

Unsupervised Point Cloud Registration via Training-Time Semantic Guidance

Kezheng Xiong^{1,2,*}, Shiyun Xu^{1,2,*}, Sheng Ao^{1,2}, Siqi Shen^{1,2}, Cheng Wang^{1,2}, and Chenglu Wen^{1,2,✉}

¹ Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China

² Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University, China
 {xiongkezheng, shiyunxu}@stu.xmu.edu.cn,
 {aosh, siqishen, cwang, clwen}@xmu.edu.cn

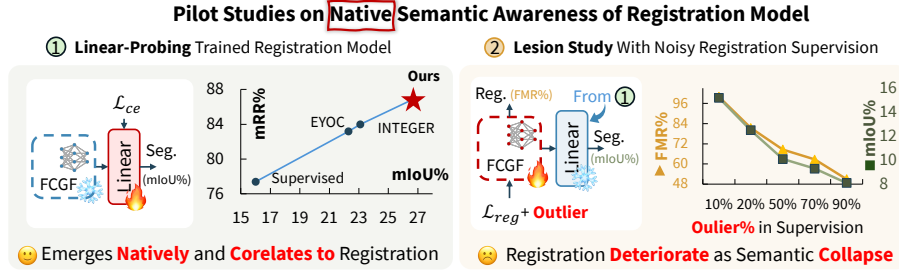


Fig. 1: Motivation: Through linear probing on SemanticKITTI [7], we find that a *fragile yet potent* semantic awareness *emerges natively* in registration networks even without explicit semantic supervision, and it correlates with registration accuracy.

Abstract. Unsupervised registration of large-scale LiDAR point clouds remains challenging due to the geometric ambiguity inherent in outdoor scenes, which degrades pseudo-label quality and leads to suboptimal convergence, particularly for sparse, low-resolution scans such as those from nuScenes. We reveal that registration models **intrinsically** encode semantic awareness that strongly correlates with registration accuracy, albeit without explicit semantic supervision. However, this native awareness is fragile: noisy supervision arising from geometric ambiguity in unsupervised settings rapidly erodes the learned semantic structure, causing performance collapse. To this end, we propose 🦖 **CAESAR**, a teacher-student framework guided by an off-the-shelf 3D segmentation model *exclusively during training*. We observe that potential inlier matches are often buried just beneath a few spurious neighbors in the noisy feature space, motivating Dual-Cue Guided Re-Matching to recover them through reselection rather than simply rejecting. Building on this, a train-only Semantic-Geometric Label Mining performs

✉ Corresponding author.

* Equal contribution.

lightweight, batch-specific teacher refinement and mines reliable pseudo-labels under semantic guidance. We further introduce Semantic Predictive Distillation to consolidate the student’s semantic awareness in the feature space. Extensive experiments on KITTI and nuScenes demonstrate state-of-the-art performance, with pronounced gains on the challenging nuScenes benchmark. Crucially, **CAESAR** incurs **zero inference overhead** and requires **no semantic annotations on the registration data**.

Keywords: Unsupervised registration · Semantic information · LiDAR data processing

1 Introduction

Point cloud registration is fundamental to 3D perception in robotics and autonomous driving, with broad applications in downstream tasks such as SLAM and scene reconstruction [29, 64]. While learning-based approaches achieve impressive accuracy, they require large, meticulously annotated datasets that are costly to obtain. Unsupervised methods [34, 41, 48] offer compelling alternatives by generating correspondence pseudo-labels from internal consistency cues, thereby eliminating the need for pose annotations.

However, the geometric ambiguity inherent in outdoor scenes (*e.g.*, repetitive building facades, featureless roads) degrades pseudo-label reliability in existing frameworks. Self-distillation methods [34], even those that mine latent contextual cues [48], suffer from a *self-referential loop*. Confirmation bias from noisy pseudo-labels gradually corrupts the teacher, which in turn degrades student supervision. This loop is particularly damaging for sparse, low-resolution LiDAR data (*e.g.*, nuScenes), where geometric cues alone are insufficient. Since internally derived signals fail to prevent collapse on their own, stable *external guidance* is needed to break this loop and preserve feature-space integrity.

Semantic information is a natural candidate for such guidance, having long been recognized for resolving geometric ambiguity [21, 52, 56]. However, existing approaches predominantly operate in fully supervised settings [21, 52, 56], leaving the unsupervised regime largely unexplored. Moreover, they exhibit notable practical limitations. Their dependence on *inference-time semantic segmentation models* incurs substantial computational overhead and hinders generalization to domains where such models may be unavailable [21, 51]. Reliance on

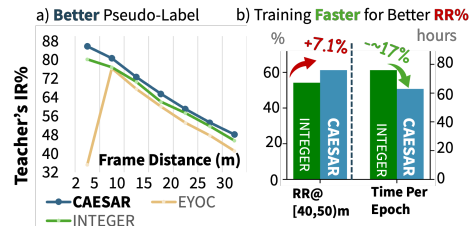


Fig. 2: Performance: Our method leverages this native ability to improve unsupervised point cloud registration even with reduced training time (Sec. 4.3).

hard segmentation predictions for outlier rejection [52] further induces false negatives and thus discard valuable correspondences, because they are *notoriously sensitive to class-boundary noise* and deteriorate in out-of-domain scenes.

Beyond these practical limitations, we argue that existing approaches [21, 51, 52] conceptually mischaracterize the role of semantics, treating it as an auxiliary side-channel that remains permanently coupled to the test-time pipeline. Yet, semantic awareness is arguably an *intrinsic* property of the registration model itself. To establish reliable correspondences, a network must implicitly learn to recognize coherent structures such as buildings or vegetation. If this hypothesis holds, it unlocks a highly efficient paradigm, in which we could effectively preserve the network’s innate semantic structure during training, thereby achieving superior performance with *strictly zero* inference overhead.

Driven by this potential, we conduct pilot studies (see Fig. 1). Through linear probing, we confirm that registration networks indeed exhibit *native semantic awareness* that correlates well with *registration* performance. However, the lesion analysis reveals a critical fragility. Even moderate noise in supervision disrupts both geometric alignment and the latent semantic structure. We term this failure mode *semantic collapse*, a phenomenon that is particularly devastating under the inherently noisy supervision of unsupervised registration.

To this end, we introduce 🔥CAESAR to Crystallize and consolidate such native Semantic Awareness, a teacher-student framework that leverages a frozen, pretrained 3D segmentation model *exclusively during training* to provide stable semantic anchors. As a result, it prevents semantic collapse without incurring any inference cost and even *reduces* training time (Fig. 2). To mine reliable pseudo-labels with the teacher, we observe that correct correspondences are often buried just beneath a few spurious feature-space neighbors, motivating Dual-Cue Guided Re-Matching (DCRM) to recover them via reselection from candidate sets. Building on this, we propose the *train-only* Semantic-Geometric Label Mining (SGLM), which efficiently refines the teacher’s feature space specifically for each mini-batch by optimizing only a compact parameter set, and then mines reliable pseudo-labels under semantic guidance. To train the student, we introduce Semantic Predictive Distillation (SPD), which consolidates the student’s semantic awareness by encouraging semantic understanding in the feature space.

