseudo-labels, which lead to better student model learning.

E Details on Design Choices on PROMOTER

E.1 AdaPI

Ablations (main paper Table 5) support our design choices: removing dynamic routing causes the *largest* single-component performance drop of pseudo-label quality; simpler fixed-depth schedules (*e.g.*, the linearly-falling strategy, which contradicts the behavior observed in Fig. 5) underperform. The staged expert formulation creates a clear abstraction hierarchy in which deeper experts refine the outputs of shallower ones, enabling specialized cascades that filter confounding signals while preserving geometric cues. Without this hierarchical structure, the model can struggle to separate reliable information from noise, increasing the risk of feature-space collapse (Fig. 8).

Detailed Architecture. The AdaPI module comprises a stack of L_e blocks forming staged experts and a lightweight router.

- **Expert Blocks:** Each block follows standard Transformer [31] design and contains a cross-attention layer, enabling global interaction between the point features and the information, as depicted in Fig. 3 of the main paper. We apply standard positional encodings [30] to the point features to enhance spatial awareness. The cross-attention allows the model to learn complex relationships between the visual and geometric information, which is crucial for effective fusion and adaptation to the registration task.
- **Router:** The router $\mathcal{R}$ is implemented as a multi-layer perceptron (MLP). It takes a point's original feature as input and produces a softmax distribution over the L_e experts, which is used to select and weight the top-k_e experts for that point.

Losses. During teacher–student training, the AdaPI module is trained with the original registration loss (*e.g.*, Hardest Contrastive Loss [25] or Circle Loss [55, 34], depending on the registration network) and a load-balancing loss [42]. The overall loss is defined as:

$$\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{load}} \cdot \mathcal{L}_{\text{load}}. \tag{13}$$

We use the pseudo-labels mined by PGL for the registration loss $\mathcal{L}_{\text{reg}}$, and λ_{load} is a hyperparameter that balances the two losses. In our experiments, we follow Fedus et al. [42] to set the hyperparameters for this load-balancing loss.

Table 7: Summary of registration models used in our experiments.

Model	Architectural Design Paradigm	#Param. (M)
FCGF[25]	Fully Convolutional	8.76
D3Feat[55]	Joint Detection & Description	14.79
Predator[34]	Attention-based	22.77

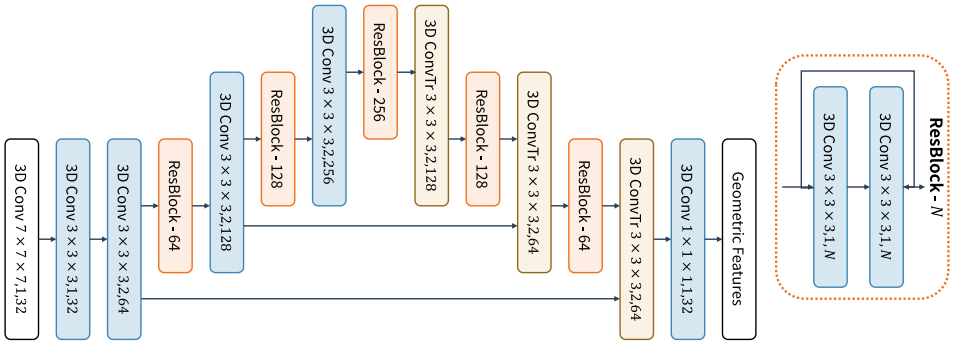


Figure 7: **The Architecture of FCGF [25].** The network uses a symmetric Res-UNet encoder-decoder structure with sparse convolutions and skip connections to generate per-point feature descriptors.

F Involved Registration Networks

To assess the scalability and versatility of PROMOTER, we extensively evaluate it on three representative deep learning models covering various model sizes and architectural paradigms: the fully-convolutional FCGF [25], the attention-based Predator [34], and the joint detection-description D3Feat [55]. Table 7 summarizes these models. We provide brief architectural descriptions below; please refer to the original papers for complete details.

F.1 Detailed Network Architectures

FCGF. The primary registration network used in our experiments is FCGF [25], a widely-adopted baseline in unsupervised point cloud registration [5, 6]. As depicted in Figure 7, its architecture is based on a Res-UNet structure implemented using sparse convolutions to efficiently process sparse 3D data. The network features a symmetric encoder-decoder design with three levels of skip connections that preserve fine-grained geometric details across scales. The encoder progressively downsamples the input through strided convolutions while increasing the feature dimension, and the decoder upsamples features back to the original resolution. After the final convolutional layer, the output features are L2-normalized to produce unit-length descriptors suitable for correspondence matching via inner product or cosine similarity.

