Adaptive Agent Selection and Interaction Network for Image-to-Point Cloud Registration

Zhixin Cheng[1], Xiaotian Yin[3], Jiacheng Deng[1], Bohao Liao[1], Yujia Chen[3], Xu Zhou[4], Baoqun Yin[1], Tianzhu Zhang[1,2] *

[1]School of Information Science and Technology, University of Science and Technology of China
[2]National Key Laboratoray of Deep Space Exploration, Deep Space Exploration Laboratory
[3]Institute of Advanced Technology, University of Science and Technology of China
[4]Sangfor Technologies
chengzhixin, xiaotianyin, dengjc, liaobh, yujia_chen@mail.ustc.edu.cn, zhouxu@sangfor.com.cn, bqyin, tzzhang@ustc.edu.cn

Abstract

Typical detection-free methods for image-to-point cloud registration leverage transformer-based architectures to aggregate cross-modal features and establish correspondences. However, they often struggle under challenging conditions, where noise disrupts similarity computation and leads to incorrect correspondences. Moreover, without dedicated designs, it remains difficult to effectively select informative and correlated representations across modalities, thereby limiting the robustness and accuracy of registration. To address these challenges, we propose a novel cross-modal registration framework composed of two key modules: the Iterative Agents Selection (IAS) module and the Reliable Agents Interaction (RAI) module. IAS enhances structural feature awareness with phase maps and employs reinforcement learning principles to efficiently select reliable agents. RAI then leverages these selected agents to guide cross-modal interactions, effectively reducing mismatches and improving overall robustness. Extensive experiments on the RGB-D Scenes v2 and 7-Scenes benchmarks demonstrate that our method consistently achieves state-of-the-art performance.

"Not everything that can be counted counts, and not everything that counts can be counted."
— Albert Einstein

Introduction

Image-to-point cloud registration (I2P) aims to estimate the rigid transformation that aligns the point cloud with the camera coordinate system, given an image and a point cloud from the same scene. This process typically involves cross-modal feature matching to establish correspondences, followed by pose estimation to compute the rotation and translation matrices. As a critical step in various vision tasks, such as 3D reconstruction (Deng, Lu, and Zhang 2024; Chang et al. 2025; Zha et al. 2025a,b), SLAM (Durrant-Whyte and Bailey 2006; He et al. 2023; Wang et al. 2025a,b), and visual localization (Chen et al. 2025; Wu et al. 2024; Wang, Yang, and Zhang 2023), I2P plays a vital and foundational role. However, a significant and inherent gap exists between the two modalities: images are dense, structured 2D grids, while point clouds are sparse, unordered, and irregular 3D data. Effectively bridging this gap to improve feature aggregation and alignment across these modalities remains a major challenge in cross-modal registration.

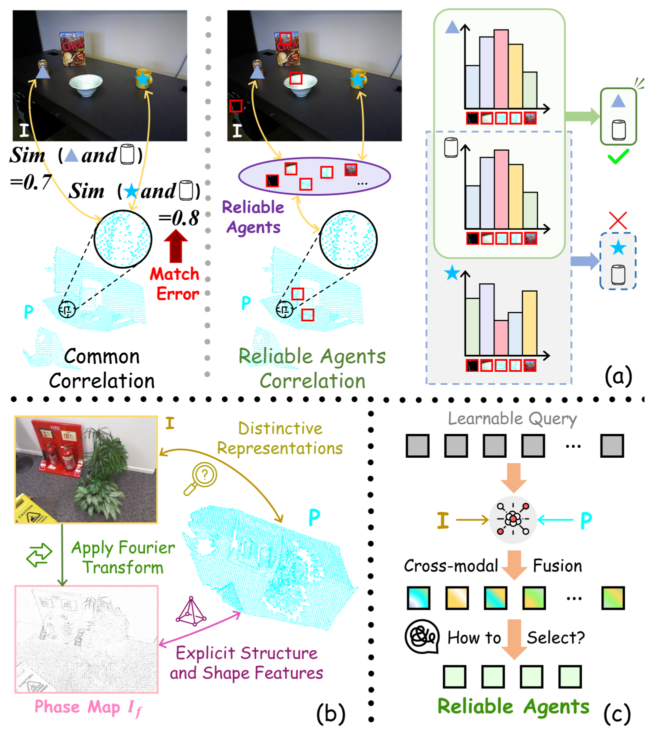

Figure 1: (a) Visual comparison of common and reliable agents correlations (b) Visualization of phase map enhancement. (c) Procedure and challenges of the reliable agents interaction method.