Extensive experiments on KITTI and nuScenes demonstrate that CAESAR achieves state-of-the-art results, with especially notable improvements on the challenging nuScenes benchmark. Crucially, the framework incurs *zero inference overhead*, requiring no additional parameters or computation at test time. Our contributions are summarized as follows:

- We identify *semantic collapse* as a key failure mode in unsupervised registration and are the first to propose preventing it via semantic anchoring using a frozen segmentation model during training.
- We present CAESAR, introducing DCRM for recovering correct correspondences via reselection, SGLM for efficient batch-specific teacher refinement and reliable pseudo-label mining, and SPD for consolidating the student’s semantic awareness.

- We achieve new state-of-the-art results on KITTI and nuScenes with zero extra inference cost and reduced training time, while demonstrating robustness across diverse segmentation model choices.

2 Related Work

Point Cloud Registration Learning-based methods typically follow two paradigms.

Direct regression [2, 28, 50] from global descriptors is efficient but can degrade under low overlap, while *correspondence-based* methods [15, 18, 24, 39, 49, 54, 56] establish point-level correspondences before pose estimation, offering greater robustness. Robust estimation under heavy outlier ratios is another critical direction. Learned estimators [5, 13, 33] surpass RANSAC [19] but require supervision; recent non-parametric methods exploit spatial compatibility [12] or graph structures [60, 61] for efficient, training-free inlier selection.

Unsupervised Registration The high cost of pose annotations motivates unsupervised methods. Early attempts use Gaussian mixture models [27, 36], neighborhood consensus [41], or ICP-based reward proxies [30], yet often struggle with the sparsity and low overlap characteristic of outdoor LiDAR data. Teacher-student frameworks [53] reduce pseudo-label noise via spatial filtering [34] and are further improved by contextual mining [48], but still rely solely on internal cues from the registration model, causing quality degradation when geometric ambiguity intensifies. We address this by harnessing the model’s intrinsic semantic awareness through external semantic anchoring.

Semantic Information in Registration Prior works leverage semantic cues for correspondence matching [52], outlier rejection [63], or dynamic-object suppression [16]. Outlier rejection [63] using discrete, hard class predictions is brittle at category boundaries. Feature-fusion approaches [21, 51] combine semantic and geometric descriptors but require frozen segmentation models at inference or 2D–3D projection pipelines, adding considerable overhead. In contrast, we show that registration models *intrinsically* develop semantic awareness, a capability that is potent yet fragile under noisy pseudo-labels. We thus propose to stabilize and strengthen this inherent awareness using external guidance *only during training*, eliminating inference-time cost while achieving superior performance.

3 Method

Problem Formulation Given two LiDAR point clouds $\mathcal{P} = \{\mathbf{p}_i\} \in \mathbb{R}^{m \times 3}$ and $\mathcal{Q} = \{\mathbf{q}_j\} \in \mathbb{R}^{n \times 3}$, point cloud registration aims to estimate the rigid transformation $\mathbf{T} = \{\mathbf{R}, \mathbf{t}\}$ with $\mathbf{R} \in \text{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ that aligns $\mathcal{P}$ to $\mathcal{Q}$.

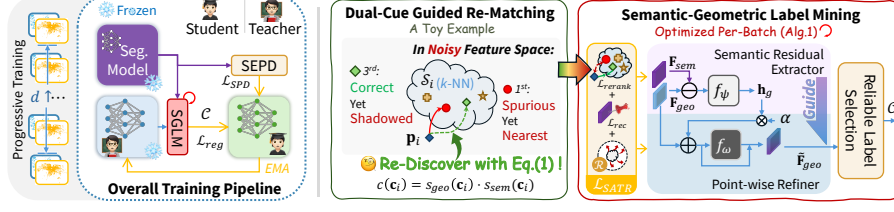


Fig. 3: Overall Pipeline. DCRM recovers correct correspondences obscured by feature-space noise by jointly considering geometric saliency and semantic similarity. The *train-only* SGLM refines the teacher’s feature space and mines high-quality pseudo-labels. The student is trained with its original registration loss on $\mathcal{C}$ and the SPD objective to strengthen semantic awareness.

Overall Pipeline As shown in Fig. 3, CAESAR semantically anchors the teacher-student framework with a frozen, pretrained 3D segmentation model $\mathcal{F}_{\text{sem}}$ [14, 32]. Given a point cloud pair $(\mathcal{P}, \mathcal{Q})$, the teacher extracts geometric features $(\mathbf{F}_{\text{geo}}^{\mathcal{P}}, \mathbf{F}_{\text{geo}}^{\mathcal{Q}})$, while $\mathcal{F}_{\text{sem}}$ provides semantic embeddings $(\mathbf{F}_{\text{sem}}^{\mathcal{P}}, \mathbf{F}_{\text{sem}}^{\mathcal{Q}})$ and per-point class logits $(\mathbf{z}^{\mathcal{P}}, \mathbf{z}^{\mathcal{Q}})$.

For the teacher, DCRM recovers correct correspondences obscured by a few spurious nearest neighbors in the noisy feature space, jointly leveraging geometric saliency and semantic consistency. Building on this, SGLM performs lightweight, batch-specific teacher refinement, optimizing only a compact parameter set to yield semantically anchored features $(\hat{\mathbf{F}}_{\text{geo}}^{\mathcal{P}}, \hat{\mathbf{F}}_{\text{geo}}^{\mathcal{Q}})$. It then mines high-fidelity pseudo-labels $\mathcal{C}$ by jointly enforcing geometric and semantic consistency. The student model is trained with its original registration loss $\mathcal{L}_{\text{reg}}$ (e.g., Hardest Contrastive Loss [15]) on $\mathcal{C}$, while a novel SPD strengthens the student’s semantic awareness by encouraging contextual understanding rather than point-wise mimicry. Following [34], we adopt progressive training to gradually extend from adjacent to distant frame pairs.

Crucially, $\mathcal{F}_{\text{sem}}$ is employed *solely at training time*, incurring zero inference overhead and requiring no semantic annotations on the registration data. The framework is robust to semantic model quality, enabling training even when off-the-shelf models are imperfect or trained on other domains (see Sec. 4.3).

Progressive Training We adopt Progressive Training [34, 48], which gradually adapts the student to increasingly distant point cloud pairs. For each pair, the frame interval $I \in \{1, \dots, I_{\text{max}}\}$ is stochastically determined, where I_{max} increases linearly from 1 to a predefined maximum I_{end} over the course of training epochs. $I=1$ represents easy adjacent frames, while larger I corresponds to low-overlap (e.g., $\leq 40\%$) and challenging pairs.

3.1 Dual-Cue Guided Re-Matching

Motivation Existing methods match each point to its nearest neighbor (NN) in feature space [15, 26, 34, 48] and then *reject* suspected outliers via filtering [34] or clustering [48]. This paradigm faces a fundamental dilemma: when noise dominates, filtering either fails due to vast geometric ambiguity, or discards most matches, thereby “starving” [11, 12] the registration process that requires adequate correspondences [62].

We make a *key observation* (Fig. 4). Although the nearest-neighbor *point* may be incorrect, the true match frequently still lies within the top- K candidate set. This motivates a paradigm shift **from rejection to reselection** (see the toy example in Fig. 3 (Mid.)). Rather than discarding noisy matches [34, 36], DCRM selects the most plausible correspondence from each candidate set, thereby recovering ample high-quality matches even under severe ambiguity, which is a crucial prerequisite for accurate registration [5, 12].

