TOKENI2P: LEARNING POSE-AWARE TOKENS FOR CORRESPONDENCE-FREE IMAGE-TO-POINT CLOUD REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-point cloud registration aims to recover the 6-DoF camera pose between a 2D image and a 3D point cloud, yet remains challenging due to their substantial modality gap. Existing methods, whether correspondence-based or correspondence-free, typically decompose registration into multiple stages, first constructing explicit correspondences or intermediate geometric proxies and subsequently recovering the pose. Such intermediate representations introduce additional complexity and, more critically, expose pose estimation to errors propagated from preceding stages. In this paper, we introduce **TokenI2P**, a single-stage, correspondence-free framework that reformulates 2D-3D registration as *pose-aware cross-modal compression*. At its core is a compact set of learnable *pose-aware tokens* that persistently interact with image and point cloud features, alternating between 2D appearance and 3D geometry to progressively accumulate pose-discriminative evidence. The resulting token state directly parameterizes the relative transformation, eliminating explicit correspondence construction, intermediate geometric proxies, and PnP/RANSAC-like pose solvers within a unified feed-forward architecture. To bias this compact representation toward shared cross-modal geometry, we further introduce dual-granularity mutual-information maximization. At the frame level, mutual-information maximization preserves globally shared scene information, while at the point level, geometry-guided mutual-information maximization strengthens fine-grained cross-modal dependencies between associated point-pixel features. Together, pose-aware token interaction determines *how* heterogeneous 2D-3D evidence is compressed, while mutual-information maximization guides *what* shared geometric information is preserved. Extensive experiments on KITTI and nuScenes show that TokenI2P achieves competitive registration accuracy with substantially reduced pipeline complexity and model size, while delivering the fastest inference among all compared methods, striking a favorable balance between accuracy and efficiency.

1 INTRODUCTION

Image-to-point cloud registration aims to estimate the 6-DoF camera pose (*i.e.*, 3D rotation and 3D translation) of a query image with respect to a 3D point cloud. It serves as a fundamental component in extensive downstream applications, such as visual localization Sarlin et al. (2023); Li et al. (2016), autonomous driving Rizk et al. (2020); Wolcott & Eustice (2014), and AR/VR Klein & Murray (2007); Sarlin et al. (2022). Despite its practical importance, this task remains highly challenging due to the substantial modality gap between 2D images and 3D point clouds: images encode dense appearance observations on a regular grid, whereas point clouds provide sparse and irregular geometric measurements in 3D space. Bridging such heterogeneous representations while preserving the geometric information required for accurate pose estimation remains a central challenge.

Existing methods largely follow a *correspondence-driven* paradigm. They first establish explicit point-to-pixel correspondences and then recover the camera pose using PnP/RANSAC or differentiable pose solvers Zhou et al. (2023); Feng et al. (2019); Ren et al. (2022); Li et al. (2023); Kang et al. (2024). This decomposition makes final pose estimation fundamentally dependent on correspondence quality. However, reliable cross-modal matching becomes particularly difficult under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

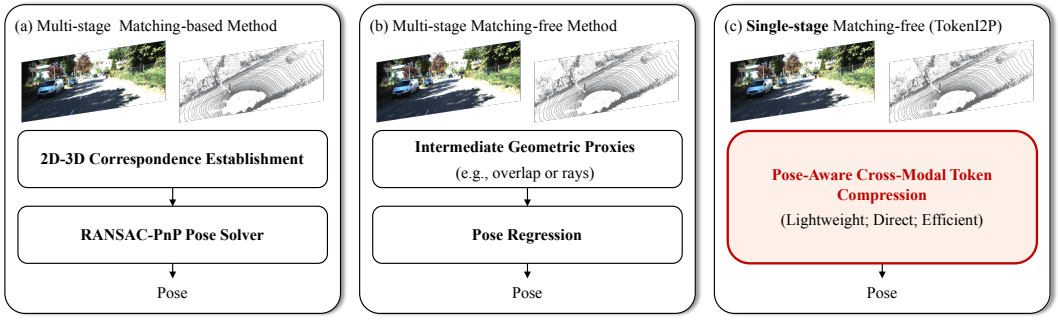

Figure 1: **Comparison of image-to-point cloud registration paradigms. (a)** Correspondence-based methods recover pose from explicit 2D-3D matches followed by RANSAC-PnP, whereas **(b)** correspondence-free methods typically predict intermediate geometric proxies (*e.g.*, overlap or rays) for subsequent pose recovery. Such multi-stage pipelines increase complexity and propagate intermediate errors to the final pose. By contrast, TokenI2P compresses heterogeneous 2D-3D observations into compact pose-aware tokens for direct *single-stage*, *correspondence-free* pose regression, eliminating intermediate geometric dependencies and streamlining registration.

limited overlap, appearance variations, and noisy observations. Erroneous correspondences thus propagate directly into the pose solver, often causing severe degradation in registration accuracy.

An appealing alternative is *correspondence-free* registration, avoiding explicit point-to-pixel matching. Nevertheless, existing correspondence-free methods largely retain a multi-stage design, replacing correspondences with *intermediate geometric proxies*. For instance, DeepI2P predicts camera-frustum relations followed by inverse-projection optimization Li & Lee (2021b); MaFreeI2P evaluates pose hypotheses through cross-modal cost volumes and progressive pose-space search Yao et al. (2024); and RayI2P predicts overlapping regions and patch-wise 3D rays before ray-guided pose regression Li et al. (2026). Although explicit correspondences are avoided, pose estimation remains conditioned on the accuracy of preceding intermediate predictions, leaving error propagation and pipeline complexity fundamentally unresolved.

