

Diff²I2P: Differentiable Image-to-Point Cloud Registration with Diffusion Prior

Juncheng Mu^{1,4*} Chengwei Ren^{2,3,4*} Weixiang Zhang³ Liang Pan^{4†} Xiao-Ping Zhang^{2,3†} Yue Gao^{1†}

{¹School of Software, ²Shenzhen Ubiquitous Data Enabling Key Lab, ³SIGS}, Tsinghua University, ⁴Shanghai AI Laboratory

{mujc21, rcw22, zhang-wx22}@mails.tsinghua.edu.cn, paul007p12020@gmail.com, xpzhang@ieee.org, gaoyue@tsinghua.edu.cn

* Equal contribution † Corresponding authors

Abstract

Learning cross-modal correspondences is essential for image-to-point cloud (I2P) registration. Existing methods achieve this mostly by utilizing metric learning to enforce feature alignment across modalities, disregarding the inherent modality gap between image and point data. Consequently, this paradigm struggles to ensure accurate cross-modal correspondences. To this end, inspired by the cross-modal generation success of recent large diffusion models, we propose **Diff²I2P**, a fully **Differentiable I2P** registration framework, leveraging a novel and effective **Diffusion** prior for bridging the modality gap. Specifically, we propose a **Control-Side Score Distillation (CSD)** technique to distill knowledge from a depth-conditioned diffusion model to directly optimize the predicted transformation. However, the gradients on the transformation fail to backpropagate onto the cross-modal features due to the non-differentiability of correspondence retrieval and PnP solver. To this end, we further propose a **Deformable Correspondence Tuning (DCT)** module to estimate the correspondences in a differentiable way, followed by the transformation estimation using a differentiable PnP solver. With these two designs, the Diffusion model serves as a strong prior to guide the cross-modal feature learning of image and point cloud for forming robust correspondences, which significantly improves the registration. Extensive experimental results demonstrate that Diff²I2P consistently outperforms SoTA I2P registration methods, achieving over 7% improvement in registration recall on the 7-Scenes benchmark. Code will be available at <https://github.com/mujc2021/Diff2I2P>.

1. Introduction

Cross-modal registration between images and point clouds is a crucial task in computer vision with broad applications in robotics, AR/VR, etc. Given an image and a point cloud of the same scene, the goal is to estimate a rigid transformation that aligns the point cloud with the image's camera co-

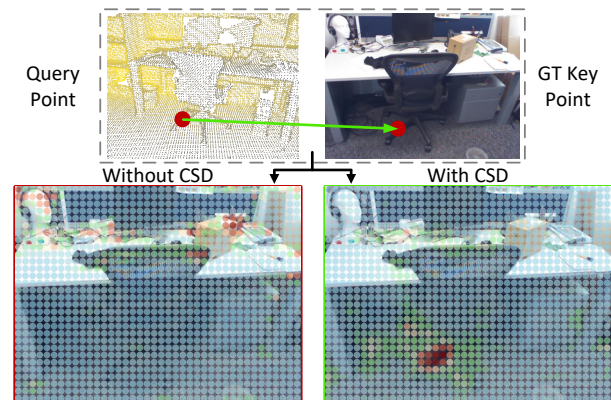


Figure 1. Our proposed **Control-Side Score Distillation (CSD)** effectively promotes the learning of cross-modal features by introducing a novel Diffusion prior to bridging the modality gap. To illustrate this, we select a query point from the *chair leg* in the point cloud, where the features are primarily geometry-dominated and challenging for the image backbone to capture effectively. Its ground-truth corresponding pixel is marked in the top right image. The bottom row visualizes the comparison of cosine similarity between image and point cloud features with and without CSD. With CSD, it correctly locates the pixel with the most similar image features while the one without it fails.

ordinate system. Unlike single-modal registration, such as image registration [22, 26, 42, 43, 48, 65] and point cloud registration [1, 14, 18, 38, 59, 61], which have been extensively studied for decades, cross-modal registration encounters greater challenges due to limited overlap, severe noise, modality misalignment, etc.

Most existing approaches [11, 19, 24, 25, 34, 40, 53] follow the technical roadmap of image or point cloud registration, i.e., *matching and transformation*. They first extract a set of cross-modal correspondences, then the transformation estimation problem is solved as a Perspective-n-Point (PnP) problem using PnP-RANSAC [12, 23]. Therefore, retrieving an accurate putative correspondence set is essential for robust registration. Recent advances [11, 25, 53, 64] have led to substantial progress in learning-based matching methods, which typically employ metric learning tech-

niques, such as contrastive and triplet losses, to enforce alignment between image and point cloud features.

Despite the rapid progress in learning-based cross-modal registration, a notable obstacle remains hindering the performance of current methods, the **modality gap**. Since feature extraction methods like [16, 17] for 2D images and [36, 37, 50] for 3D points mostly only focus on learning local features within their own modality, makes it difficult for the 2D backbones to effectively learn the 3D geometric features from the image. Conversely, the colorless nature of the point cloud makes the 3D backbone fail to adequately capture the texture features of the scene. However, existing methods predominantly focus on metric learning to forcibly alleviate modality misalignment, overlooking this gap and resulting in limited performance.

To tackle these, we propose a novel Control-Side Score Distillation (CSD) technique that distills the 2D texture and 3D geometry knowledge from a depth-conditioned Diffusion [44, 62] model to promote the cross-modal feature learning for registration. This model takes as input an RGB image with a depth map as the condition, guiding the denoising process by integrating 2D image textures with 3D point geometry to generate a novel high-fidelity image. CSD is inspired by the fact that the misaligned depth and image pair will impair the noise prediction capability of this pretrained Diffusion model, leading to severe generation artifacts as shown in Fig. 2. This failure can be modeled by the CSD loss in a Score Distillation Sampling (SDS) manner as follows. The point cloud is first projected into a depth image based on the predicted transformation. Then the input image and depth image are fed into Diffusion for SDS on the ControlNet side. In this way, the alignment can be effectively guided by this cross-modal Diffusion, enabling the image and point cloud backbones to learn distinctive inter-modal features. We illustrate the contribution of CSD in Fig. 1. The selected query point on the top left contains geometry-dominant features that are difficult for the 2D backbone to learn. The pixel features trained without CSD fail to capture geometry features, leading to ambiguous feature matching. While features trained with CSD effectively model the 3D geometry of the chair leg, achieving accurate and concentrated feature matching.