Correspondence Scoring To realize this selection strategy, DCRM scores each candidate correspondence by jointly requiring both *geometric saliency* and *semantic similarity*, respectively denoted as s_{geo} and s_{sem} :

$$c(\mathbf{c}_i) = s_{\text{geo}}(\mathbf{c}_i) \cdot s_{\text{sem}}(\mathbf{c}_i), \quad (1)$$

The two terms are naturally complementary, as registration models struggle with repetitive geometric structures, whereas segmentation models are uncertain near class boundaries. By requiring high scores in *both* aspects simultaneously, Eq. (1) mitigates each source’s failure mode.

Specifically, the term s_{geo} rewards structural distinctiveness [4, 38]. From each point’s spatial neighborhood $\mathcal{N}_k(\mathbf{p}_i)$, the covariance eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3$ yield a saliency ratio $\rho^{\mathbf{p}_i} = \lambda_1 / (\lambda_1 + \lambda_2 + \lambda_3)$, which peaks at non-planar regions such as edges and corners [8]. The pairwise geometric score is then:

$$s_{\text{geo}}(\mathbf{c}_i) = \rho^{\mathbf{p}_i} \cdot \rho^{\mathbf{q}_i} \quad (2)$$

Meanwhile, rather than solely comparing hard class assignments [52], which are brittle near boundaries, s_{sem} uses a weighted Jensen-Shannon divergence (JSD) [20]:

$$s_{\text{sem}}(\mathbf{c}_i) = w_e \exp(-w_c \cdot \text{JSD}(\mathbf{z}^{\mathbf{p}_i}, \mathbf{z}^{\mathbf{q}_i})), \quad (3)$$

where $\mathbf{z}$ denotes C -class softmax probabilities from $\mathcal{F}_{\text{sem}}$. The confidence weight $w_c = (H(\mathbf{z}^{\mathbf{p}_i}) + H(\mathbf{z}^{\mathbf{q}_i})) / 2 \log C$ penalizes uncertain predictions [44] via the

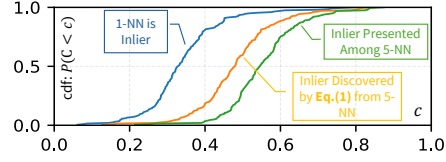


Fig. 4: Cumulative Distribution Function (cdf) of Pairs from KITTI. Variable C refers to the *ratio of correct candidate sets* under certain criteria. During early training, over 80% pairs have lower than 40% correct NN matches, whereas most pairs include twice as many correct matches within their top-5 candidate sets, demonstrating the potential of reselection.

entropy $H(\cdot)$ of the class distribution [40]. The embedding weight $w_e = (1 + \cos(\mathbf{F}_{\text{sem}}^{\mathbf{p}_i}, \mathbf{F}_{\text{sem}}^{\mathbf{q}_i}))/2$ captures holistic semantic context via the cosine similarity of $\mathcal{F}_{\text{sem}}$'s embeddings, further mitigating boundary noise and providing complementary cues for disambiguation.

Constrained Correspondence Reselection For each point $\mathbf{p}_i$, we construct a candidate set from its k -NN in geometric feature space:

$$\mathcal{S}_i = \{\mathbf{s}_k \mid \mathbf{s}_k = (\mathbf{p}_i, \mathfrak{N}_{\mathcal{Q}}^k(\mathbf{p}_i))\}_{k=1}^K, \quad (4)$$

where $\mathfrak{N}_{\mathcal{Q}}^k(\mathbf{p}_i)$ is the k -th NN of $\mathbf{p}_i$ in $\mathcal{Q}$. We select the highest-scoring candidate with $\mathbf{c}_i = \arg \max_{\mathbf{s}_k \in \mathcal{S}_i} c(\mathbf{s}_k)$. Applying similarity thresholds [34, 48], we obtain the refined set $\hat{\mathcal{C}} = \{\mathbf{c}_i\}$. Crucially, $\hat{\mathcal{C}}$ and its candidate sets $\hat{\mathcal{S}} = \{\mathcal{S}_i\}$ serve as a *feature-space quality diagnostic*. In a discriminative feature space, correct matches should rank near the top. We formalize this via the *average rank* (AR):

$$\text{AR}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{(\mathbf{p}_i, \mathbf{q}_j) \in \mathcal{C}} \min_k \{k \mid \mathfrak{N}_{\mathcal{Q}}^k(\mathbf{p}_i) = \mathbf{q}_j\}. \quad (5)$$

A low AR indicates a discriminative feature space where correct matches surface near the top of each candidate set; a high AR signals that true correspondences remain obscured by noise. Beyond its diagnostic role, AR also motivates an effective, differentiable training objective in SGLM (Sec. 3.2).

3.2 Semantic-Geometric Label Mining

While DCRM recovers correspondences obscured by noise, it cannot compensate once true matches fall beyond reselection's reach as semantic collapse intensifies (Sec. 1, Fig. 1). The *train-only* SGLM addresses this root cause by stabilizing the feature space itself, thereby *preserving* the teacher's latent semantic awareness. It proceeds in two stages. Semantic-aware Teacher Refining first reinforces the teacher's features using stable semantic anchors from $\mathcal{F}_{\text{sem}}$ by optimizing a compact parameter set specific to each batch, and Reliable Label Selection then selects reliable pseudo-labels from the refined correspondences.

Stage 1: Semantic-aware Teacher Refining (SATR) When pseudo-label noise severely accumulates, reselection alone cannot reverse the ensuing semantic erosion. Existing efforts either freeze the teacher entirely [34], forfeiting adaptability to evolving noise, or adapt the full teacher per-batch [48], incurring prohibitive cost and risking overfitting. With $\mathcal{F}_{\text{sem}}$ as a stable reference, we instead design a concise train-only adapter to reinforce the eroding semantic structure in the feature space, thereby preventing collapse during training. SATR is far more efficient, since it optimizes a mere $0.07M$ parameters (vs. $8.76M$ adapted in previous state-of-the-art [15]) while achieving better performance. The efficiency is ensured not only by the compact design, but also through the AR metric (Eq. (5)), which directly exposes noise-induced degradation and can thus guide a refinement process to restore a discriminative feature space.

Architecture Semantic embeddings reside in a far higher-dimensional space ($d_{\text{sem}} = 512$ [14] vs. $d_{\text{geo}} = 32$ [15]). Naïve fusion or concatenation [21,31,35,51] risks diluting geometric details or conflating geometrically distinct points of the same class (*e.g.*, separate trees/vehicles). Therefore, SATR is designed to extract a compact *semantic residual* containing the absent, discriminative information present in $\mathcal{F}_{\text{sem}}$. Targeting the ambiguity revealed by DCRM (Eq. (5)) within candidate sets, its parameters are freshly initialized for each batch and optimized *locally* in a regularized way (Eqs. (9) and (10)). This controlled and adaptive refinement enables **CAESAR** to remain effective even with imperfect semantics (see Sec. 4.3).