D3Feat. D3Feat [55] introduces a joint learning framework that simultaneously performs dense feature detection and description within a single network. The architecture is based on a convolutional UNet-like structure with skip connections that operates directly on unstructured point clouds. A key innovation of D3Feat is its dual-role design: a single network outputs a dense feature map that serves both to derive L2-normalized feature descriptors and to compute keypoint detection scores.

The detection score for each point is computed as $s_i = \max_k(\alpha_i^k \beta_i^k)$, combining a density-invariant saliency score (α_i^k) and a channel-wise max score (β_i^k):

$$\alpha_i^k = \ln(1 + \exp(D_i^k - \frac{1}{|N_{x_i}|} \sum_{x_j \in N_{x_i}} D_j^k)), \quad \beta_i^k = \frac{D_i^k}{\max_t(D_i^t)}, \tag{14}$$

where D_i^k is the k-th channel of the dense feature at point i, and N_{x_i} denotes the neighborhood of point x_i. This formulation makes the detector robust to varying point densities commonly found in outdoor LiDAR data. Joint optimization is enabled by a self-supervised *detector loss*, which is guided by on-the-fly feature matching results during training, alongside a standard contrastive loss for the descriptors. Please refer to the original paper [55] for complete details.

Predator. Predator [34] is a sophisticated registration model specifically designed for low-overlap point cloud pairs, which are common in outdoor scenarios. Its architecture employs a two-stream encoder-decoder network, distinguished by a novel *overlap-attention module* in the bottleneck that enables early co-contextual information exchange between the latent superpoint representations of the two input point clouds.

After a standard convolutional backbone extracts hierarchical features, the model employs a series of graph neural network blocks and cross-attention blocks to condition the features of each point cloud on the other. This cross-conditioning allows the network to reason about correspondences and overlap regions jointly. Due to the large number of parameters introduced by these attention-based components (22.77M total, nearly $3\times$ FCGF's 8.76M), Predator is challenging to train in an unsupervised manner, a challenge that PROMOTER specifically addresses.

From these conditioned features, the model computes not only overlap scores (o) for each superpoint but also *cross-overlap scores*, which estimate the overlap probability of a superpoint's correspondence in the other cloud, formulated as:

$$\tilde{o}_i^{P'} = w_i^T o^{Q'}, \tag{15}$$

where w_i are the attention weights derived from the cross-attention module. The decoder then upsamples these conditioned representations to produce final per-point feature descriptors, along with refined overlap (o) and matchability (m) scores that guide a probabilistic sampling strategy during correspondence establishment. During training, in addition to the Circle Loss [34] for guiding feature learning, Predator introduces an *overlap loss* and a *matchability loss* to optimize the overlap and matchability predictions, respectively. Please refer to the original paper [34] for complete details.

G Discussion On Table 3

The convergence failure of EfficientNet-B7 (main paper Table 3, row e) is *unlikely* due to a deficiency in PROMOTER's design. Instead, it is consistent with our core hypothesis: *success depends on inherent geometric priors in VFMs, not merely model capacity or the availability of additional modalities.*

In Table 3, DINOv2 (86M params) succeeds (86.2 mRR), while EfficientNet-B7 (66M params) fails to converge. Both are general-purpose vision models with comparable capacity, yet only DINOv2 works. A plausible explanation lies in the geometric priors encoded in these models: DINOv2's web-scale self-supervised training with sophisticated augmentations implicitly embeds geometric understanding, enabling depth estimation and correspondence matching via linear probing as shown in their paper [26]. EfficientNet-B7, trained solely with supervised classification on ImageNet, learns to be *invariant* to geometric transformations (rotations, scales, viewpoints), which is precisely the cues critical for registration. Its cross-entropy objective provides *little incentive* to capture geometric relationships.