Although CSD effectively promotes cross-modal feature learning, the non-differentiability of the correspondence formation process prevents the gradients of transformation from backpropagating to the backbones. Therefore, we further propose a Deformable Correspondence Tuning (DCT) module, which takes as input the correspondences’ features and coordinates and predicts the point offsets for each correspondence. By leveraging cross-modal features, DCT estimates offsets in a differentiable manner while optimizing the alignment of point-pixel pairs. Finally, we utilize a differentiable PnP solver, BPnP [3], to link the deformed

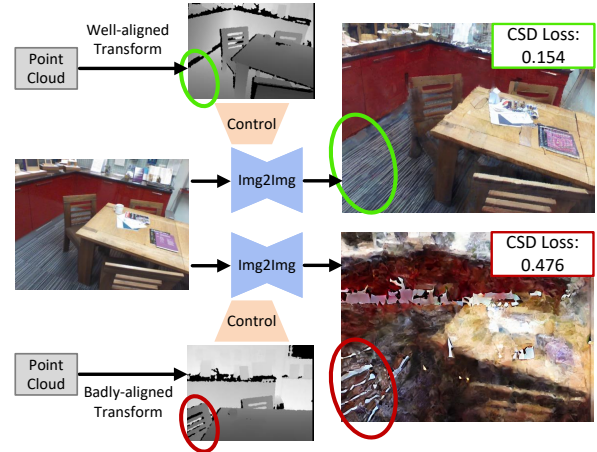


Figure 2. The illustration of CSD leveraging diffusion prior to effectively evaluate the predicted transformation. Well-aligned depth and image pairs (top) result in a lower CSD loss and produce clear output images. In contrast, misaligned cases introduce noticeable artifacts (red circle), leading to a higher CSD loss.

correspondences with CSD by differentially estimating the transformation. In this way, gradients can flow through the offsets and backpropagate to the backbones. Once the training is complete, we simply discard the Diffusion model and infer solely using the distilled feature backbones, resulting in fast runtime and low VRAM cost.

Based on these, we propose Diff²I2P, a fully differentiable cross-modal registration method capable of bridging the modality gap between image and point cloud. Extensive experiments on both 7-Scenes [13] and RGB-D Scenes V2 [21] benchmarks demonstrate our scene-agnostic superiority. To sum up, our main contributions are three-fold:

- We design a fully differentiable image-to-point cloud cross-modal registration method that leverages diffusion prior to bridge the modality gap.
- We propose a novel Control-Side Score Distillation (CSD) technique to distill the alignment knowledge from a Depth-conditioned Diffusion for promoting cross-modal feature learning.
- We propose a Deformable Correspondence Tuning (DCT) Module to enable differentiable feature matching while refining the correspondence set.

2. Related Work

Image-to-Image Registration. Image registration is of significant importance in computer vision and has been studied for decades. Conventional methods employ keypoint detection approaches by applying handcrafted [29, 45] or learned [8, 10, 30, 33, 41, 47, 58] descriptors to establish image correspondences. Then, they estimate the transformation using algorithms such as Bundle Adjustment [15, 51] or Perspective-n-Point (PnP) [23]. Recently, detection-free

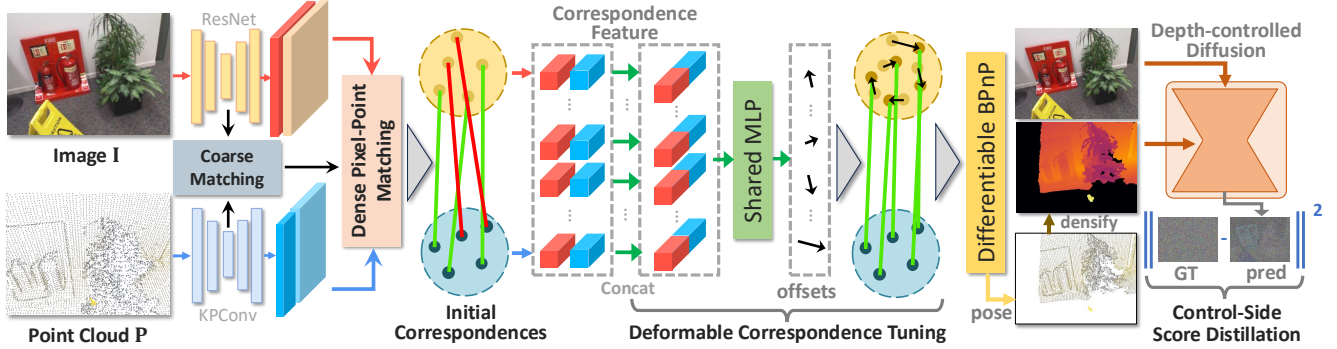


Figure 3. **Pipeline of Diff²I2P.** We construct a fully differentiable pipeline for cross-modal registration. Given the input image and point cloud, we first perform feature matching to obtain the initial correspondences. Next, the Deformable Correspondence Tuning module takes their features as input and predicts point cloud coordinate offsets, ensuring the differentiability meanwhile refining the correspondences. Subsequently, we utilize a BPnP [3] for differentiable transformation estimation and the predicted transform is used to project the point cloud into a depth map. The input image and this depth map are then sent to compute the Control-Side Score Distillation (CSD) loss for distilling cross-modal features, with its gradient backpropagates through the entire registration network for sufficient optimization.

methods [22, 26, 42, 43, 48, 65] have also demonstrated promising results.

Point Cloud Registration. The development of point cloud registration is similar to image registration. Early methods primarily relied on keypoint detection [9, 46], extracting point feature descriptions for registration, such as PPF [9] and FPFH [46]. Subsequently, detection-free methods [1, 5, 14, 18, 61] demonstrate greater potential. Recently, some approaches [32, 38, 39, 57, 59, 60, 63] have begun to employ a coarse-to-fine strategy for registration. They perform feature extraction [36, 37, 50] and matching to obtain correspondences, and utilize robust estimators [2, 6, 12] to estimate the transformation.

Image-to-Point Cloud Registration. Image-to-point cloud registration [11, 24, 40, 55, 64] is more challenging than single-modal registration. 2D3DMATR [25] employs a coarse-to-fine strategy, using transformers to assist in feature extraction, and solve the transformation with PnP [23] solver. However, this force alignment on cross-modal feature learning is insufficient. FreeReg [54] utilizes the Diffusion model to extract cross-modal correspondences, allowing for registration without the need for training. However, it directly applies the Diffusion model for feature extraction, leading to extremely high runtime and computational cost. Our Diff²I2P significantly improves the registration accuracy by bridging the modality gap, while maintaining fast inference times.