This observation motivates a more direct question: *Can the relative pose be inferred directly from heterogeneous 2D/3D observations without explicitly establishing correspondences or intermediate geometric proxies?* To this end, we introduce **TokenI2P**, a single-stage, correspondence-free framework that reformulates image-to-point cloud registration as *pose-aware cross-modal compression*. At its core is a compact set of learnable *pose-aware tokens* that maintain a persistent latent state and alternately interact with image and point cloud features, progressively accumulating pose-discriminative evidence from 2D appearance and 3D geometry. The resulting token state directly parameterizes the relative pose, integrating cross-modal interaction, feature aggregation, and pose regression within a unified feed-forward architecture. Unlike existing pipelines, TokenI2P requires neither explicit correspondence construction nor intermediate geometric proxies, eliminating subsequent PnP/RANSAC or iterative pose optimization.

A compact token bottleneck, however, inevitably limits representational capacity, making *what to preserve* as important as *how to compress*. For relative pose estimation, the most informative cross-modal evidence lies in the scene structure jointly observed by the image and point cloud. We therefore introduce dual-granularity mutual-information maximization to bias the learned representations toward information shared across the two modalities. At the frame level, mutual-information maximization preserves globally shared scene information between the image and its visible point cloud region. At the point level, geometry-guided mutual-information maximization strengthens fine-grained cross-modal dependencies between geometrically associated point-pixel features. Together, pose-aware token interaction determines *how* heterogeneous 2D-3D evidence is compressed, while mutual-information maximization guides *what* shared geometric information is preserved.

Despite its streamlined design, TokenI2P achieves competitive registration accuracy on KITTI and nuScenes with substantially reduced pipeline complexity and model size, while delivering the fastest inference among all compared methods. These results suggest that accurate image-to-point cloud registration does not necessarily require explicit cross-modal correspondences or handcrafted in-

termediate geometric interfaces, but can instead emerge from compact, pose-aware cross-modal representations. To summarize, our main contributions are as follows:

- A Novel Single-Stage Paradigm: We propose TokenI2P, which unifies image-to-point cloud registration into an efficient, single-stage end-to-end feed-forward paradigm, eliminating multi-stage error accumulation and heavy pipeline complexity inherent in mainstream correspondence-based or correspondence-free methods.
- Pose-Aware Cross-Modal Compression: We introduce a compact set of role-specialized pose-aware tokens that persistently alternate between 2D appearance and 3D geometry, progressively compressing cross-modal evidence into a latent representation for direct pose regression.
- Mutual-Information Guided Cross-Modal Representation Learning: We introduce frame-wise and point-wise mutual-information maximization to guide pose-aware compression toward shared cross-modal geometry, preserving global scene consistency and fine-grained geometric dependencies for pose reasoning.

2 Related Work

Matching-based Image-to-Point Cloud Registration. Matching-based methods estimate camera poses through cross-modal image–point cloud associations. 2D3D-MatchNet Feng et al. (2019) learns descriptors for sparse 2D and 3D keypoint matching, while P2-Net Wang et al. (2021) jointly learns keypoint detection and feature description in a shared latent space. CorrI2P Ren et al. (2022) estimates dense correspondences within the detected overlap and recovers the pose using EPnP-RANSAC. VP2P-Match Zhou et al. (2023) combines voxel and point representations with differentiable probabilistic PnP, whereas 2D3D-MATR Li et al. (2023) uses coarse-to-fine patch and pixel-to-point matching followed by PnP-RANSAC. Later work improves correspondence-based registration through pretrained features, correspondence refinement, differentiable optimization, and geometric constraints Kang et al. (2024); Wang et al. (2024); Wu et al. (2024); Mu et al. (2025); An et al. (2025); Cheng et al. (2026); An et al. (2026). GraphI2P Bie et al. (2025) lifts image observations into virtual 3D points and refines correspondences through graph-based consistency. ICL Li et al. (2025) uses overlap detection and learnable correspondence queries to generate paired 2D-3D keypoints for pose regression. Despite these advances, pose estimation remains sensitive to the quality of intermediate correspondences.

Matching-free Image-to-Point Cloud Registration. Matching-free methods estimate relative pose without explicitly constructing point-to-pixel correspondences. CMRNet Cattaneo et al. (2019) jointly processes an RGB image and a depth map projected from LiDAR using an initial pose to regress a pose correction. DeepI2P Li & Lee (2021b) uses camera-frustum classification followed by inverse-projection optimization. MaFreeI2P Yao et al. (2024) evaluates pose hypotheses with cross-modal cost volumes, while RayI2P Li et al. (2026) estimates pose from patch-wise 3D ray representations. Despite avoiding explicit correspondences, these methods still rely on intermediate geometric representations or procedures. In contrast, TokenI2P uses learnable pose-aware tokens to capture cross-modal information and directly regresses the relative pose in a single stage.

3 Method

3.1 Overview of TokenI2P

We design a simple transformer-based encoder-decoder architecture for direct image-to-point cloud registration. As shown in Fig. 2, TokenI2P consists of two modality-specific perception encoders and a pose-aware token decoder. The model directly maps 2D/3D observations to their relative camera pose without explicit point-pixel correspondences or intermediate geometric representations.

The perception encoders (Sec. 3.2) independently extract dense image and point cloud features, which are projected into a common feature space and equipped with modality-specific positional encodings. To bridge the two modalities, we introduce a compact set of role-specialized pose-aware tokens that maintain a persistent latent state throughout the decoder (Sec. 3.3). By alternately attending to image and point cloud features, these tokens progressively compress complementary 2D appearance and 3D geometric evidence into a compact pose representation, which is directly decoded into the relative camera pose.