Why This Supports PROMOTER. PROMOTER's premise is to extract and leverage inherent geometric priors from VFMs, *i.e.*, priors that enable 3D scene understanding, cross-view correspondence, and geometric consistency from 2D observations [75, 13]. Visual-geometric VFMs (VGGT, DAMv2) perform strongly via explicit visual-geometric training; DINOv2 succeeds via implicit geometric learning at scale. When the visual model lacks such geometric information (as in the EfficientNet case), PROMOTER's components may struggle to extract useful signals: AdaPI's adaptive integration assumes *some* reliable geometric cues to filter; PGL's perceptual consistency requires cross-view geometric coherence. Without these, pseudo-labels degrade, increasing the risk of feature-space collapse (Fig. 8). The convergence failure is thus consistent with the view that our framework's performance depends on the availability of geometric priors in the visual model.

H Additional Experimental Results

H.1 AdaPI's Efficacy as a Feature-Space Stabilizer

AdaPI stabilizes the difficult early phase of unsupervised training on the extremely sparse and irregular point clouds from nuScenes. We attribute this to its effective integration of VFM-derived priors. Fig. 8 shows the distribution of feature-space similarity $d_f = \cos(\mathbf{f}^{\mathbf{p}_i}, \mathbf{f}^{\mathbf{q}_j})$ versus ground-truth geometric residual $d_g = \|(\mathbf{R}_{\text{gt}}\mathbf{p}_i + \mathbf{t}_{\text{gt}}) - \mathbf{q}_j\|_2$ for all point pairs on the nuScenes dataset. A top-left, *densely concentrated* region (indicated by lighter colors) reflects a *well-structured feature space*

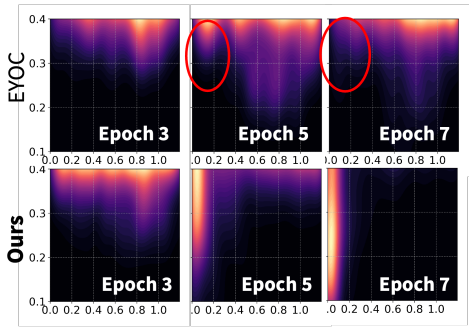


Figure 8: **Distributions of d_f (y axis) vs. d_g (x axis)** show the stabilized training process in early epochs. Our method (Bottom) maintains a concentrated distribution of high feature similarity for geometrically correct correspondences, while EYOC (Top) suffers from feature-space collapse, dispersing the distribution. The color intensity indicates the density in each region.

Table 8: Spatial Compatibility Measure Choices in PGL

	SC^2-PCR	FastMAC	PointDSC
tIR@1st Epoch	82.8	82.1	81.7
Time(s) Per Pair	0.63	0.55	0.67

in which geometrically correct correspondences exhibit high feature similarity. In early training, EYOC suffers from feature-space "collapse": its initially learned features degrade within a few epochs, dispersing the high-density region. PROMOTER maintains a stable, increasingly concentrated distribution, demonstrating the benefits of AdaPI in addressing data sparsity and irregularity.

H.2 Quantitative Results

Spatial Compatibility Measure Choices in PGL. Table 8 shows consistent results across three spatial compatibility methods: our default SC^2 [46], graph-based FastMAC [48], and learning-based PointDSC [36] (with official KITTI pretrained weights). All achieve comparable pseudo-label quality (tIR@1st Epoch $\sim$82%) and similar training efficiency. This confirms that PGL's effectiveness stems from its *perceptual consistency scoring* rather than the specific geometric filter chosen. The VFM-derived perceptual score provides complementary signals that enhance pseudo-label quality regardless of the underlying robust estimator.

Adaptability Across Registration Models on nuScenes. Table 9 evaluates PROMOTER with different registration models (all using the DAMv2 VFM) on the challenging nuScenes dataset. D3Feat achieves the best performance, benefiting from its joint detection-description framework and consistent with KITTI observations. Notably, Predator's 69.3 mRR represents a *notable improvement*: in our experiments, prior unsupervised methods struggle to train this high-capacity model (22.77M params) on nuScenes's sparse 32-beam data, yet PROMOTER produces reasonable results. This indicates improved scalability while leaving room for future refinement in leveraging sophisticated architectures for unsupervised outdoor LiDAR registration.

Using Multiple Camera Views. Table 10 demonstrates PROMOTER's flexibility with multi-view inputs. On KITTI, dual stereo cameras provide minimal gain (86.5→86.7 mRR), consistent with

Table 9: **Scaling with Different Registration Models on nuScenes.** #Params denote the number of parameters of the registration model.