Score Distillation Sampling. Score Distillation Sampling (SDS) [35] is widely used in 3D asset generation tasks. Typically, Neural Radiance Fields (NeRF) [31] are employed to represent 3D objects, with SDS distilling knowledge from pre-trained diffusion models [44] for generation. Some methods [4, 27, 49] further leverage SDS to achieve improved generation results. With the emergence of ControlNet [62], we draw inspiration from SDS to propose a

novel Control-Side Score Distillation technique, which distills knowledge from pre-trained ControlNet to optimize control-side (e.g. depth conditioning control) parameters.

3. Method

3.1. Overview

Given a pair of 2D image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a 3D point cloud $\mathbf{P} \in \mathbb{R}^{N \times 3}$, the goal of cross-modal image-to-point cloud registration is to predict a rigid transformation $\mathcal{T} = [\mathbf{R} | \mathbf{t}]$, where $\mathbf{R} \in \mathcal{SO}^3$ is the 3D rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the 3D translation vector. The most common pipeline is correspondence-based, which involves finding correspondences between a set of points and pixels, followed by minimizing the 2D projection error between them:

$$\min_{\mathbf{R}, \mathbf{t}} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \bar{\mathcal{C}}} \|\mathcal{K}(\mathbf{R}\mathbf{x}_i + \mathbf{t}; \mathbf{K}) - \mathbf{y}_i\|^2, \quad (1)$$

where $\bar{\mathcal{C}} = \{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbf{P}, \mathbf{y}_i \in \mathbf{I}\}$ is the correspondences set, \mathbf{K} is the camera intrinsic matrix, and $\mathcal{K} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection from 3D point space to 2D image plane. This is a PnP problem and can be solved by algorithms like RANSAC [12] and EPnP [23].

However, following the process above, previous work suffers from inaccurate correspondences due to two aspects. **First**, the registration process is non-differentiable due to the *argmax* operator used in feature matching and the PnP-RANSAC method in transformation estimation. These prevent the direct supervision of the transformation for sufficient optimization. **Second**, these methods typically apply metric learning to forcibly align cross-modal feature spaces. The inherent modality gap makes it challenging for backbones to effectively learn cross-modal features. To tackle these, we propose a novel module called Deformable Correspondence Tuning (DCT) cooperating with a differentiable

PnP solver to enable differentiable registration while refining the impaired correspondences. Then we further propose a Control-Side Score Distillation (CSD) loss leveraging a novel cross-modal Diffusion prior to bridge the modality gap. The overall pipeline is shown in Fig. 3.

3.2. Differentiable Registration

Pixel-Point Matching. We perform multi-scale pixel-point matching in a coarse-to-fine manner following 2D3D-MATR [25]. Specially, we organize the multi-scale features of the point cloud and image in an FPN [28] style, with KPconv [50] and ResNet [16] employed to extract features from the point cloud and image. Then we adopt a multi-scale patch matching strategy to alleviate the scale mismatching problem. After selecting the point-pixel pairs with *topk* similarity, we obtain the initial correspondence set $\mathcal{C}_{\text{init}} = \{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbf{P}, \mathbf{y}_i \in \mathbf{I}\}$.

Deformable Correspondence Tuning. Inspired by the deformable convolution [7, 50, 56, 66] and attention [67], our key idea of Deformable Correspondence Tuning (DCT) is to introduce learnable point offsets $\Delta \mathbf{p}$ to refine the correspondences. These offsets are predicted using the features from both image and point cloud, which direct the gradient of the correspondence to the feature backbones. This not only preserves the gradient flow for differentiable registration but also fine-tunes the correspondences to prevent potential erroneous matches. For each predicted correspondence $c_i = (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{C}_{\text{init}}$, DCT predicts a point offset $\Delta \mathbf{p}_i \in \mathbb{R}^3$ for \mathbf{x}_i to make it better aligned with \mathbf{y}_i under the ground-truth transformation $\mathcal{T} = [\mathbf{R} | \mathbf{t}]$. After that, our final correspondence set \mathcal{C} can be updated by adding the offsets to points' coordinate:

$$\mathcal{C} = \{(\mathbf{x}_i + \Delta \mathbf{p}_i, \mathbf{y}_i) | (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{C}_{\text{init}}\}. \quad (2)$$

To optimize the point offset $\Delta \mathbf{p}_i$, we consider the global positions with features from both modalities of each correspondence. Specifically, we concatenate the fine-level features of $(\mathbf{x}_i, \mathbf{y}_i)$ to enable feature sharing between the point and pixel spaces and then provide explicit position information for them, i.e., $\mathbf{F}^i = [\mathbf{F}_x^i; \mathbf{F}_y^i; \mathbf{x}_i; \mathcal{K}^{-1}(\mathbf{y}_i; \mathbf{K})] \in \mathbb{R}^{2N+6}$, where $\mathbf{F}_x^i \in \mathbb{R}^N$ and $\mathbf{F}_y^i \in \mathbb{R}^N$ are their corresponding features. Then we simply adopt an MLP to predict the offset with $\Delta \mathbf{p}_i = \text{MLP}(\mathbf{F}^i)$.

DCT aims to predict offsets that move points to positions most likely to align with their corresponding pixels. It alleviates the limitation of feature matching, which only considers feature similarity, by incorporating correspondence correlations in coordinate space. More importantly, DCT enables the backpropagation of gradients from the correspondences to feature extraction backbones, making the differentiable registration feasible.

Differentiable PnP Solver. We employ BPnP [3] as the differentiable solver due to its better numerical stability in

the context of our work. BPnP efficiently derives accurate gradients of PnP solver based on the Implicit Function Theorem [3]. By applying BPnP to correspondence set \mathcal{C} , we can estimate the transformation \mathcal{T} in a differentiable and effective manner, enabling accurate registration. Please refer to the supplementary materials for details about BPnP.

3.3. Control-Side Score Distillation

The key design of CSD is to present an effective evaluation that explicitly optimizes the predicted transformation. The most intuitive approach is to simply apply the estimated transformation \mathcal{T} to the point cloud \mathbf{P} and compute the mean squared error with the image \mathbf{I} :

$$\mathcal{L}_{\text{trans}} = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{C}} \|(\mathbf{R}(\mathbf{x}_i) + \mathbf{t}) - \mathcal{K}^{-1}(\mathbf{y}_i; \mathbf{K})\|^2, \quad (3)$$

where $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{y}_i) | \mathbf{x}_i \in \mathbf{P}, \mathbf{y}_i \in \mathbf{I}\}$ is the ground truth correspondences, \mathbf{K} is the camera intrinsic matrix, and \mathcal{K}^{-1} is the projection from 2D image plane to 3D point space. *However*, our experiments find that this direct optimization of $\mathcal{L}_{\text{trans}}$ suffers from **severe training instability** due to the inherent difficulty of cross-modal training. Misalignment caused by poorly estimated transformations can result in an excessively large $\mathcal{L}_{\text{trans}}$, leading to convergence failure with gradient explosions. To this end, we discard the ground-truth supervision and instead propose a more effective objective namely CSD loss \mathcal{L}_{CSD} to distill cross-modal knowledge from the pretrained Diffusion model.