Specifically, the SATR comprises a *residual extractor* Φ_g and a *point-wise refiner* Φ_r , connected via a learnable gate α . Φ_g projects geometric features $\mathbf{F}_{\text{geo}}^{\mathcal{P}} \in \mathbb{R}^{N \times d_{\text{geo}}}$ and semantic embeddings $\mathbf{F}_{\text{sem}}^{\mathcal{P}} \in \mathbb{R}^{N \times d_{\text{sem}}}$ into a shared latent space and isolates discriminative semantic signals absent from the geometric representation through a sparse convolutional block [15] f_ψ :

$$\mathbf{h}_g^{\mathcal{P}} = f_\psi(\mathbf{F}_{\text{sem}}^{\mathcal{P}} \mathbf{W}_{\text{sem}} - \mathbf{F}_{\text{geo}}^{\mathcal{P}} \mathbf{W}_{\text{geo}}) \mathbf{W}_{\text{out}}. \quad (6)$$

To keep these residuals faithful to the semantic source, Φ_g is trained to recover the original semantic embeddings via an auxiliary projection f_ϕ :

$$\mathcal{L}_{\text{rec}}^{\mathcal{P}} = \|\mathbf{F}_{\text{sem}}^{\mathcal{P}} - f_\phi(\mathbf{F}_{\text{geo}}^{\mathcal{P}} + \text{sg}(\tanh(\alpha)) \cdot \mathbf{h}_g^{\mathcal{P}})\|_2^2; \quad (7)$$

Owing to the complementary stop-gradient placement (cf. Eq. (8)), $\mathcal{L}_{\text{rec}}$ propagates gradients only through Φ_g . The same applies to $\mathcal{Q}$, giving $\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{rec}}^{\mathcal{P}} + \mathcal{L}_{\text{rec}}^{\mathcal{Q}}$. Next, Φ_r incorporates the residuals into the geometric space via the learnable gate α and a lightweight MLP f_ω :

$$\mathbf{h}_r^{\mathcal{P}} = \mathbf{F}_{\text{geo}}^{\mathcal{P}} + \tanh(\alpha) \cdot \text{sg}(\mathbf{h}_g^{\mathcal{P}}), \quad \tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{P}} = \mathbf{h}_r^{\mathcal{P}} + f_\omega(\mathbf{h}_r^{\mathcal{P}}), \quad (8)$$

where $\text{sg}(\cdot)$ blocks gradients from Φ_r to Φ_g . Together with the $\text{sg}(\cdot)$ placement in Eq. (7), these stop-gradients realize *alternating optimization*, in which each loss updates one module while implicitly treating the other as fixed, allowing Φ_g and Φ_r to specialize without conflicting gradients.

To sharpen feature discriminability with the incorporated semantic cues, we propose a contrastive re-ranking loss that exploits the candidate-set structure from DCRM, pulling the selected correspondence $\mathbf{c}_i \in \hat{\mathcal{C}}$ closer while repelling the remaining $K-1$ candidates [37]. This is achieved via a *differentiable surrogate* of the AR diagnostic that smoothly relaxes Eq. (5):

$$\mathcal{L}_{\text{rerank}} = \frac{1}{|\hat{\mathcal{C}}|} \sum_{\mathbf{c}_i \in \hat{\mathcal{C}}} \log \left(1 + \sum_{\mathbf{q}_k \in \mathcal{S}_i \setminus \{\mathbf{q}_j\}} \exp \left(\frac{s(\mathbf{p}_i, \mathbf{q}_k) - s(\mathbf{p}_i, \mathbf{q}_j)}{\tau} \right) \right) \quad (9)$$

Please refer to the supplementary material for full mathematical proof of this relaxation. Here, $s(\cdot, \cdot)$ denotes cosine similarity between refined features and τ is a temperature. The same computation applies to $\mathcal{Q}$. Unlike standard contrastive registration losses [15,26,39,59], $\mathcal{L}_{\text{rerank}}$ operates *locally* within each $\mathcal{S}_i$, confining the influence of erroneous selections to their immediate neighborhoods.

Volumetric Regularized Objective While $\mathcal{L}_{\text{rerank}}$ and $\mathcal{L}_{\text{rec}}$ ensure each component serves its intended role, one concern remains. Incorporating high-dimensional semantic residuals into a compact d_{geo} -dimensional space risks diluting its volume [57], neutralizing the geometric discriminative capacity upon which CAESAR relies for pseudo-labeling. We prevent this with an exploration regularizer that maximizes the geometric volume of the feature distribution [47, 57]:

$$\mathcal{R}^{\mathcal{P}} = -\gamma \frac{1}{2} \log \det(\mathbf{I} + \frac{1}{\epsilon^2} \Gamma_{\text{batch}}^{\mathcal{P}}), \quad (10)$$

where $\Gamma_{\text{batch}}^{\mathcal{P}} \in \mathbb{R}^{d \times d}$ is the batch-wise covariance matrix and γ, ϵ control regularization strength. Aggregating all terms yields the full optimization objective:

$$\mathcal{L}_{\text{SATR}} = \underbrace{\mathcal{R}^{\mathcal{P}} + \mathcal{R}^{\mathcal{Q}}}_{\text{Exploration}} + \underbrace{\mathcal{L}_{\text{rerank}} + \mathcal{L}_{\text{rec}}}_{\text{Exploitation}}. \quad (11)$$

This defines an effective exploration–exploitation loop. Volume maximization creates room for semantic residuals, re-ranking ($\mathcal{L}_{\text{rerank}}$) thereby sharpens local discrimination, and reconstruction $\mathcal{L}_{\text{rec}}$ prevents the semantic guidance from drifting, keeping the adapted space both expressive and semantically grounded.

Batch-Specific Optimization With the objective in Eq. (11), the optimization of SATR is driven by Algorithm 1, freshly initialized for every mini-batch and optimized in an Expectation-Maximization (EM) procedure [17]. We apply zero-initialization to f_{ω} ’s final layer and α , ensuring gradual, stable refinement from the original features. In the E-step, Φ_r and Φ_g are frozen while DCRM mines correspondences $\hat{\mathcal{C}}$ and their candidate sets $\hat{\mathcal{S}}$ from the current refined features. In the M-step, these quantities are held fixed and Φ_r , Φ_g are updated by minimizing $\mathcal{L}_{\text{SATR}}$, with the stop-gradient decoupling (Eqs. (7) and (8)) implicitly alternating optimization between the two components, yielding a two-level optimization hierarchy. The E/M alternation terminates once AR (Eq. (5)) stabilizes, which serves as a natural convergence criterion, since AR is the metric to reflect feature-space ambiguities that $\mathcal{L}_{\text{rerank}}$ smoothly surrogates. A final DCRM pass then yields the refined proposals $\tilde{\mathcal{C}}$ for Reliable Label Selection. The trained modules are discarded after each batch, preventing confirmation bias from accumulating across batches.

Thresholds and other hyperparameters are detailed in the supplementary material with sensitivity analyses. Empirically, Algorithm 1 converges stably and rapidly, typically within 3–5 iterations, yielding improved pseudo-labels while notably reducing training time compared to full model adaptation [48] (Sec. 4.3).