Model	#Params	VFM	mRR	$d \in [40, 50)$		
				RR	RRE	RTE
FCGF	8.76		72.0	45.5	0.54	0.31
D3Feat	14.79	DAMv2	76.9	51.4	0.51	0.29
Predator	22.77		69.3	36.8	0.63	0.33

Table 10: **Impact of Using Multiple Camera Views in AdaPI.** Results are reported on KITTI and nuScenes datasets.

Configuration	Dataset	mRR	RR@$[40, 50)$m
Single View (Default)	KITTI	86.5	63.3
Dual View		86.7	63.8
Single View (Default)	nuScenes	72.0	45.5
Dual View		77.1	59.5

Table 11: **Additional Generalizability Results from KITTI to nuScenes Compared with Supervised Methods.** The best *unsupervised* results are highlighted in **bold**.

Method	mRR	RR@$d \in$				
		$[5, 10)$	$[10, 20)$	$[20, 30)$	$[30, 40)$	$[40, 50)$
Supervised Methods						
FCGF	60.1	93.9	84.5	58.9	39.2	23.9
FCGF+C	60.3	95.9	84.5	61.1	35.8	24.1
Predator	32.9	91.1	50.2	11.2	5.9	6.1
Unsupervised Methods						
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**

our single-camera justification (Table 6). On nuScenes, dual cameras (front+back) yield clear improvements (72.0→77.1 mRR, 45.5→59.5% RR@[40,50)), indicating that PROMOTER can exploit richer information when available. This asymmetry is expected: nuScenes's 360° camera array has less overlap than KITTI's forward-facing stereo pair, making additional views more informative.

Additional Generalizability Baselines. Table 11 extends the main paper's generalization results with additional supervised baseline comparisons. PROMOTER achieves 66.1 mRR when transferring KITTI→nuScenes, outperforming all unsupervised methods and several supervised baselines in the challenging cross-dataset transfer.

Interestingly, FCGF generalizes better than when trained from scratch on nuScenes, a phenomenon also observed in prior work [5, 6], potentially attributable to nuScenes's lower resolution (32 beams, 70 m range) making direct training harder. PROMOTER performs well in *both* scenarios, with strong generalization and competitive from-scratch training results (main paper Table 1). By leveraging VFM geometric priors via AdaPI+PGL, PROMOTER learns robust features even from very sparse data, supporting its applicability across diverse LiDAR configurations.

H.3 Additional Qualitative Results

Fig. 9 visualizes KITTI registration at $d \in [40, 50)$ m—a challenging low-overlap regime. PROMOTER often produces alignments visibly closer to ground truth than the strong baseline EYOC [5], where the baseline struggles. Notably, Sample (3) demonstrates robustness to repetitive structures (roads, buildings) that are prone to ambiguous matching, since PROMOTER's VFM-enhanced features help resolve such ambiguities.

I Limitations, Broader Impact, and Future Directions

While PROMOTER substantially advances unsupervised registration scalability and enables the training of larger models [34] than prior work [5, 6], a performance gap remains for very high-capacity architectures such as Predator [34] relative to simpler models like FCGF. This may stem from: (1) residual pseudo-label noise from PGL (though substantially reduced compared to prior methods), which is amplified by larger parameter spaces; and (2) Predator's architectural sensitivity to noisy early-stage supervision.