The specific process of \mathcal{L}_{CSD} calculation is illustrated in Fig. 4. We first apply the predicted transformation to the point cloud and project it to the image plane to get the depth map, which will serve as the input condition for the Diffusion model. However, this depth map is relatively sparse, differing from the dense depth maps used to train the ControlNet [62]. Therefore, we introduce a differentiable densification operator \mathcal{F} to densify the sparse depth map for conditioning:

$$\mathbf{D} = \mathcal{F}(\mathcal{K}(\mathbf{R} \cdot \mathbf{P} + \mathbf{t}, \mathbf{K})), \quad (4)$$

where \mathcal{F} represents the morphology operations of dilation and erosion to generate the dense depth map \mathbf{D} .

We then organize the noisy latent \mathbf{z}_t by encoding the image \mathbf{I} into latent \mathbf{z} using VAE [20] and adding a sampled noise ϵ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to it with a timestep t controlling the noise strength. Then \mathbf{z}_t along with the dense depth map \mathbf{D} are fed into the noise predictor $\hat{\epsilon}_\phi$ to predict the added noise. Since we use a depth-conditioned Diffusion model, here $\hat{\epsilon}_\phi$ represents the original UNet with ControlNet. Following SDS [35], we first define a weighted denoising score matching objective $\mathcal{L}_{\text{Diff}}$ to compute the loss between the predicted noise with the ground truth,

$$\mathcal{L}_{\text{Diff}}(z_t, \mathbf{D}) = \mathbb{E}_{t \sim \mathcal{U}(0,1)} \left[w(t) \|\mathbf{m} \circ (\hat{\epsilon}_\phi(z_t, \mathbf{D}; t) - \epsilon)\|_2^2 \right], \quad (5)$$

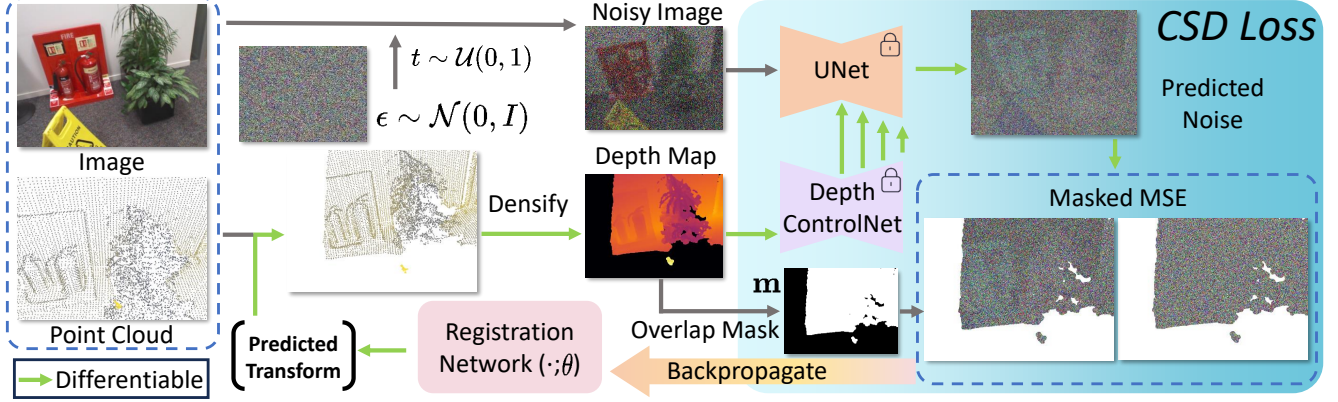


Figure 4. Illustration of Control-Side Score Distillation. We employ a depth-conditioned Diffusion model [44, 62] to bridge the modality gap between images and point clouds. It effectively distills knowledge from the ControlNet to promote cross-modal feature learning for accurate registration. Its gradient flows through the differentiable path marked in green lines to efficiently optimize the registration network.

where ϕ is the learnable parameters of $\hat{\epsilon}_\phi$, $w(t)$ is a weighting function that depends on the timestep t , and \circ denotes the Hadamard product. $\mathbf{m} \in \{0, 1\}^{h \times w}$ is the overlap mask indicating whether a pixel falls within the non-empty region of the projected depth, where h, w are the size of latent. We exclude the loss from the empty region since there is no effective depth condition to guide the denoising process.

Then its gradient with respect to the parameters of registration network θ can be derived as follows:

$$\nabla_{\theta} \mathcal{L}_{\text{Diff}}(z_t, \mathbf{D}(\theta)) = \mathbb{E}_{t, \epsilon} \left[w(t) \cdot \mathbf{m} \circ (\hat{\epsilon}_\phi(z_t, \mathbf{D}) - \epsilon) \underbrace{\frac{\partial \hat{\epsilon}_\phi(z_t, \mathbf{D})}{\partial \mathbf{D}}}_{\text{Jacobian}} \frac{\partial \mathbf{D}}{\partial \theta} \right]. \quad (6)$$

Inspired by SDS [35], omitting the above Jacobian $\in \mathbb{R}^{(h \times w) \times (H \times W)}$ helps alleviate the high computational cost due to the massive number of parameters in the UNet and ControlNet [62] of depth-conditioned Diffusion model. However, the size mismatch between $\mathbf{D} \in \mathbb{R}^{H \times W}$ and $\epsilon \in \mathbb{R}^{h \times w}$ makes this omission nontrivial in CSD. To tackle this, we introduce an intermediate variable $\mathbf{d}(\theta) \in \mathbb{R}^{h \times w}$, which is the resized depth map of \mathbf{D} to match the latent size. Then $\nabla_{\theta} \mathcal{L}_{\text{Diff}}$ can be reformulated as:

$$\nabla_{\theta} \mathcal{L}_{\text{Diff}}(z_t, \mathbf{D}(\theta)) = \mathbb{E}_{t, \epsilon} \left[w(t) \cdot \mathbf{m} \circ (\hat{\epsilon}_\phi(z_t, \mathbf{D}) - \epsilon) \underbrace{\frac{\partial \hat{\epsilon}_\phi(z_t, \mathbf{D})}{\partial \mathbf{D}} \frac{\partial \mathbf{D}}{\partial \mathbf{d}}}_{\text{size matched Jacobian}} \frac{\partial \mathbf{d}}{\partial \theta} \right]. \quad (7)$$