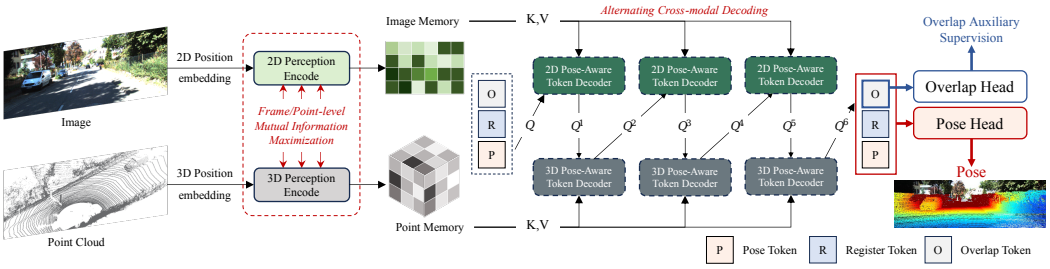

Figure 2: **Overview of the TokenI2P pipeline.** Modality-specific encoders produce image and point cloud feature memories. Pose-aware tokens alternately attend to these memories, and the resulting tokens are jointly aggregated for direct relative pose regression.A training-only overlap head predicts point-wise co-visibility to provide auxiliary supervision.

.

During training, we further introduce mutual-information guided cross-modal representation learning (Sec. 3.4), which maximizes cross-modal mutual information at complementary frame and point levels to guide the encoded features toward shared cross-modal geometry. Frame-wise mutual information preserves globally shared scene information, while point-wise mutual information strengthens fine-grained geometric dependencies across the two modalities.

3.2 2D/3D Perception Encoders

At the perception level, we seek to fulfill two objectives: 1) preserving the modality-specific spatial structures of the image and point cloud; and 2) providing spatially grounded feature memories for subsequent cross-modal pose reasoning. To this end, we employ dedicated 2D and 3D perception encoders to extract dense appearance and geometric features, respectively, and augment them with their native spatial coordinates before feeding them to the pose-aware token decoder.

2D Perception Encoder. Given an RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we employ a residual convolutional network (He et al., 2016) to extract a dense feature map $\mathbf{F}_I \in \mathbb{R}^{H'\times W'\times C_I}$. The feature map is projected to a D-dimensional representation and flattened into $N_I = H'W'$ image tokens. Since the spatial arrangement of image features provides essential cues for cross-modal geometric reasoning, we retain their locations using a learnable 2D positional encoding. Specifically, for each image feature $\mathbf{f}_i^I$ at normalized image coordinate $\mathbf{u}_i \in \mathbb{R}^2$, we construct

$$\mathbf{x}_i^I = \phi_I(\mathbf{f}_i^I) + \psi_I(\mathbf{u}_i), \tag{1}$$

where ϕ_I denotes the feature projection and ψ_I the 2D positional encoder. The resulting image memory is denoted by $\mathbf{X}_I = \{\mathbf{x}_1^I, ..., \mathbf{x}_{N_I}^I\}$.

3D Perception Encoder. Given a point cloud $\mathbf{P} = \{\mathbf{p}_j \in \mathbb{R}^3 \mid j = 1, ..., N\}$, we use a hierarchical KPConv network (Thomas et al., 2019) to progressively aggregate local geometric context and obtain a subsampled set of point features $\mathbf{F}_P = \{\mathbf{f}_1^P, ..., \mathbf{f}_{N_p}^P\}$ $(N_P \ll N)$. Analogous to the image branch, each point feature is projected to the same D-dimensional feature dimension, and we preserve its explicit geometric location through a learnable 3D positional encoding as follows,

$$\mathbf{x}_j^P = \phi_P(\mathbf{f}_j^P) + \psi_P(\mathbf{p}_j), \tag{2}$$

where ϕ_P and ψ_P denote the point-feature projection and 3D positional encoder, respectively. This yields the point cloud memory $\mathbf{X}_P = \{\mathbf{x}_1^P, ..., \mathbf{x}_{N_P}^P\}$. Rather than prematurely aligning the two modalities into a single feature representation, the two encoders preserve their native 2D and 3D spatial structures while exposing them through a common feature dimension. The resulting $\mathbf{X}_I$ and $\mathbf{X}_P$ therefore serve as separate spatially grounded memories, from which the pose-aware token decoder progressively retrieves cross-modal evidence for pose estimation.

3.3 Pose-Aware Token Decoder

Given the encoded image and point cloud memories $\mathbf{X}_I$ and $\mathbf{X}_P$, we employ a pose-aware token decoder to integrate heterogeneous 2D appearance and 3D geometric evidence for direct pose estimation. Instead of explicitly constructing point-pixel correspondences, the decoder maintains a

compact latent state that is progressively refined through interactions with the two modalities. To enable effective cross-modal pose reasoning, we equip this latent state with role-specialized tokens and update them through alternating 2D-3D decoding.

Role-specialized Latent Tokens. Inferring a relative pose from heterogeneous 2D and 3D observations requires reasoning over complementary cross-modal evidence. Rather than compressing all such evidence into a single latent token, we maintain a compact set of *role-specialized latent tokens* that capture different aspects of the pose-relevant information. Specifically, the latent state consists of a pose token $\mathbf{q}_{\text{pose}}$, N_r register tokens $\mathbf{Q}_{\text{reg}}$, and N_o overlap tokens $\mathbf{Q}_{\text{ov}}$:

$$\mathbf{Q}^0 = [\,\mathbf{q}^0_{\text{pose}};\, \mathbf{Q}^0_{\text{reg}};\, \mathbf{Q}^0_{\text{ov}}\,], \qquad N_q = 1 + N_r + N_o, \tag{3}$$

where $\mathbf{Q}^0 \in \mathbb{R}^{N_q \times D}$. The three token groups provide complementary representations for pose reasoning. The *pose token* maintains a pose-centric latent state that is progressively refined through cross-modal interaction. The *register tokens* aim to provide additional latent capacity for accumulating global contextual evidence across the image and point cloud (Darcet et al., 2024). The *overlap tokens* are explicitly supervised to identify 3D regions co-visible with the image, encouraging them to preserve fine-grained cross-modal geometric evidence (as demonstrated below). Importantly, these roles are complementary rather than independent: all token groups are jointly updated throughout the alternating 2D-3D decoder and are collectively aggregated for final pose regression.