Stage 2: Reliable Label Selection (RLS) With refined features from the teacher, this stage produces reliable pseudo-labels $\mathcal{C}$ for student training. Rather than relying on a single transformation estimated from noisy initial correspondences [34, 48], we generate K transformation hypotheses $\{\mathbf{T}_k\}_{k=1}^K$ from the refined proposals $\tilde{\mathcal{C}}$ via spatial compatibility measures [12, 61] and select the best by

Algorithm 1 Batch-Specific Optimization for Semantic-aware Teacher Refining

Require: Geometric/semantic features $\mathbf{F}_{\text{geo}}^{\mathcal{P}}, \mathbf{F}_{\text{geo}}^{\mathcal{Q}} / \mathbf{F}_{\text{sem}}^{\mathcal{P}}, \mathbf{F}_{\text{sem}}^{\mathcal{Q}}$; logits $\mathbf{z}^{\mathcal{P}}, \mathbf{z}^{\mathcal{Q}}$
Ensure: Refined proposals $\tilde{\mathcal{C}}$

- 1: Initialize trainable parameters from Φ_r, Φ_g $\triangleright$ freshly initialized per batch
- 2: Calculate $\tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{P}}, \tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{Q}}$ by Eqs. (6) and (8) on $\mathbf{F}_{\text{geo}}^{\mathcal{P}}, \mathbf{F}_{\text{geo}}^{\mathcal{Q}}, \mathbf{F}_{\text{sem}}^{\mathcal{P}}, \mathbf{F}_{\text{sem}}^{\mathcal{Q}}$
- 3: **repeat**
- 4: **E-step:** Fix Φ_r, Φ_g ; let $\hat{\mathcal{C}}, \hat{\mathcal{S}} \leftarrow \text{DCRM}(\tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{P}}, \tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{Q}})$
- 5: **M-step:** Fix $\hat{\mathcal{C}}, \hat{\mathcal{S}}$; update Φ_r, Φ_g by gradient descent via $\nabla_{\Phi_r, \Phi_g} \mathcal{L}_{\text{SATR}}$
- 6: Calculate $\tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{P}}, \tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{Q}}$ by Eqs. (6) and (8) on $\mathbf{F}_{\text{geo}}^{\mathcal{P}}, \mathbf{F}_{\text{geo}}^{\mathcal{Q}}, \mathbf{F}_{\text{sem}}^{\mathcal{P}}, \mathbf{F}_{\text{sem}}^{\mathcal{Q}}$
- 7: **until** $\text{AR}(\hat{\mathcal{C}})$ difference converges within ϵ
- 8: $\tilde{\mathcal{C}} \leftarrow \text{DCRM}(\tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{P}}, \tilde{\mathbf{F}}_{\text{geo}}^{\mathcal{Q}})$ $\triangleright$ final proposals for Reliable Label Selection

re-weighting each hypothesis’s consensus set (*i.e.*, sets of selected inliers [11, 12]) $\mathcal{C}_k$ with $c(\cdot)$ defined in Eq. (1):

$$k^* = \arg \max_k \left(\text{IC}(\mathbf{T}_k) \cdot \frac{1}{|\mathcal{C}_k|} \sum_{\mathbf{c}_i \in \mathcal{C}_k} c(\mathbf{c}_i) \right). \quad (12)$$

Here, $\text{IC}(\cdot)$ denotes the inlier count metric commonly used to evaluate transformation hypotheses [11, 12, 19, 60, 61]. This ensures the selected transformation is grounded in correspondences that are both geometrically salient and semantically coherent. Using $\mathbf{T}_{k^*}$, we follow [34, 48] to obtain $\mathcal{C}'_{k^*}$ via NN search. The final pseudo-labels are then selected as the top- M highest-scoring correspondences:

$$\mathcal{C} = \{\mathbf{c}_i \mid \mathbf{c}_i \in \mathcal{C}'_{k^*}, i \in \text{TopM}(\{c(\mathbf{c}_i)\})\}. \quad (13)$$

This final selection step prevents *systematically* incorrect pseudo-labels caused by inaccurate transformations that are more prone to memorize [10, 45], thereby enhancing supervision quality and ultimately improving registration performance.

3.3 Initialization and Training Pipeline

Teacher Initialization and Update Unlike prior methods that require additional synthetic pre-training [48] or assume identity transformations between adjacent frames as ground truth [34] for initialization, the proposed approach of semantic anchoring enables reliable pseudo-labeling on adjacent frame pairs, even from a randomly initialized teacher. In the first stage of progressive training [34], we therefore train the student without SPD to avoid propagating feature-level noise, synchronizing the teacher with the student after each epoch. In subsequent stages, the teacher is frozen and updated once per epoch via exponential moving average (EMA) of the student weights.

Semantic Predictive Distillation for Student Training the student with the registration loss $\mathcal{L}_{\text{reg}}$ (*e.g.*, Hardest Contrastive Loss [15]) on the pseudo-labels $\mathcal{C}$ transfers geometric correspondence knowledge. However, $\mathcal{L}_{\text{reg}}$ carries no explicit signal about the semantic structure captured by $\mathcal{F}_{\text{sem}}$. Directly regressing

high-dimensional semantic embeddings is suboptimal due to the feature-space gap ($d_{\text{sem}} \gg d_{\text{geo}}$, cf. Sec. 3.2), which collapses feature diversity and dilutes geometric details. We instead formulate SPD as contextual prediction [3, 42], where the student is encouraged to *reconstruct* a point’s semantic embedding from its geometric context, thereby internalizing semantic reasoning. Concretely, distillation targets are restricted to high-confidence points $\tilde{\mathbf{p}}$ whose correspondences $\mathfrak{C} = \{(\tilde{\mathbf{p}}_i, \tilde{\mathbf{q}}_j)\} \subseteq \mathcal{C}$ pass mutual nearest-neighbor verification in the teacher’s refined feature space. For each selected point, we stochastically drop r feature channels and add Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ [46] to the student features, encouraging robustness and preventing overfitting. The objective is defined as:

$$\mathcal{L}_{\text{SPD}}^{\mathcal{P}} = \sum_{\tilde{\mathbf{p}}_i \in \tilde{\mathcal{P}}} c_i \cdot \text{SmoothL1}(f_{\theta}(\mathbf{F}_{\text{stu}}^{\{\mathcal{P} \setminus \tilde{\mathcal{P}}\}} + \epsilon), \mathbf{F}_{\text{sem}}^{\tilde{\mathbf{p}}_i}). \quad (14)$$

Here, f_{θ} is a convolutional [14] auxiliary decoder trained alongside the student to reconstruct the corresponding semantic embedding. c_i weight predictions by estimated quality (Eq. (1)); Smooth L1 [23] provides robustness to spurious targets. The loss is computed symmetrically for $\mathcal{Q}$, giving $\mathcal{L}_{\text{SPD}} = \mathcal{L}_{\text{SPD}}^{\mathcal{P}} + \mathcal{L}_{\text{SPD}}^{\mathcal{Q}}$.

Overall Student Objective The student model is trained end-to-end as a weighted sum of both losses, with λ_{SPD} balancing the two terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda_{\text{SPD}} \mathcal{L}_{\text{SPD}}, \quad (15)$$

4 Experiments

Following existing unsupervised methods for outdoor LiDAR data [34, 48], we evaluate CAESAR on KITTI [22] and nuScenes [9]. Implementation details are provided in the supplementary material.

4.1 Setup

Metrics Following [26, 39, 59], we report *Relative Rotation Error* (RRE), *Relative Translation Error* (RTE), and *Registration Recall* (RR) with thresholds $\text{RRE} < 5^\circ \wedge \text{RTE} < 2\text{m}$. For distant scenarios, we report $\text{RR}@[d_1, d_2]$ for distance $d \in [d_1, d_2]$ and *mean RR* (mRR) [34, 48], defined as the average RR across all distance intervals. We use the teacher’s *Inlier Ratio* at the first teacher-student training epoch (tIR@1st Epoch) to evaluate pseudo-label quality [34, 48].