By combining the term $\frac{\partial \mathbf{D}}{\partial \mathbf{d}} \in \mathbb{R}^{(H \times W) \times (h \times w)}$, we get the size matched Jacobian $\in \mathbb{R}^{(h \times w) \times (h \times w)}$, which can be omitted as a unit matrix to save the computation cost. Our implementation finds that this omission leads to an effective optimization for the registration network θ (Sec. 4.4). We formally derive the gradient of \mathcal{L}_{CSD} with respect to the

parameter of registration network θ as:

$$\nabla_{\theta} \mathcal{L}_{\text{CSD}}(\mathbf{x}, \mathbf{D}(\theta)) = \mathbb{E}_{t, \epsilon} \left[w(t) \cdot \mathbf{m} \circ (\hat{\epsilon}_\phi(\mathbf{x}, \mathbf{D}; t) - \epsilon) \frac{\partial \mathbf{D}}{\partial \theta} \right]. \quad (8)$$

During training, we directly compute this $\nabla_{\theta} \mathcal{L}_{\text{CSD}}$ and backpropagate it to the registration network θ for efficient optimization. Please refer to the supplementary materials for more details.

3.4. Loss functions

Our model is trained using the sum of three losses, a metric learning loss \mathcal{L}_m for supervising the descriptive image and point cloud features, an offset loss \mathcal{L}_o for supervising the deformable correspondences, and a CSD loss \mathcal{L}_{CSD} for bridging the modality gap, i.e., $\mathcal{L} = \alpha \mathcal{L}_m + \beta \mathcal{L}_o + \gamma \mathcal{L}_{\text{CSD}}$, where α, β and γ are coefficients of losses.

Following 2D3D-MATR [25], we utilize \mathcal{L}_m for correspondences supervision, which is composed of a scaled circle loss \mathcal{L}_c on coarse-level features and a standard circle loss \mathcal{L}_f on fine-level features. In practical implementation, we find that simply extending the original \mathcal{L}_c from the Transformer to the image and point cloud backbones will augment the learning of distinctive features, so we adopt it for coarse-level supervision. The metric learning loss is defined as $\mathcal{L}_m = \mathcal{L}_{\text{aug}} + \lambda \mathcal{L}_f$. For detailed information about these losses, please refer to supplementary materials.

The offset loss \mathcal{L}_o is designed to estimate accurate offsets $\Delta \mathbf{p}$ for better coordinate tuning. For each $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{C}$, we design their distance loss $\mathcal{L}_{\mathbf{d}}^i$ as:

$$\mathcal{L}_{\mathbf{d}}^i = \left\| (\bar{\mathbf{R}}(\mathbf{x}_i + \Delta \mathbf{P}_i) + \bar{\mathbf{t}}) - \mathcal{K}^{-1}(\mathbf{y}_i; \mathbf{K}) \right\|^2, \quad (9)$$

where $\mathcal{K}^{-1} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is the projection function from 2D image plane to 3D point space. To prevent the offsets from compromising feature extraction backbones' capability, we impose an L2-norm regularization to enhance stability, i.e.,

$\mathcal{L}_r^i = \|\Delta \mathbf{p}_i\|_2$. Then, the total offset loss \mathcal{L}_o can be calculated as,

$$\mathcal{L}_o = \frac{1}{|C|} \sum_{i=1}^{|C|} (\mathcal{L}_d^i + \mu \mathcal{L}_r^i), \quad (10)$$

where μ is the coefficient to control regularization strength.

For \mathcal{L}_{CSD} , we do not explicitly compute its value. Instead, we directly derive its gradient from Eq. (8) for back-propagation.

4. Experiments

In this section, we first introduce the implementation details (Sec. 4.1), then we present extensive experimental results of Diff²I2P and other **training-based baselines** on 7-Scenes [13] (Sec. 4.2) and RGB-D Scenes V2 [21] (Sec. 4.3) for evaluation. In Sec. 4.4, we conduct extensive ablation studies to validate the contribution of each component. In Sec. 4.5, we discuss the comparison with another competitive Diffusion-based method FreeReg [54], and present more evaluations on the inference runtime and VRAM.

4.1. Implementation Details

Network architecture. We employ an FPN [28] strategy for downsampling and feature extraction of both the image and point cloud, with each modality undergoing four stages of downsampling. For feature extraction, we use ResNet [16] for image features and KPConv [50] for point cloud features. The image downsampling reduces the resolution from 480×640 to 60×80, while the point cloud downsampling begins with a voxel size of 2.5 cm, doubling at each following stage. The configuration of the transformer layers is consistent with 2D3D-MATR [25]. The training can be completed with four RTX 3090s within one day.

Datasets. We mainly evaluate Diff²I2P on 7-Scenes [13] and RGB-D Scenes V2 [21] benchmarks. 7-Scenes [13] consists of 7 indoor scenes with 46 RGB-D sequences. Each image-point cloud pair has an overlap of at least 50%. We utilized the official data split, resulting in 4,048 samples for training, 1,011 for validation, and 2,304 for testing. Our benchmarking on this dataset is consistent with that of [25]. RGB-D Scenes V2 [21] includes 14 indoor scenes, comprising 11,427 frames. Each image-point cloud pair has at least 30% overlap. We randomly split the dataset into training, validation, and test sets, containing 1978, 117, and 386 samples, respectively. To further evaluate Diff²I2P in dynamic outdoor scenarios, we also provide experiments on KITTI [52] in the supplementary material.

Baselines. Since our method requires training, we primarily compare Diff²I2P with four training-based methods, FCGF-2D3D [5], Predator-2D3D [18], P2-Net [53], and 2D3D-MATR [25]. To ensure fair comparisons, the feature extraction backbones used in all comparison methods are kept

Table 1. Registration results of Diff²I2P and baselines on 7-Scenes. The best of each indicator is in **bold**.