Alternating Cross-modal Decoding. To jointly refine the role-specialized tokens with both appearance and geometric evidence, we propagate the unified latent state through L transformer blocks that alternately attend to the image and point cloud feature memories. Each block consists of self-attention among the latent tokens, cross-attention to the corresponding modality, and a feed-forward network, all with residual connections. Specifically, the latent state at the l-th block is updated as

$$\mathbf{Q}^l = \mathcal{D}_{m_l}\left(\mathbf{Q}^{l-1}, \mathbf{X}_{m_l}\right), \qquad m_l \in \{I, P\}, \tag{4}$$

where $\mathcal{D}_{m_l}$ denotes a transformer block conditioned on modality m_l. By alternating m_l between the image and point cloud, the latent state evolves as:

$$\mathbf{Q}^0 \xrightarrow{\mathbf{X}_I} \mathbf{Q}^1 \xrightarrow{\mathbf{X}_P} \mathbf{Q}^2 \xrightarrow{\mathbf{X}_I} \cdots \xrightarrow{\mathbf{X}_P} \mathbf{Q}^L. \tag{5}$$

Importantly, the updated latent state is carried from one modality to the next, such that evidence acquired from the image can condition subsequent interaction with the point cloud, and vice versa. This persistent alternating process allows the tokens to progressively integrate complementary 2D appearance and 3D geometric evidence into a shared latent representation for pose estimation.

Overlap-aware Geometric Grounding. Alongside the joint cross-modal decoding above, the overlap tokens are further grounded with fine-grained geometric supervision, explicitly encouraging them to focus on co-visible regions that provide direct geometric evidence for relative pose estimation. Specifically, after the final decoder block, each encoded point feature attends to the *overlap tokens* $\mathbf{Q}^L_{\text{ov}}$ to retrieve an overlap-aware cross-modal context:

$$\mathbf{C}_{\text{ov}} = \mathcal{A}_{\text{ov}}\left(\mathbf{X}_P, \mathbf{Q}^L_{\text{ov}}\right), \tag{6}$$

where $\mathcal{A}_{\text{ov}}$ denotes cross-attention with the point features as queries and the overlap tokens as keys and values. The resulting $\mathbf{C}_{\text{ov}}$ provides each 3D point with cross-modal context for reasoning about its co-visibility with the image. We then concatenate the retrieved context with the corresponding point representation and 3D positional embedding to predict the point-wise overlap probability:

$$\hat{\mathbf{o}}_{\text{ov}} = \sigma\left(H_{\text{ov}}\left([\,\mathbf{X}_P; \mathbf{C}_{\text{ov}}; \mathbf{E}_P\,]\right)\right), \qquad \mathcal{L}_{\text{ov}} = \text{BCE}(\hat{\mathbf{o}}_{\text{ov}}, \mathbf{o}_{\text{ov}}), \tag{7}$$

where H_{ov} denotes a lightweight overlap prediction head, $\sigma(\cdot)$ is a sigmoid function and $\mathbf{o}_{\text{ov}}$ denotes the ground-truth point-wise overlap labels. The overlap labels are derived from the ground-truth relative pose by determining whether each 3D point is co-visible with the input image. This auxiliary supervision explicitly grounds the overlap tokens in cross-modal co-visibility, complementing global pose supervision with localized geometric evidence. Importantly, the overlap tokens themselves contribute to the final pose representation, whereas overlap prediction is used only as an auxiliary training signal and does not participate in inference.

Pose Readout and Supervision. After alternating cross-modal decoding, the role-specialized tokens encode complementary pose-relevant information. We aggregate the pose token with the mean-pooled register and overlap tokens to form a unified pose representation for direct pose regression:

$$\hat{\mathbf{T}} = (\hat{\mathbf{R}}, \hat{\mathbf{t}}) = H_{\text{pose}}\big(\big[\mathbf{q}^L_{\text{pose}};\, \phi_m(\mathbf{Q}^L_{\text{reg}});\, \phi_m(\mathbf{Q}^L_{\text{ov}})\big]\big), \tag{8}$$

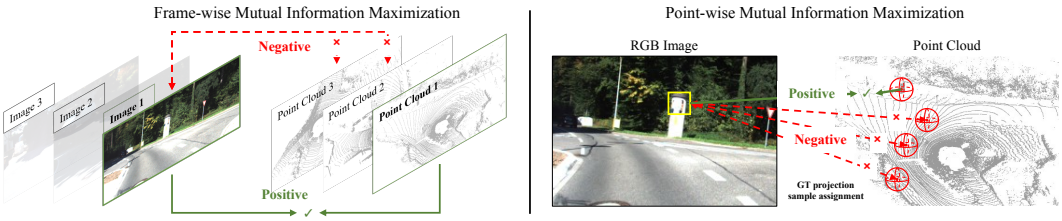

Figure 3: **Frame-wise and point-wise mutual-information maximization.** At the frame level, paired image-point cloud observations form positive samples, while cross-pair combinations form negative samples. At the point level, ground-truth pose projection establishes the geometric correspondences used to define positive and negative 2D-3D feature pairs.

where $\phi_m(\cdot)$ denotes mean pooling and H_{pose} is a lightweight MLP. The predicted pose is supervised by separate rotation and translation objectives as follows:

$$\mathcal{L}_{\text{pose}} = \|\hat{\mathbf{R}} - \mathbf{R}\|_2 + \lambda_t \|\hat{\mathbf{t}} - \mathbf{t}\|_2, \tag{9}$$

where $\mathbf{R}$ and $\mathbf{t}$ denote the ground-truth relative rotation and translation, respectively, and λ_t balances the rotation and translation objectives. By leveraging learned pose-aware token representations, including pose-centric, contextual, and overlap-aware tokens, the model directly infers the relative pose without explicit correspondence construction or geometric optimization.

3.4 Cross-Modal Mutual Information Maximization

While the pose-aware token decoder enables explicit interaction between image and point cloud representations, more efficient pose reasoning also potentially relies on the two modalities to preserve information that is mutually informative about their shared scene geometry. We therefore formulate cross-modal representation learning from a mutual-information perspective and maximize shared information at two complementary granularities, as illustrated in Fig. 3. At the frame level, we maximize mutual information between global image and point cloud representations to capture scene-level consistency. At the point level, we exploit ground-truth geometry during training to maximize mutual information between geometrically associated local 2D and 3D features.