Baselines We compare against supervised methods, including *geometry-only* FCGF [15], Predator [26], SpinNet [1], D3Feat [6], CoFiNet [55], and GeoTrans. [39], as well as *semantic-aware* baselines S²Reg [51] and ML-SemReg [52], both of which require the segmentation model at inference. For unsupervised methods, we compare with RIENet [41], EYOC [34], and INTEGER [48]. Following [34, 48], we use the FCGF model with RANSAC for transformation estimation unless otherwise specified, ensuring a fair comparison.

Table 1: Comparisons with State-of-the-Art Methods. Bold indicates the best unsupervised results. “✓” in “U” column denotes the methods are Unsupervised. Notably, S²Reg and ML-SemReg require semantic information at inference which adds computational overhead, while CAESAR uses semantics only during training, matching the inference efficiency of purely geometric methods.

Dataset	Semantic Availability Requirement	Method	U	mRR	RR@d ∈				
					[5, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)
KITTI	N/A	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	–	83.0	99.8	98.2	89.6	76.1	51.3
		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
	Inference-Time	S ² Reg	–	52.2	100.0	94.2	61.3	5.0	1.5
		ML-SemReg	–	73.8	98.1	96.3	82.0	68.1	24.3
	N/A	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
	Train-Only	CAESAR	✓	86.9	99.5	97.6	91.1	83.1	61.3
nuScenes	N/A	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
	Inference-Time	S ² Reg	–	33.5	94.0	50.3	14.0	9.1	0.0
		ML-SemReg	–	43.5	87.0	77.1	37.5	12.6	3.2
	N/A	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	✓	63.1	97.1	86.9	62.9	39.6	29.4
	Train-Only	CAESAR	✓	79.5	100.0	96.4	85.6	72.2	43.1
KITTI ↓ nuScenes	N/A	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	✓	62.6	97.5	84.6	62.6	37.8	30.2
	Train-Only	CAESAR	✓	74.1	99.5	94.6	82.3	56.3	37.1

4.2 Comparison with State-of-the-Art

Overall Performance Tab. 1 shows that CAESAR substantially outperforms all unsupervised methods, achieving state-of-the-art results. Notably, CAESAR yields pronounced improvements on nuScenes, validating that preventing *semantic collapse* is critical for sparse, low-resolution LiDAR data. We employ cross-dataset off-the-shelf $\mathcal{F}_{\text{sem}}$ (SphereFormer [32] trained on nuScenes [9] for KITTI and vice versa) by default to avoid information leakage and to benchmark robustness. Consistent gains in distant scenarios (*e.g.*, $d \in [20, 30), [30, 40)$) confirm that semantic anchoring improves discriminability when geometry becomes ambiguous. Crucially, unlike methods [51, 52] that require semantic models at inference, $\mathcal{F}_{\text{sem}}$ is employed *only at training time*, matching the inference cost of purely geometric methods. Additional results with other registration models are provided in the supplementary material.

Generalizability When trained on KITTI and tested on nuScenes, CAESAR substantially outperforms existing methods (+11.5% mRR vs. INTEGER), with

Table 2: Impact of Segmentation Models. We train on KITTI with different $\mathcal{F}_{\text{sem}}$. *In-domain* refers to $\mathcal{F}_{\text{sem}}$ trained on SemanticKITTI [7], while *out-of-domain* is those trained on nuScenes [9]. Sim-to-real $\mathcal{F}_{\text{sem}}$ uses no real-world annotation for semantics, showing CAESAR’s practicality.

# Seg. Model	Paradigm	mRR	$d \in [40, 50]$			
			RR	RRE	RTE	
► <i>In-domain Supervised $\mathcal{F}_{sem}$</i>						
a) SphereFormer	Transformer	87.3	63.3	0.97	0.27	
► <i>Out-of-domain Supervised $\mathcal{F}_{sem}$</i>						
b) SphereFormer	Transformer	86.9	61.3	1.10	0.28	
c) SPVNAS	Point-Voxel	85.6	60.8	1.17	0.30	
d) MinkUNet	Voxel-Based	85.8	61.3	1.07	0.29	
► <i>Unsupervised sim-to-real adapted $\mathcal{F}_{sem}$</i>						
e) DGT-ST	–	85.4	58.3	1.39	0.44	

Table 3: Ablation Study on KITTI.

Variants	tIR@1 st	mRR	$d \in [40, 50]$			
	Epoch		RR	RRE	RTE	
► w/o DCRM	83.7	85.1	59.8	1.34	0.55	
▼ <i>reselection w/ score variants:</i>						
w/ s_{sem}	84.8	85.8	59.3	1.47	0.38	
w/ s_{geo}	84.5	85.6	59.8	1.12	0.31	
► w/o SATR	83.6	84.8	59.3	1.23	0.51	
▼ <i>w/ SATR variants:</i>						
w/o Φ_q	84.9	85.7	59.8	1.30	0.41	
w/o $\mathcal{R}$	84.7	85.8	60.3	1.41	0.40	
$\mathcal{L}_{\text{rerank}} \rightarrow \mathcal{L}_{\text{reg}}$	83.3	85.0	59.8	1.25	0.49	
► w/o RLS	81.2	85.2	60.3	1.62	0.53	
▼ <i>w/ RLS variants:</i>						
Eq. (12) $\rightarrow$ IC	82.0	85.7	59.8	1.21	0.44	
w/o Eq. (13)	83.2	85.5	59.3	1.46	0.52	
► w/o SPD	85.6	85.0	58.8	1.26	0.41	
► $L_{\text{SPD}} \rightarrow L_1$	85.6	84.0	57.8	1.51	0.49	
Full CAESAR	85.6	86.9	61.3	1.10	0.28	

marked long-range improvements (e.g., +6.9% in RR@[40, 50]). This superior cross-dataset transferability, despite significant domain gaps (64- vs. 32-beam LiDAR and distinct scene characteristics), suggests that CAESAR learns more generalizable features, likely because stable semantic guidance prevents overfitting to domain-specific noise.

4.3 Analysis

Impact of Segmentation Model Choice and Quality Tab. 2 shows that CAESAR maintains robust performance across diverse $\mathcal{F}_{\text{sem}}$ choices, including voxel-based [14], point-voxel hybrid [43], transformer-based [32], and *unsupervised* sim-to-real domain-adapted [58] models. Performance is stable with **both in-domain and out-of-domain semantics**, demonstrating robustness to domain shift in semantic priors. Crucially, the unsupervised DGT-ST’s semantic quality **is significantly inferior** to that of supervised methods (SemanticKITTI mIoU: 43.1% vs. SphereFormer’s 74.8%); yet CAESAR still **mines useful semantic guidance** from it and outperforms the previous state-of-the art [48], albeit with a modest drop. This validates that even imperfect semantics suffice for CAESAR to prevent semantic collapse, demonstrating robustness and thereby opening avenues for exploring task synergy. Further robustness analyses with weaker/degraded semantics are provided in the supplementary material.