Model	Chess	Fire	Heads	Office	Pupk	Kitc	Stairs	Mean
Inlier Ratio ↑								
FCGF [5]	34.2	32.8	14.8	26.0	23.3	22.5	6.0	22.8
P2Net [53]	55.2	46.7	13.0	36.2	32.0	32.8	5.8	31.7
Predator [18]	34.7	33.8	16.6	25.9	23.1	22.2	7.5	23.4
MATR [25]	72.1	66.0	31.3	60.7	50.2	52.5	18.1	50.1
Ours	74.1	68.8	39.2	65.6	52.1	54.2	18.1	53.2
Feature Matching Recall ↑								
FCGF [5]	99.7	98.2	69.9	97.1	83.0	87.7	16.2	78.8
P2Net [53]	100.0	99.3	58.9	99.1	87.2	92.2	16.2	79.0
Predator [18]	91.3	95.1	76.7	88.6	79.2	80.6	31.1	77.5
MATR [25]	100.0	99.6	98.6	100.0	92.4	95.9	58.1	92.1
Ours	100.0	100.0	100.0	100.0	93.4	96.2	55.4	92.2
Registration Recall ↑								
FCGF [5]	89.5	79.7	19.2	85.9	69.4	79.0	6.8	61.4
P2Net [53]	96.9	86.5	20.5	91.7	75.3	85.2	4.1	65.7
Predator [18]	69.6	60.7	17.8	62.9	56.2	62.6	9.5	48.5
MATR [25]	96.9	90.7	52.1	95.5	80.9	86.1	28.4	75.8
Ours	99.0	95.6	74.0	98.9	86.8	90.2	36.5	83.0

consistent. We also present more discussions and comparisons with another Diffusion-based baseline FreeReg [54] in Sec. 4.5.

Metrics. We primarily adopt four metrics. Inlier Ratio (IR), which calculates the proportion of correctly estimated pixel-point correspondences (defined as those within 5cm in 3D distance). Patch Inlier Ratio (PIR), which measures the proportion of correctly estimated patch correspondences (defined as those with overlap bigger than 30%). Feature Matching Recall (FMR), which measures the percentage of all point cloud-image pairs with an inlier ratio exceeding a threshold (*i.e.*, 10%). Registration Recall (RR), which assesses the proportion of all point cloud-image pairs with an RMSE less than a specified threshold (*i.e.*, 10 cm).

4.2. Evaluations on 7-Scenes

Quantitative results. We train Diff²I2P on the 7-Scenes [13] dataset and the experimental results are shown in Tab. 1. For IR, Diff²I2P consistently outperforms previous methods, indicating more accurate feature matching due to the differentiable supervision with CSD loss. Our method also achieves the best FMR across most scenes, reaching 100% in four scenes, and outperform the baselines in the most of cases. Regarding the most crucial metric RR, Diff²I2P shows comprehensive performance improvements, with an average RR exceeding that of 2D3D-MATR [25] by 7.2% and achieving a 22% increase in the challenging *Heads* scene. These results demonstrate the breakthrough advancements of our proposed fully differentiable pipeline with CSD supervision.

Qualitative results. Fig. 5 visualizes the correspondences

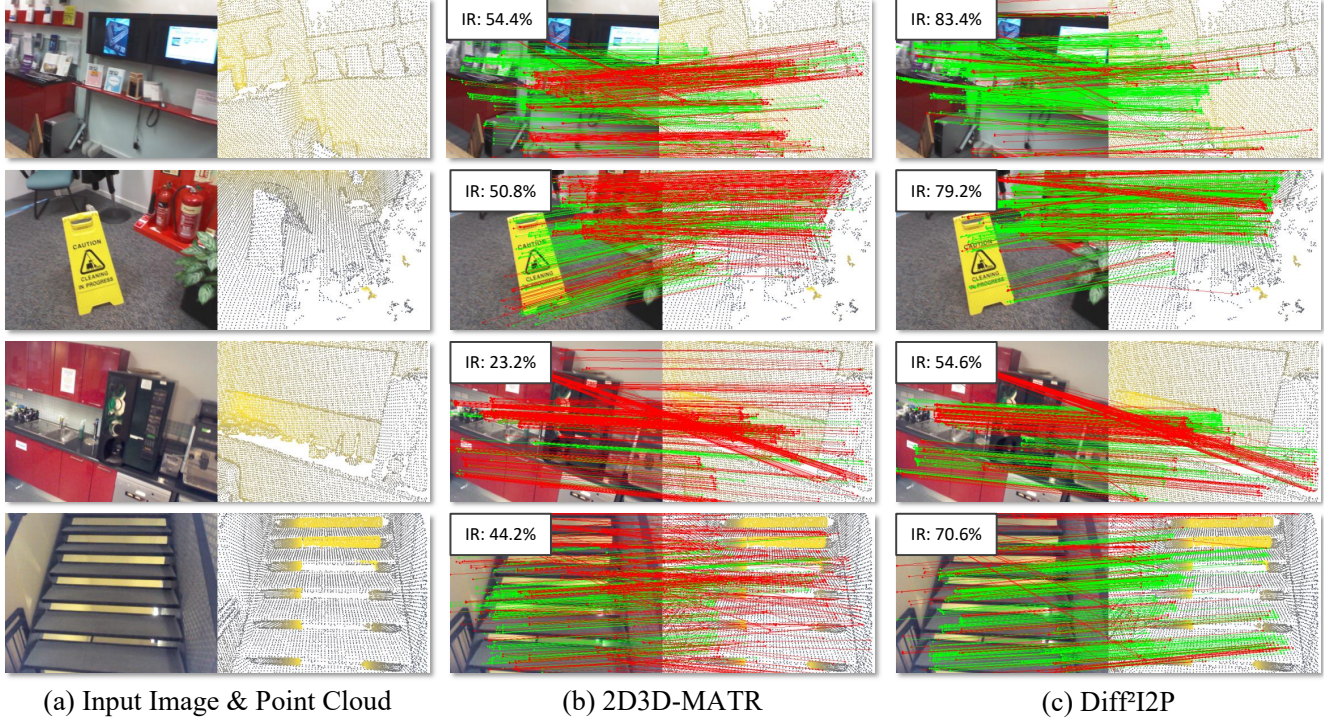


Figure 5. Qualitative results on the 7-Scenes dataset. The red lines indicate erroneous correspondences (3D distance greater than 5 cm), while the green lines represent correct correspondences.

estimated by 2D3D-MATR and Diff²I2P on 4 scenes from the 7-Scenes dataset. Column (a) shows the input image and point cloud pairs, while columns (b) and (c) show the correspondences from 2D3D-MATR and our Diff²I2P, respectively. We selected the top 500 matching pairs based on their matching scores for visualization. The first and second rows depict scenes from *redkitchen* and *fire*, which contain relatively rich colors and clear geometric features. Both methods achieve a high IR in these cases, with Diff²I2P better identifying regions with prominent textures and geometric features, achieving an IR of approximately 80%. In the third and fourth rows, the scene contains complicated and repetitive geometric patterns under complex lighting, making it challenging for precise feature matching. 2D3D-MATR generates numerous incorrect matches, whereas our method achieves significantly more accurate matching, with an inlier ratio (IR) twice as high in comparison. This indicates the differentiable supervision with CSD loss effectively facilitates the learning of cross-modal features.