Frame-wise Mutual Information Maximization. Relative pose is determined by the scene evidence jointly observed by the image and point cloud. At the frame level, we therefore seek to preserve as much information as possible that is shared between their global representations. Mutual information provides a natural information-theoretic measure of such shared information, quantifying the statistical dependence between two representations. This motivates us to maximize the mutual information between the global 2D and 3D representations, thereby strengthening the shared scene-level evidence available to the pose-aware token decoder for cross-modal pose reasoning.

Formally, let $\mathbf{g}_I$ and $\mathbf{g}_P$ denote the global 2D and 3D representations, respectively. Their mutual information is defined and lower-bounded by InfoNCE (Oord et al., 2018) as follows:

$$I(\mathbf{g}_I; \mathbf{g}_P) = \mathbb{E}_{p(\mathbf{g}_I, \mathbf{g}_P)} \left[\log \frac{p(\mathbf{g}_I \mid \mathbf{g}_P)}{p(\mathbf{g}_I)} \right] \geq \log B + \mathbb{E} \left[\log \frac{f(\mathbf{g}_I^+, \mathbf{g}_P)}{\sum_{j=1}^{B} f(\mathbf{g}_I^j, \mathbf{g}_P)} \right], \tag{10}$$

where the compatibility function $f(\cdot, \cdot)$ estimates the density ratio $p(\mathbf{g}_I|\mathbf{g}_P)/p(\mathbf{g}_I)$. We instantiate it as $f(\mathbf{g}_I, \mathbf{g}_P) = \exp(s(\mathbf{g}_I, \mathbf{g}_P)/\tau)$, where $s(\cdot, \cdot)$ denotes cosine similarity and τ is a learnable temperature. Here, B denotes the number of contrastive samples, consisting of one positive representation $\mathbf{g}_I^+$ paired with $\mathbf{g}_P$ and $B-1$ negative representations from mismatched samples.

In the context of the image-to-point cloud registration, this contrastive construction arises naturally from the co-visibility between the two modalities. Given a mini-batch of B paired image-point cloud training samples $\{(\mathbf{I}_i, \mathbf{P}_i)\}_{i=1}^{B}$, each paired observation $(\mathbf{I}_i, \mathbf{P}_i)$ contains a shared field of view, whereas cross-sample pairs $(\mathbf{I}_i, \mathbf{P}_j)$, $i \neq j$, are generally non-covisible. We therefore treat the global representations of paired observations as positives and those of mismatched observations as negatives. To construct frame-level global representations, we aggregate the encoded image features using learnable attention pooling, yielding a global image representation $\mathbf{g}_I$. For the point

cloud, we first identify the points visible in the image under the ground-truth pose and average their encoded features to obtain the corresponding global point cloud representation $\mathbf{g}_P$. The resulting representations therefore summarize the scene content jointly observed by the two modalities.

By substituting these samples into the InfoNCE estimator, maximizing the lower bound on the frame-wise mutual information amounts to minimizing the following frame-wise objective:

$$\mathcal{L}_{\mathrm{FMI}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(s(\mathbf{g}_I^i,\mathbf{g}_P^i)/\tau\big)}{\sum_{j=1}^{B}\exp\Big(s(\mathbf{g}_I^j,\mathbf{g}_P^i)/\tau\Big)} - \frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(s(\mathbf{g}_P^i,\mathbf{g}_I^i)/\tau\big)}{\sum_{j=1}^{B}\exp\Big(s(\mathbf{g}_P^j,\mathbf{g}_I^i)/\tau\Big)}. \tag{11}$$

These two symmetric terms optimize the contrastive objective in opposite directions, using the image and point cloud representations alternately as queries, encouraging consistent cross-modal dependence from both modalities.

Point-wise Mutual Information Maximization. While frame-wise mutual information promotes globally shared information between the two modalities, accurate pose estimation further requires fine-grained geometric dependencies between local 2D and 3D features. We therefore extend the same information-maximization principle to the point level, encouraging geometrically associated image and point features to preserve their local cross-modal dependence.

Specifically, given a 3D point $\mathbf{p}_j$ visible in the image, we use the ground-truth relative pose to project it onto the image plane and obtain its corresponding image location $\pi(\mathbf{p}_j)$. Let $\mathbf{f}_j^P$ and $\mathbf{f}_{\pi(j)}^I$ denote the encoded point feature and its geometrically associated image feature, respectively. Their point-wise mutual information and the corresponding InfoNCE lower bound can be expressed as

$$I(\mathbf{f}^I;\mathbf{f}^P) = \mathbb{E}_{p(\mathbf{f}^I,\mathbf{f}^P)}\left[\log\frac{p(\mathbf{f}^I\mid\mathbf{f}^P)}{p(\mathbf{f}^I)}\right] \geq \log N_v + \mathbb{E}\left[\log\frac{f(\mathbf{f}_{\pi(j)}^I,\mathbf{f}_j^P)}{\sum_{k=1}^{N_v} f(\mathbf{f}_{\pi(k)}^I,\mathbf{f}_j^P)}\right], \tag{12}$$

where N_v denotes the number of valid point–image associations. Here, $(\mathbf{f}_j^P,\mathbf{f}_{\pi(j)}^I)$ forms the positive pair, whereas $\{(\mathbf{f}_j^P,\mathbf{f}_{\pi(k)}^I)\}_{k\neq j}$ serve as negative pairs. Following the frame-wise objective, we instantiate the compatibility function as $f(\mathbf{f}^P,\mathbf{f}^I)=\exp(s(\mathbf{f}^P,\mathbf{f}^I)/\tau_p)$. Since different local associations exhibit varying degrees of ambiguity, we further assign each positive pair an entropy-guided weight w_j according to its cross-modal similarity distribution. By substituting these samples into the InfoNCE estimator, maximizing the lower bound on the point-wise mutual information amounts to minimizing the following point-wise objective:

$$\mathcal{L}_{\mathrm{PMI}} = -\sum_{j=1}^{N_v} w_j\log\frac{\exp\big(s(\mathbf{f}_{\pi(j)}^I,\mathbf{f}_j^P)/\tau_p\big)}{\sum_{k=1}^{N_v}\exp\Big(s(\mathbf{f}_{\pi(k)}^I,\mathbf{f}_j^P)/\tau_p\Big)} - \sum_{j=1}^{N_v} w_j\log\frac{\exp\big(s(\mathbf{f}_j^P,\mathbf{f}_{\pi(j)}^I)/\tau_p\big)}{\sum_{k=1}^{N_v}\exp\Big(s(\mathbf{f}_k^P,\mathbf{f}_{\pi(j)}^I)/\tau_p\Big)}. \tag{13}$$

where we optimize symmetrically in both point-to-image and image-to-point directions, and w_j denotes the entropy-guided importance weight for the j-th sample. Specifically, we define the cross-modal matching distribution as $p_{jk}=\mathrm{softmax}_k(s(\mathbf{f}_j^P,\mathbf{f}_{\pi(k)}^I)/\tau_p)$ and measure its ambiguity by the entropy $H_j=-\sum_k p_{jk}\log p_{jk}$. We then compute $w_j\propto\exp([\alpha H_j+(1-\alpha)(1-p_{jj})]/\tau_w)$, where H_j characterizes the ambiguity of the cross-modal matching distribution, $1-p_{jj}$ measures the difficulty of the positive association, and α balances their contributions. The weights are normalized over all valid associations, such that ambiguous and difficult point-image pairs receive greater emphasis during point-wise mutual information maximization.

4 Experiments

Implementation Details. TokenI2P is trained for 25 epochs on two RTX 4090 GPUs with a batch size of 2. The four-stage image and point cloud encoders use channels $[512,256,128,128]$ and $[1024,512,256,128]$, respectively, with a base voxel size of $0.15\,\mathrm{m}$. The 256-dimensional query bridge comprises six Transformer layers with eight heads each, using one pose token, four register tokens, 32 overlap tokens, and 0.1 dropout. Scene- and local-level MI losses are weighted by 0.05 and 0.1, respectively. Adam uses an initial learning rate of 10^{-4}, weight decay of 10^{-6}, and a per-epoch decay factor of 0.95.

Table 1: Registration accuracy on KITTI and nuScenes. The Stage column distinguishes multi-stage and single-stage pose-estimation pipelines. † denotes methods using ZoeDepth Bhat et al. (2023) for monocular depth estimation.

Category	Stage	Method	KITTI			nuScenes		
			RTE (m)↓	RRE (°)↓	Acc (%)↑	RTE (m)↓	RRE (°)↓	Acc (%)↑
Matching-based	Multi	CorrI2P	3.78 ± 65.16	5.89 ± 20.34	72.42	3.04 ± 60.76	3.73 ± 9.03	49.00
		VP2P-Match	0.75 ± 1.13	3.29 ± 7.99	83.04	0.89 ± 1.44	2.15 ± 7.03	88.33
		FreeReg†	0.95 ± 1.05	2.06 ± 3.21	91.68	–	–	–
		CoFiI2P	0.31 ± 0.20	1.24 ± 0.84	–	1.21 ± 9.55	2.54 ± 8.93	–
		ICL	0.20 ± 0.21	1.24 ± 2.34	97.49	0.63 ± 0.44	2.13 ± 3.75	90.94
		GraphI2P†	0.32 ± 0.81	1.65 ± 1.32	99.61	0.49 ± 1.22	1.73 ± 1.63	**99.48**
Matching-free	Multi	Grid Cls. + PnP	3.64 ± 3.46	19.19 ± 28.96	11.22	3.02 ± 2.40	12.66 ± 21.01	2.45
		DeepI2P (3D)	4.06 ± 3.54	24.73 ± 31.69	3.77	2.88 ± 2.12	20.65 ± 12.24	2.26
		DeepI2P (2D)	3.59 ± 3.21	11.66 ± 18.16	25.95	2.78 ± 1.99	4.80 ± 6.21	38.10
		RayI2P	$\mathbf{0.09 \pm 0.08}$	0.63 ± 0.71	99.75	$\mathbf{0.39 \pm 0.29}$	1.48 ± 1.72	96.61
	Single	**Ours**	0.11 ± 0.09	$\mathbf{0.44 \pm 0.36}$	**100**	0.61 ± 0.51	$\mathbf{1.24 \pm 1.39}$	97.27

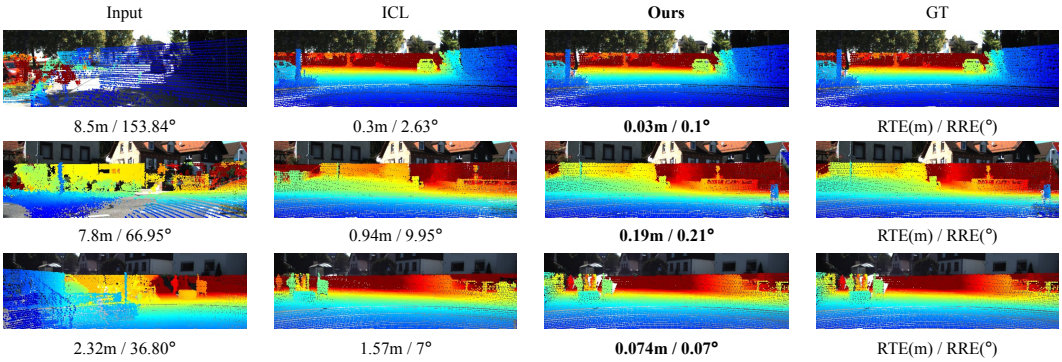

Figure 4: Qualitative comparison of registration results.