Qualitative Results We qualitatively analyze the impact of CAESAR on feature-space structure and matching quality. In Fig. 5 (Left), 2D t-SNE [25] visualizations colored by class labels from SemanticKITTI [7] show that CAESAR’s features are well separated compared to EYOC’s. Consequently, in Fig. 5 (Mid.), the features extracted by CAESAR are more discriminative, resulting in fewer outlier correspondences (Fig. 5 (Right)). Note that EYOC produces dissimilar features

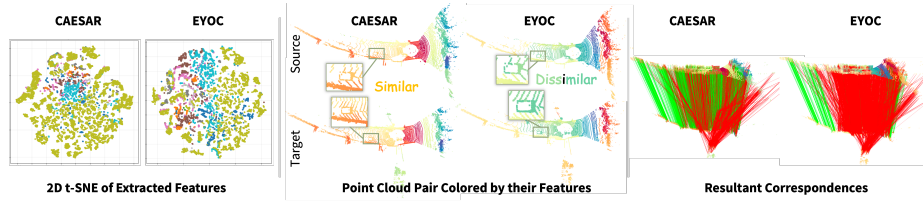


Fig. 5: (Left:) 2D t-SNE Visualization of the Source Point Cloud. Colored by Class Labels. CAESAR’s features are more semantically discriminative than EYOC’s. **(Mid & Right:) Visualization with the Corresponding Target Point Cloud.** Point clouds are colored by features as done in [15]. Red lines denotes outlier.

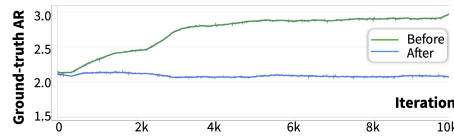


Fig. 6: Average Rank During Training. Semantic-aware Teacher Refining effectively lowers and stabilizes the AR of potentially correct matches during training.

Table 4: Training/Testing Time on KITTI. $\Delta\%$ is the difference with previous state-of-the-art INTEGER [48].

Methods	mRR%	RR@d $\in$ [40, 50]	Training (h)	Testing (s/pair)
EYOC	83.2	52.3	44.9	~ 0.16
INTEGER	84.0	54.2	76.0	
CAESAR	86.9	61.3	62.9	
$\Delta\%$	+2.9%	+7.1%	-17.2%	–

for the same semantic entity (highlighted regions), whereas CAESAR’s features remain consistent. This confirms that CAESAR effectively prevents semantic collapse, producing more stable and discriminative features that improve matching quality, especially in long-range scenarios where geometric cues are weak.

SATR’s Effectiveness for Stability We evaluate Semantic-aware Teacher Refining using the *average rank of ground-truth* (AR_{gt}) of inliers in Fig. 6. Its definition follows Eq. (5) but is restricted to correspondences whose ground-truth inliers appear within the candidate set. As frame distance increases during progressive training [34, 48], the AR before refinement (green line) rises during progressive training due to the ambiguity induced by lower overlap. However, SGLM stabilizes feature-space discriminability, as evidenced by the consistent AR after refinement (blue line).

Training/Testing Efficiency Training and testing times are shown in Tab. 4. The iterative nature and full-teacher adaptation strategy of INTEGER [48] make its training computationally intensive. By alleviating the self-referential collapse via semantic anchoring, CAESAR eliminates the need for extensive iterative refinement that optimizes the full teacher model, delivering stronger results while reducing training time. Testing efficiency matches that of the original FCGF [15] because all additional modules are discarded at inference.

4.4 Ablation Study

We extensively ablate CAESAR’s design in Tab. 3. DCRM is crucial: neither s_{sem} nor s_{geo} alone suffices, confirming the complementarity of semantic and geometric cues. Semantic-aware Teacher Refining significantly stabilizes training; removing it degrades mRR by 2.1%. Among its variants, replacing $\mathcal{L}_{\text{rerank}}$ with $\mathcal{L}_{\text{reg}}$ causes the sharpest decline (−1.9% mRR), confirming that the candidate-set-local surrogate of AR is critical; the global loss fails to utilize the per-set ambiguity and risks corrupting the full feature space with noise. Reliable Label Selection is essential for pseudo-label quality; its removal causes a considerable drop in tIR@1st, triggering cascading errors in teacher-student training that are particularly severe in long-range scenarios. Removing SPD notably reduces final mRR, while replacing it with vanilla L_1 loss degrades performance even further, demonstrating that naïve feature-space mimicry harms the student’s learning.

5 Conclusion

In this paper, we observe that registration models spontaneously develop latent semantic structure, yet noisy pseudo-labels quickly erode it, triggering semantic collapse. CAESAR addresses this fragility by stabilizing the emergent awareness at training time through an off-the-shelf segmentation model, without adding cost at inference or requiring semantic annotations on the registration data. For the teacher, DCRM recovers reliable correspondences and exposes local ambiguities with semantic information; SGLM then refines the feature space within each mini-batch and mines accurate pseudo-labels. For the student, SPD complements standard registration supervision by encouraging holistic semantic reasoning in the learned features. Results on KITTI and nuScenes establish new state-of-the-art registration performance, with particularly large margins on the sparser nuScenes benchmark, underscoring the effectiveness of semantic anchoring when geometric cues alone are insufficient. *Limitations and future work* are discussed in the supplementary material.

6 Acknowledgements

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

References

1. Ao, S., Hu, Q., Yang, B., Markham, A., Guo, Y.: Spinnet: Learning a general surface descriptor for 3d point cloud registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11753–11762 (2021)
2. Aoki, Y., Goforth, H., Srivatsan, R.A., Lucey, S.: Pointnetlk: Robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7163–7172 (2019)

3. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
4. Avidar, D., Malah, D., Barzohar, M.: Point cloud registration using a viewpoint dictionary. In: 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE). pp. 1–5. IEEE (2016)
5. Bai, X., Luo, Z., Zhou, L., Chen, H., Li, L., Hu, Z., Fu, H., Tai, C.L.: Pointdsc: Robust point cloud registration using deep spatial consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15859–15869 (2021)
6. Bai, X., Luo, Z., Zhou, L., Fu, H., Quan, L., Tai, C.L.: 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. pp. 6359–6367 (2020)
7. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019)
8. Belton, D.: Improving and extending the information on principal component analysis for local neighborhoods in 3d point clouds. the international archives of the photogrammetry. Remote Sensing and Spatial Information Sciences **37**, B5 (2008)
9. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
10. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press (2006)
11. Chen, Z., Sun, K., Yang, F., Guo, L., Tao, W.: Sc2-pcr++: Rethinking the generation and selection for efficient and robust point cloud registration. IEEE transactions on pattern analysis and machine intelligence **45**(10), 12358–12376 (2023)
12. Chen, Z., Sun, K., Yang, F., Tao, W.: 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. pp. 13221–13231 (2022)
13. Choy, C., Dong, W., Koltun, V.: Deep global registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2514–2523 (2020)
14. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
15. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8958–8966 (2019)
16. Cui, J., Chen, J., Li, L.: Sage-icp: Semantic information-assisted icp. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 8537–8543. IEEE (2024)
17. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
18. Deng, H., Birdal, T., Ilic, S.: Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In: Proceedings of the European conference on computer vision (ECCV). pp. 602–618 (2018)