4.3. Evaluations on RGB-D Scenes V2

We train Diff²I2P on the RGBD Scenes V2 [21] dataset and compare it with the baselines. The experimental results are shown in Tab. 2. For intermediate metrics such as PIR, IR, and FMR, Diff²I2P demonstrates a significant improvement over the baselines, highlighting the effectiveness of feature matching. This improvement benefits from the cross-modal

Table 2. Registration results of Diff²I2P and baselines on RGB-D Scenes V2. The best of each indicator is in **bold**.

Method	PIR (%) [↑]	IR (%) [↑]	FMR (%) [↑]	RR (%) [↑]
FCGF-2D3D [5]	20.1	10.3	29.2	32.5
P2-Net [53]	30.4	14.5	63.7	41.7
Predator-2D3D [18]	32.6	15.8	68.1	33.6
2D3D-MATR [25]	57.6	36.3	76.0	56.9
Diff ² I2P (ours)	60.8	36.9	77.1	60.5

supervision provided by the CSD loss, enabling Diff²I2P to more accurately estimate the transformation, and consequently achieve a higher registration recall.

4.4. Ablation Studies

We conduct ablation studies on the network structure of Diff²I2P to validate the contribution of each component. All experiments are performed on the 7-Scenes [13].

Deformable Correspondence Tuning. The absence of the DCT module would disrupt the differentiability of training, we can't ablate it in the training phase. Therefore, we ablate it during inference instead. The visualization of the estimated correspondences with and without DCT refinement are illustrated in Fig. 6. The predicted offsets refine the incorrect correspondences and further improve their overall alignment. Here we emphasize that the *principal function* of DCT is to ensure **training differentiability**, i.e., enabling the gradient of the predicted transformation to be backpropagated to the backbone. Without DCT, the distillation of the Diffusion model can not be achieved.

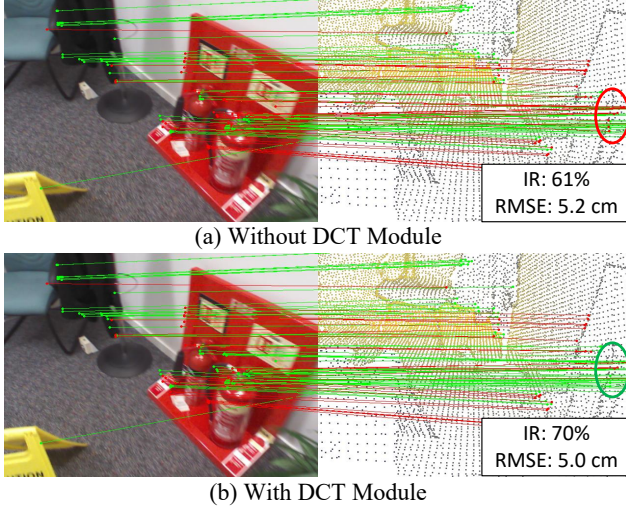


Figure 6. Illustration of the DCT for refining correspondences. (a) shows the correspondences before offsets refinement. (b) shows the correspondences after offset adjustments. Although DCT can fine-tune the correspondence to improve registration accuracy, we still emphasize that its *principal function* is to ensure **training differentiability**, i.e., enabling the gradient of the predicted transformation to be backpropagated to the backbone. Without DCT, the distillation of the Diffusion model can not be achieved.

Loss Functions. We ablate the proposed loss functions, including the circle loss used in [25], our augmented circle loss, CSD loss, CSD loss without the overlap mask, and the transformation loss. The experimental results are shown in Tab. 3, where these losses are abbreviated as *circle*, *aug*, *CSD*, *CSD**, and *trans*, respectively. Here we retain the offset loss in all experiments for training the DCT module.

In Tab. 3, experiment (e) shows the loss combination selected for Diff²I2P, which achieves the highest registration recall of 83.0% on the 7-Scenes dataset. Experiment (a) ablates our augmented circle loss, and compared to (e), all metrics show a decline. This demonstrates the crucial role of the aug-circle loss in providing more comprehensive guidance for correspondence estimation. Experiment (b) uses the transformation loss, which applies the predicted transformation to the point cloud and computes the MSE loss with their corresponding image points. However, this loss is highly unstable, particularly when the transformation bias is large, which can result in an extremely high loss and lead to gradient explosions. Experiments (c) and (d) show the results of canceling the CSD loss and changing to the unmasked version of CSD loss, respectively. Both demonstrate a noticeable degradation in registration recall compared to experiment (e), with other metrics fluctuating slightly. These experiments solidly validate our strategy of loss selection. This further validates that with well-designed architecture and loss design, the strong cross-modal features of the Diffusion model are effectively distilled into our fea-

Table 3. Ablation study of the loss functions on 7Scenes. CSD* indicates the CSD without the overlap mask.

Type	circle	aug	CSD*	CSD trans	PIR	IR	FMR	RR
(a)	✓			✓	86.5	53.1	92.0	82.5
(b)		✓		✓	Gradient Explosion			
(c)		✓			86.3	53.0	91.6	81.3
(d)		✓	✓		86.4	52.6	91.4	81.7
(e)		✓		✓	86.7	53.2	92.1	83.0

Table 4. Comparisons with FreeReg [54].

Method	RR (%)↑	Time (s)↓	Size (MB)↓	VRAM (GB)↓
2D3D-MATR [25]	75.8	0.072	118.62	3.1
FreeReg [54]	71.2	9.634	-	14.2
Diff ² I2P (ours)	83.0	0.074	118.75	3.1

ture backbone, enabling robust and accurate registration.

4.5. Comparison with FreeReg

FreeReg [54] is another strong baseline based on Diffusion model. It extracts cross-modal correspondence in a training-free manner by utilizing intermediate Diffusion features for matching. The detailed comparison of performance and computational costs is outlined in Tab. 4.

Discussion. Despite the inherent unfairness in comparing our method with FreeReg due to the disparity in training requirements, we include it as an informative baseline to highlight the strengths of our method in two key aspects. *First*, unlike directly using intermediate features for matching, we distill the cross-modal features from the Diffusion model into the backbones. This reduces the randomness during inference, leading to better performance. *Second*, FreeReg requires a full invocation of the large Diffusion model for each inference, resulting in slow inference speed and high VRAM consumption. In contrast, our method distills knowledge from the Diffusion model only during the training phase, enabling more efficient inference with higher performance.