Metrics. Following VP2P-Match Zhou et al. (2023), we report mean Relative Translation Error (RTE), mean Relative Rotation Error (RRE), and registration accuracy (Acc). RTE and RRE are computed over all test pairs, including failed registrations; Acc is the proportion of pairs with RTE below $2\,\mathrm{m}$ and RRE below 5°.

4.1 Comparison with Existing Methods

We evaluate the proposed method on two widely used outdoor image-to-point cloud registration benchmarks, KITTI Odometry Geiger et al. (2012) and nuScenes Caesar et al. (2020). We compare our method with representative image-to-point cloud registration approaches from both matching-based and matching-free paradigms. The matching-based methods include CorrI2P Ren et al. (2022), VP2P-Match Zhou et al. (2023), FreeReg Wang et al. (2024), CoFiI2P Kang et al. (2024), ICL Li et al. (2025), and GraphI2P Bie et al. (2025). The matching-free methods include Grid Cls. + PnP and DeepI2P Li & Lee (2021a) and RayI2P Li et al. (2026).

Quantitative Comparison. Table 1 reports the quantitative results on KITTI and nuScenes. Among all compared methods, TokenI2P is the only single-stage approach that directly regresses the relative pose from cross-modal observations. Despite this streamlined inference pipeline, it achieves the lowest RRE (0.44°) and the highest registration accuracy (100%) on KITTI, with an RTE of $0.11\,\mathrm{m}$, close to the best result of $0.09\,\mathrm{m}$. On nuScenes, TokenI2P again achieves the lowest RRE (1.24°) while maintaining a high registration accuracy of 97.27%.

Table 2: Efficiency comparisons.

Method	Model Size (M)	Inference Time (s)
DeepI2P (2D) Li & Lee (2021a)	100.12	23.47
DeepI2P (3D) Li & Lee (2021a)	100.12	35.61
CorrI2P Ren et al. (2022)	141.07	8.96
VP2P-Match Zhou et al. (2023)	30.73	0.19
CoFiI2P Kang et al. (2024)	39.09	0.18
ICL Li et al. (2025)	45.21	0.24
RayI2P Li et al. (2026)	49.24	0.11
Ours	**25.75**	**0.10**

Efficiency Comparison. Table 2 shows that TokenI2P has the smallest model (25.75M parameters) and the fastest inference (0.10 s) among the compared methods. Its single-stage token-based pose regression avoids separate correspondence estimation and pose solving, yielding a compact inference pipeline with lower computational overhead and reduced latency in practical deployment.

Qualitative Comparison. Fig. 4 compares the input projection, ICL, TokenI2P, and ground truth across three road scenes. We project the depth-colored point cloud into the image using the corresponding pose and known calibration parameters. While ICL largely corrects the initial misalignment, residual offsets remain around geometric boundaries, particularly where roadside structures meet the road surface and around nearby vertical objects. TokenI2P produces projections that more closely follow the ground truth in these regions, with better-aligned road–structure boundaries and local object contours across the three examples. This visual improvement is consistent with its lower registration error in the quantitative comparison.

4.2 ABLATION STUDY

Ablation on Mutual Information Losses. Table 3 shows that removing both MI objectives increases RTE and RRE to 0.30 m and 1.08°, whereas either objective alone improves registration. Using both yields the best results of 0.11 m and 0.44°, supporting the complementary roles of scene-level consistency and local geometric alignment.

Effects of Role-Specialized Tokens. We evaluate the contribution of the three role-specialized token groups in the pose-aware decoder by removing each of them individually. As shown in Table 4, removing any token group degrades the registration performance. In particular, the largest degradation is observed without the overlap tokens, indicating the importance of preserving fine-grained geometric evidence from co-visible regions. The pose token and register tokens also consistently improve the results by maintaining pose-centric information and complementary global cross-modal context, respectively. These results demonstrate the complementary roles of the three token groups in pose estimation.

Table 3: Ablation of MI objectives on KITTI.

Scene-level	Local-level	RTE (m)↓	RRE (°)↓
✗	✗	0.30±0.21	1.08±1.19
✓	✗	0.24±0.16	0.72±0.66
✗	✓	0.27±0.15	0.76±0.62
✓	✓	**0.11±0.09**	**0.44±0.36**

Table 4: Effect of each token group on KITTI.

Method	RTE (m)↓	RRE (°)↓	Acc (%)↑
w/o Pose	0.14±0.11	0.59±0.52	99.89
w/o Register	0.28±0.18	1.21±1.08	99.73
w/o Overlap	0.20±0.12	0.61±0.5	99.91
Full	**0.11±0.09**	**0.44±0.36**	**100.00**

Table 5: Ablation on token configurations on KITTI.

(N_p, N_r, N_q)	RTE (m)↓	RRE (°)↓	Acc (%)↑
$(1, 2, 16)$	0.23±0.14	0.85±0.7	99.91
$(\mathbf{1}, \mathbf{4}, \mathbf{32})$	**0.11±0.09**	**0.44±0.36**	**100.00**
$(2, 8, 64)$	0.41±0.26	1.51±1.39	98.76

Ablation on Token Configuration. We further investigate the effect of token configuration by jointly varying the numbers of pose, register, and overlap tokens. As shown in Table 5, $(1, 4, 32)$ achieves the best performance, with an RTE of 0.11 m and an RRE of 0.44°. A smaller configuration $(1, 2, 16)$ yields higher errors, whereas increasing the token numbers to $(2, 8, 64)$ also results in degraded performance. This indicates that an appropriate latent capacity is important for pose estimation, while simply introducing more tokens does not necessarily improve the representation. We therefore adopt $(1, 4, 32)$ as the default configuration.