19. Fischler, M.A., Bolles, R.C.: 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)
20. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: *International symposium on Information theory, 2004. ISIT 2004. Proceedings.* p. 31. IEEE (2004)
21. Fung, S., Lu, X., Edirimuni, D.d.S., Pan, W., Liu, X., Li, H.: Semreg: Semantics constrained point cloud registration. In: *European Conference on Computer Vision.* pp. 293–310. Springer (2024)
22. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE conference on computer vision and pattern recognition.* pp. 3354–3361. IEEE (2012)
23. Girshick, R.: Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision.* pp. 1440–1448 (2015)
24. Gojcic, Z., Zhou, C., Wegner, J.D., Wieser, A.: The perfect match: 3d point cloud matching with smoothed densities. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* pp. 5545–5554 (2019)
25. Hinton, G.E., Roweis, S.: Stochastic neighbor embedding. *Advances in neural information processing systems* **15** (2002)
26. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition.* pp. 4267–4276 (2021)
27. Huang, X., Li, S., Zuo, Y., Fang, Y., Zhang, J., Zhao, X.: Unsupervised point cloud registration by learning unified gaussian mixture models. *IEEE Robotics and Automation Letters* **7**(3), 7028–7035 (2022)
28. Huang, X., Mei, G., Zhang, J.: 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.* pp. 11366–11374 (2020)
29. Jiang, H., Salzmann, M., Dang, Z., Xie, J., Yang, J.: Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems* **36**, 21285–21297 (2023)
30. Jiang, H., Shen, Y., Xie, J., Li, J., Qian, J., Yang, J.: Sampling network guided cross-entropy method for unsupervised point cloud registration. In: *Proceedings of the IEEE/CVF international conference on computer vision.* pp. 6128–6137 (2021)
31. Kong, X., Yang, X., Zhai, G., Zhao, X., Zeng, X., Wang, M., Liu, Y., Li, W., Wen, F.: Semantic graph based place recognition for 3d point clouds. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* pp. 8216–8223. IEEE (2020)
32. Lai, X., Chen, Y., Lu, F., Liu, J., Jia, J.: Spherical transformer for lidar-based 3d recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* pp. 17545–17555 (2023)
33. Lee, J., Kim, S., Cho, M., Park, J.: Deep hough voting for robust global registration. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* pp. 15994–16003 (2021)
34. Liu, Q., Zhu, H., Wang, Z., Zhou, Y., Chang, S., Guo, M.: 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.* pp. 11366–11374 (2024)

35. Liu, S., Wang, T., Zhang, Y., Zhou, R., Li, L., Dai, C., Zhang, Y., Wang, L., Wang, H.: Deep semantic graph matching for large-scale outdoor point cloud registration. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–12 (2024)
36. Mei, G., Tang, H., Huang, X., Wang, W., Liu, J., Zhang, J., Van Gool, L., Wu, Q.: Unsupervised deep probabilistic approach for partial point cloud registration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13611–13620 (2023)
37. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
38. Pauly, M., Gross, M., Kobbelt, L.P.: Efficient simplification of point-sampled surfaces. In: *IEEE Visualization, 2002. VIS 2002*. pp. 163–170. IEEE (2002)
39. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11143–11152 (2022)
40. Shannon, C.E.: A mathematical theory of communication. *The Bell system technical journal* **27**(3), 379–423 (1948)
41. Shen, Y., Hui, L., Jiang, H., Xie, J., Yang, J.: Reliable inlier evaluation for unsupervised point cloud registration. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 36, pp. 2198–2206 (2022)
42. Soh, W.K., Rajapakse, J.C.: Noise-induced self-supervised hybrid unet transformer for ischemic stroke segmentation with limited data annotations. *Scientific Reports* **15**(1), 19783 (2025)
43. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: *European conference on computer vision*. pp. 685–702. Springer (2020)
44. Tarubinga, E., Kalafatovich, J., Lee, S.W.: Farcluss: Fuzzy adaptive rebalancing and contrastive uncertainty learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2506.11142* (2025)
45. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* **11**(12) (2010)
46. Wang, F., Jiang, X., Galliani, S., Vogel, C., Pollefeys, M.: Glace: Global local accelerated coordinate encoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21562–21571 (2024)
47. Wu, Z., Zhang, J., Pai, D., Wang, X., Singh, C., Yang, J., Gao, J., Ma, Y.: Simplifying dino via coding rate regularization. *arXiv preprint arXiv:2502.10385* (2025)
48. Xiong, K., Xiang, H., Xu, Q., Wen, C., Shen, S., LI, J.J., Wang, C.: Mining and transferring feature-geometry coherence for unsupervised point cloud registration. *Advances in Neural Information Processing Systems* **37**, 35468–35491 (2024)
49. Xiong, K., Zheng, M., Xu, Q., Wen, C., Shen, S., Wang, C.: Speal: Skeletal prior embedded attention learning for cross-source point cloud registration. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 6279–6287 (2024)
50. Xu, H., Liu, S., Wang, G., Liu, G., Zeng, B.: Omnet: Learning overlapping mask for partial-to-partial point cloud registration. In: *Proceedings of the IEEE/CVF ICCV*. pp. 3132–3141 (2021)
51. Xu, Z., Gao, X., Jiang, X., Cheng, S., Zhang, Q., Li, W., Gao, X.: S2reg: Structure-semantics collaborative point cloud registration. *Pattern Recognition* **161**, 111290 (2025)

52. Yan, S., Shi, P., Li, J.: Ml-semreg: Boosting point cloud registration with multi-level semantic consistency. In: European Conference on Computer Vision. pp. 19–37. Springer (2024)
53. Yang, H., Dong, W., Carlone, L., Koltun, V.: Self-supervised geometric perception. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (2021)
54. Yew, Z.J., Lee, G.H.: Regtr: End-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF CVPR. pp. 6677–6686 (2022)
55. Yu, H., Li, F., Saleh, M., Busam, B., Ilic, S.: Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems* **34**, 23872–23884 (2021)
56. Yu, J., Ren, L., Zhang, Y., Zhou, W., Lin, L., Dai, G.: 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. pp. 17702–17711 (2023)
57. Yu, Y., Chan, K.H.R., You, C., Song, C., Ma, Y.: Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in neural information processing systems* **33**, 9422–9434 (2020)
58. Yuan, Z., Zeng, W., Su, Y., Liu, W., Cheng, M., Guo, Y., Wang, C.: Density-guided translator boosts synthetic-to-real unsupervised domain adaptive segmentation of 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23303–23312 (2024)
59. Zeng, Z., Wu, Q., Zhang, X., Wu, L.Y., An, P., Yang, J., Wang, J., Wang, P.: Unlocking generalization power in lidar point cloud registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22244–22253 (2025)
60. Zhang, X., Yang, J., Zhang, S., Zhang, Y.: 3d registration with maximal cliques. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17745–17754 (2023)
61. Zhang, Y., Zhao, H., Li, H., Chen, S.: Fastmac: Stochastic spectral sampling of correspondence graph. *arXiv preprint arXiv:2403.08770* (2024)
62. Zhao, G., Ao, S., Zhang, Y., Xu, K., Guo, Y.: Progressive correspondence regenerator for robust 3d registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1210–1219 (2025)
63. Zhao, G., Guo, Z., Ma, H.: Sgor: outlier removal by leveraging semantic and geometric information for robust point cloud registration. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9388–9395. IEEE (2024)
64. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the CVPR Conference. pp. 12786–12796 (2022)