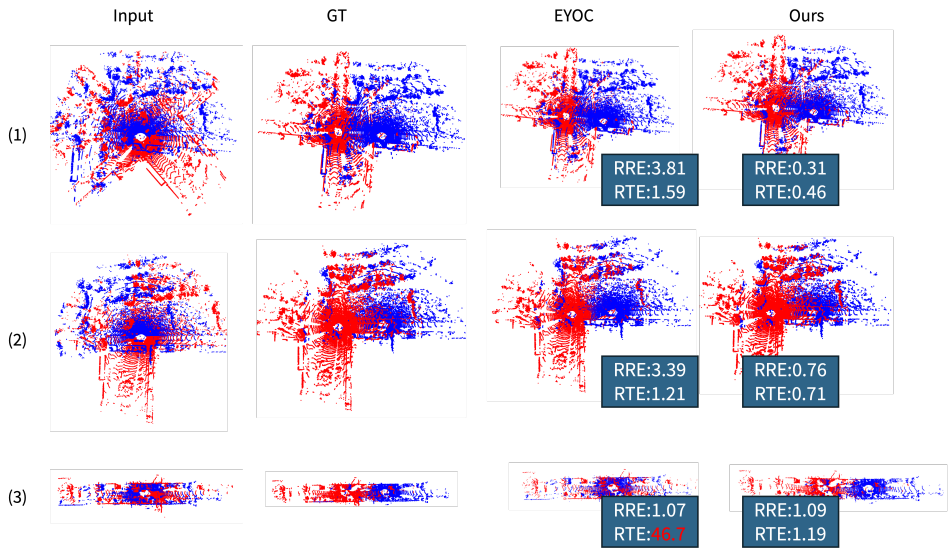


Figure 9: **Qualitative Registration Results on KITTI with Frame Distances $d \in [40, 50]$m.** We visualize challenging low-overlap registration scenarios. PROMOTER often achieves alignments that are visibly closer to the ground truth compared to the baseline method, especially in cases with simple and repetitive structures that are prone to producing ambiguous matching (Sample (3)).

Future work could explore multi-stage pseudo-label refinement for iterative quality improvement. Additionally, extending PROMOTER to incorporate temporal consistency across longer sequences (beyond pairwise registration) could further enhance pseudo-label quality by leveraging multi-frame geometric constraints. These directions could help unlock the full potential of sophisticated modern registration architectures in unsupervised settings.

Broader Impact Statement. This work aims to improve unsupervised LiDAR point cloud registration by transferring visual-geometric priors, reducing reliance on expensive pose annotations while preserving inference efficiency. The primary positive impact lies in enabling more robust 3D perception and mapping for robotics and autonomous systems, which can enhance safety and reliability. Our research does not involve personal data or human subjects, and we will release code and training details to support transparency and reproducibility. We acknowledge that, in its current form, our method shares the limitations of the visual foundation models it relies on, which may restrict its robustness in safety-critical environments. As these limitations are not specific to our method but rather inherent to the underlying foundation models, we believe there are no particular societal consequences that we must specifically highlight here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's three core contributions, including the train-only AdaPI, the VFM-guided PGL, and MAT for student training. They also scope the claims properly by emphasizing state-of-the-art results on KITTI and nuScenes and zero additional inference overhead.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The appendix includes a dedicated limitations section that explicitly discusses the remaining gap on very high-capacity models, residual pseudo-label noise, and future directions such as multi-stage refinement and temporal consistency.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is primarily empirical and does not present formal theorems or proof obligations. The few theory-style remarks are informal intuitions, so this item is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The manuscript has specified the architecture, training procedure, datasets/splits, losses, metrics, and evaluation protocol in enough detail to reproduce the main experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be released after acceptance, and the appendix provides the implementation details, hyperparameters, and evaluation setup needed to reproduce the results. The experiments rely on standard public KITTI and nuScenes benchmarks with official splits.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The appendix specifies the dataset splits, training recipe, and the evaluation protocol, which is sufficient to understand and replicate the reported behavior.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports extensive benchmark tables and ablations, but it does not include error bars or significance tests. The fixed-split evaluation is standard for this setting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix states the training hardware, memory footprint, optimizer setup, epoch count, and batch sizes, and the main paper reports training-time comparisons against prior work.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public autonomous-driving benchmarks, does not involve personal data or human subjects, and is presented as a perception method for improving robustness and efficiency. The manuscript does not indicate any conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The appendix includes an explicit broader-impact statement describing the positive impact of more robust 3D perception and mapping for robotics and autonomous systems, while also acknowledging inherited limitations in safety-critical settings.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk dual-use model or a scraped dataset. Its VFM-related modules are training-only and are discarded at inference, so specialized misuse safeguards are not needed here.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits all external datasets, models, and libraries through citations and uses standard public benchmarks with official splits and established implementations. No uncredited or repackaged third-party asset package is introduced.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper introduces a method rather than a new dataset or benchmark asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work uses public driving datasets and does not involve crowdsourcing or human-subject experiments, so this item is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subject study or participant recruitment is involved in this work, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method is a point-cloud registration framework and does not use LLMs as an important or non-standard component. The manuscript contains no LLM-dependent methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.