5 CONCLUSION

We presented TokenI2P, a single-stage image-to-point cloud registration framework that estimates pose from compact, role-specialized tokens without explicit matching or pose solving at inference. Alternating 2D-3D decoding integrates appearance and geometry, while mutual-information objectives and overlap supervision strengthen cross-modal representations during training. Experiments on KITTI and nuScenes demonstrate that TokenI2P achieves strong registration performance with the smallest model and fastest inference among the compared methods. These results support direct token-based pose reasoning as an accurate and efficient alternative to multi-stage registration. Future work will improve translation estimation and evaluate generalization beyond driving scenes.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing of the manuscript. The tools were used to improve grammar, clarity, and readability without altering the technical content, methodology, experimental results, or scientific claims. All AI-assisted revisions were carefully reviewed by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on image-to-point cloud registration for computer vision and robotic perception. The study is conducted entirely on publicly available benchmark datasets and does not involve human-subject experiments, private user information, or the collection of personally identifiable data. We use the datasets only for research purposes and follow their corresponding terms of use and licensing requirements.

The proposed method is designed for geometric perception and sensor registration. While such techniques may be incorporated into broader autonomous or robotic systems, this work does not directly address deployment in safety-critical environments. Any practical use should therefore consider application-specific reliability, failure modes, and system-level safety requirements. We report the experimental protocol and evaluation criteria to facilitate reproducibility and transparent assessment of the method. The authors take responsibility for the ethical conduct and presentation of this research.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed architecture, training configuration, loss functions, dataset construction, and evaluation protocol in the main paper and supplementary material. The datasets used in our experiments are publicly available, and the data splits and preprocessing settings are specified to enable consistent evaluation. We also report the key implementation details and hyperparameters required to reproduce the experiments. To further support reproducibility, we plan to release the source code, trained model weights, configuration files, and data preparation instructions upon publication. These resources are intended to enable independent verification of the reported results and facilitate future research based on our method.

REFERENCES

Pei An, Jiaqi Yang, Muyao Peng, You Yang, Qiong Liu, Xiaolin Wu, and Liangliang Nan. Mincd-pnp: Learning 2d-3d correspondences with approximate blind pnp. In *ICCV*, 2025.

Pei An, Junfeng Ding, Jiaqi Yang, Yulong Wang, Jie Ma, and Liangliang Nan. Hg-i2p: Bridging modalities for generalizable image-to-point-cloud registration via heterogeneous graphs. In *CVPR*, 2026.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Lin Bie, Shouan Pan, Siqi Li, Yining Zhao, and Yue Gao. GraphI2P: Image-to-point cloud registration with exploring pattern of correspondence via graph learning. In *CVPR*, 2025.

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 *CVPR*, 2020.

Daniele Cattaneo, Matteo Vaghi, Augusto Luis Ballardini, Simone Fontana, Domenico Giorgio Sorrenti, and Wolfram Burgard. CMRNet: Camera to lidar-map registration. In *ITSC*, 2019.

Zhixin Cheng, Bohao Liao, Jiacheng Deng, Xiaotian Yin, Xinjun Li, Yujia Chen, Baoqun Yin, and Tianzhu Zhang. Rethinking 2d-3d registration: A novel network for high-value zone selection and representation consistency alignment. In *CVPR*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

Mengdan Feng, Sixing Hu, Marcelo H Ang, and Gim Hee Lee. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *ICRA*, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Shuhao Kang, Youqi Liao, Jianping Li, Fuxun Liang, Yuhao Li, Xianghong Zou, Fangning Li, Xieyuanli Chen, Zhen Dong, and Bisheng Yang. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. *IEEE Robotics and Automation Letters*, 2024.

Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, 2007.

Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *CVPR*, 2021a.

Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. *arXiv preprint arXiv:2104.03501*, 2021b.

Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, and Kai Xu. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *ICCV*, 2023.

Xinjun Li, Wenfei Yang, Jiacheng Deng, Zhixin Cheng, Xu Zhou, and Tianzhu Zhang. Implicit correspondence learning for image-to-point cloud registration. In *CVPR*, 2025.

Xinjun Li, Wenfei Yang, Zhixin Cheng, Jiacheng Deng, Fei Wang, Chen Qian, and Tianzhu Zhang. Rayi2p: Learning rays for image-to-point cloud registration. In *International Conference on Learning Representations*, 2026.

Yunpeng Li, Noah Snavely, Daniel P Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *Large-Scale Visual Geo-Localization*. 2016.

Juncheng Mu, Chengwei Ren, Weixiang Zhang, Liang Pan, Xiao-Ping Zhang, and Yue Gao. Diff2I2P: Differentiable image-to-point cloud registration with diffusion prior. In *ICCV*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Siyu Ren, Yiming Zeng, Junhui Hou, and Xiaodong Chen. Corri2p: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

Mostafa Rizk, Ahmad Mroue, Mohammad Farran, and Jamal Charara. Real-time slam based on image stitching for autonomous navigation of uavs in gnss-denied regions. In *AICAS*, 2020.

Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *ECCV*, 2022.

Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *CVPR*, 2023.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019.

Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, and Andrew Markham. P2-Net: Joint description and detection of local features for pixel and point matching. In *ICCV*, 2021.

Haiping Wang, Yuan Liu, Bing Wang, Yujing Sun, Zhen Dong, Wenping Wang, and Bisheng Yang. FreeReg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *ICLR*, 2024.

Ryan W Wolcott and Ryan M Eustice. Visual localization within lidar maps for automated urban driving. In *IROS*, 2014.

Qianliang Wu, Haobo Jiang, Lei Luo, Jun Li, Yaqing Ding, Jin Xie, and Jian Yang. Diff-Reg: Diffusion model in doubly stochastic matrix space for registration problem. In *ECCV*, 2024.

Gongxin Yao, Xinyang Li, Yixin Xuan, and Yu Pan. Mafreei2p: A matching-free image-to-point cloud registration paradigm with active camera pose retrieval. In *ICME*, 2024.

Junsheng Zhou, Baorui Ma, Wenyuan Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Differentiable registration of images and lidar point clouds with voxelpoint-to-pixel matching. *NeurIPS*, 2023.

6 APPENDIX