5. Conclusion

In this paper, we present Diff²I2P, a fully differentiable pipeline for cross-modal registration that bridges the modality gap between image and point cloud. With our proposed CSD loss, the training process can distill 2D texture and 3D geometry knowledge from a pretrained depth-controlled Diffusion model for cross-modal feature learning. We further propose a DCT module to introduce differentiability to the correspondence set while refining impaired matches. By combining a differentiable BPnP solver, we construct a fully differentiable network for sufficient optimization. Extensive experiments on both 7-Scenes and RGB-D Scenes V2 benchmarks demonstrate the state-of-the-art performance of Diff²I2P, significantly surpassing the strongest baselines. We hope our proposed method inspire further work in differentiable registration and bridging the modality gap.

Acknowledgments. This work was supported by the NSFC (624B1007) and Shanghai Artificial Intelligence Laboratory. This work was also supported in part by Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), and by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Endowed Professorship Scheme of Shenzhen Pengrui Foundation.

References

- [1] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3D local features. In *CVPR*, pages 6359–6367, 2020. 1, 3
- [2] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *CVPR*, pages 15859–15869, 2021. 3
- [3] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *CVPR*, pages 8100–8109, 2020. 2, 3, 4
- [4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, pages 22246–22256, 2023. 3
- [5] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *ICCV*, pages 8958–8966, 2019. 3, 6, 7
- [6] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR*, pages 2514–2523, 2020. 3
- [7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 4
- [8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, pages 224–236, 2018. 2
- [9] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, pages 998–1005. Ieee, 2010. 3
- [10] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, pages 8092–8101, 2019. 2
- [11] 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 *IEEE international conference on robotics and automation*, pages 4790–4796. IEEE, 2019. 1, 3
- [12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1, 3
- [13] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179. IEEE, 2013. 2, 6, 7
- [14] Zan Gojcic, Caifa Zhou, Jan Wegner, and Andreas Wieser. The perfect match: 3D point cloud matching with smoothed densities. In *CVPR*, pages 5545–5554, 2019. 1, 3
- [15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 4, 6
- [17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 2
- [18] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3D point clouds with low overlap. In *CVPR*, pages 4267–4276, 2021. 1, 3, 6, 7
- [19] Peng Jiang and Srikanth Saripalli. Contrastive learning of features between images and lidar. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 411–417. IEEE, 2022. 1
- [20] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4
- [21] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *IEEE international conference on robotics and automation*, pages 3050–3057. IEEE, 2014. 2, 6, 7
- [22] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *CVPR*, pages 13153–13163, 2021. 1, 3
- [23] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Eppnp: An accurate o(n) solution to the pnp problem. *IJCV*, 81:155–166, 2009. 1, 2, 3
- [24] Jiaxin Li and Gim Hee Lee. Deepi2p: Image-to-point cloud registration via deep classification. In *CVPR*, pages 15960–15969, 2021. 1, 3
- [25] 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*, pages 14128–14138, 2023. 1, 3, 4, 5, 6, 7, 8
- [26] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *NeurIPS*, 33:17346–17357, 2020. 1, 3
- [27] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pages 300–309, 2023. 3
- [28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4, 6
- [29] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157. Ieee, 1999. 2
- [30] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan.

- Aslfeat: Learning local features of accurate shape and localization. In *CVPR*, pages 6589–6598, 2020. 2
- [31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3
- [32] Juncheng Mu, Lin Bie, Shaoyi Du, and Yue Gao. Colorpcr: Color point cloud registration with multi-stage geometric-color fusion. In *CVPR*, pages 21061–21070, 2024. 3
- [33] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *NeurIPS*, 31, 2018. 2
- [34] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *AAAI*, pages 11856–11864, 2020. 1
- [35] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3, 4, 5
- [36] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, pages 652–660, 2017. 2, 3
- [37] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 2, 3
- [38] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, pages 11143–11152, 2022. 1, 3
- [39] Chengwei Ren, Yifan Feng, Weixiang Zhang, Xiaoping Steven Zhang, and Yue Gao. Multi-scale consistency for robust 3d registration via hierarchical sinkhorn tree. *NeurIPS*, 37, 2024. 3
- [40] 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*, 33(3):1198–1208, 2022. 1, 3
- [41] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *NeurIPS*, 32, 2019. 2
- [42] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 31, 2018. 1, 3
- [43] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *ECCV*. Springer, 2020. 1, 3
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 5
- [45] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pages 2564–2571. Ieee, 2011. 2
- [46] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009. 3
- [47] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 2
- [48] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 1, 3
- [49] Jiayang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3
- [50] Hugues Thomas, Charles Ruizhongtai Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KpConv: Flexible and deformable convolution for point clouds. In *ICCV*, pages 6411–6420, 2019. 2, 3, 4, 6
- [51] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 2
- [52] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 6
- [53] Bing Wang, Changhao Chen, Zhaopeng Cui, Jie Qin, Chris Xiaoxuan Lu, Zhengdi Yu, Peijun Zhao, Zhen Dong, Fan Zhu, Niki Trigoni, et al. P2-net: Joint description and detection of local features for pixel and point matching. In *ICCV*, pages 16004–16013, 2021. 1, 6, 7
- [54] 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. *arXiv preprint arXiv:2310.03420*, 2023. 3, 6, 8
- [55] 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*, pages 160–178. Springer, 2024. 3
- [56] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, et al. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *CVPR*, pages 5652–5661, 2024. 4
- [57] Fan Yang, Lin Guo, Zhi Chen, and Wenbing Tao. One-inlier is first: Towards efficient position encoding for point cloud registration. *NeurIPS*, 35, 2022. 3
- [58] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, pages 467–483. Springer, 2016. 2
- [59] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. CofiNet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *NeurIPS*, 34:23872–23884, 2021. 1, 3

- [60] Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, and Guojun Dai. PEAL: Prior-embedded explicit attention learning for low-overlap point cloud registration. In *CVPR*, pages 17702–17711, 2023. 3
- [61] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions. In *CVPR*, pages 1802–1811, 2017. 1, 3
- [62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 2, 3, 4, 5
- [63] Yu Zhang, Junle Yu, Xiaolin Huang, Wenhui Zhou, and Ji Hou. PCR-CG: Point cloud registration via deep explicit color and geometry. In *ECCV*, pages 443–459. Springer, 2022. 3
- [64] 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*, 36, 2024. 1, 3
- [65] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, pages 4669–4678, 2021. 1, 3
- [66] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 4
- [67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4