Current solutions to I2P challenges generally fall into two categories. Recently, detection-free approaches (Kang et al. 2023; Li et al. 2023) have emerged as the mainstream, gradually replacing traditional detect-then-match methods (Feng et al. 2019; Wang et al. 2021; Ren et al. 2022). The detect-then-match paradigm involves independently detecting 2D keypoints in images and 3D keypoints

*Corresponding author.

in point clouds, and then matching them based on semantic features. However, this approach has two major limitations: first, keypoint detection is inherently modality dependent, image keypoints rely on texture and color, while point cloud keypoints depend on local geometry—making it challenging to extract consistent keypoints. Second, the significant differences between 2D and 3D descriptors impedes effective cross-modal alignment. Thus, 2D3D-MATR (Li et al. 2023) introduces the first coarse-to-fine detection-free framework. This framework extracts features from both modalities, performs coarse patch-level matching, and then refines these matches into fine-grained pixel-to-point correspondences. Finally, it estimates rigid transformations using PnP-RANSAC (Lepetit, Moreno-Noguer, and Fua 2009; Fischler and Bolles 1981). However, in the scenarios with repetitive structures (Torii et al. 2013), non-overlapping regions, or varying illumination, directly aggregating features through a conventional transformer (Vaswani et al. 2017) may introduce substantial noise, thereby degrading matching accuracy (Han et al. 2024; Cheng et al. 2025c) and ultimately affecting overall registration performance.

Building on the discussion of detection-free methods, we identify two critical issues that are essential for accurate and robust registration between images and point clouds. ***Firstly, we need to find effective ways to aggregate informative features from both modalities to reduce cross-modal mismatches.*** Traditional transformer-based approaches often compute similarities directly between image and point cloud features. However, in scenes with repetitive structures or non-overlapping regions, such similarity measures can lead to incorrect correspondences. For example, as illustrated in Figure 1(a), a soda can might be confused with a structurally similar object. By aggregating features based on key informative regions, we could reduce such ambiguities and improve registration accuracy. ***Therefore, it is crucial to reliably select representative features from both modalities to enhance registration quality.*** As images mainly capture texture and point clouds capture geometry, modality-specific encoding often leads to distinct feature representations. To bridge this gap, we introduce phase maps to enhance the image's sensitivity to structural edges, thereby facilitating the selection of representative features (Figure 1(b)). Building on this, we further explore learning-based queries to aggregate features across modalities (Figure 1(c)). However, without dedicated design, the full potential of the model for cross-modal registration may not be realized, as identifying optimal and reliable fusion agents is essential.

In response to these concerns, we propose Adaptive Agent Selection and Interaction Network (A2SI) for image-to-point cloud registration. Our framework consists of two main modules: Iterative Agent Selection (IAS) module and Reliable Agent Interaction (RAI) module. In the IAS, we extract phase maps from images to highlight structural edge information, which helps reduce the domain gap between image and point cloud features (Liao et al. 2024). Considering that reinforcement learning (Wang et al. 2024b) supports reward-guided scoring and selection, we adapt this idea to our task and propose a lightweight Tri-Stage Agents Optimization Strategy, designed to identify reliable agents that most enhance the model's registration capability. In the RAI module, we utilize these reliable agents to efficiently aggregate cross-modal features, moving away from traditional transformer-based fusion methods. The RAI module greatly minimizes the noise caused by repetitive structures and non-overlapping regions, resulting in higher-quality matching with lower computational costs.

Our main contributions are summarized as follows:

- We propose a novel Adaptive Agent Selection and Interaction Network for image-to-point cloud registration, which achieves excellent accuracy and strong generalization ability.
- We design the Iterative Agents Selection module, which leverages phase maps to enhance structural awareness in image features, and a Tri-Stage Agents Optimization Strateg selects reliable agents containing key information. In the Reliable Agents Interaction module, we replace traditional transformer fusion with agent-guided interaction for improved matching and reduced errors.
- Extensive experiments and ablation studies on two benchmarks, RGB-D Scenes v2 and 7-Scenes, demonstrate the superiority of our method, establishing it as a new state-of-the-art for image-to-point cloud registration.

Related works

In this section, we briefly overview related works on I2P registration, including stereo image registration, point cloud registration, and inter-modality registration.

Image and Point Cloud Registration

Early image registration methods relied on handcrafted keypoints and descriptors like SIFT (Ng and Henikoff 2003) and ORB (Rublee et al. 2011). Later, learning-based approaches such as SuperGlue (Sarlin et al. 2020), leveraging transformers (Vaswani et al. 2017), improved robustness but struggled in textureless regions. To address this, detector-free methods like LoFTR (Sun et al. 2021) and Efficient LoFTR (Wang et al. 2024a) adopted coarse-to-fine matching with global attention for dense correspondence estimation. Point cloud registration evolved from handcrafted descriptors like PPF (Moheimani, Vautier, and Bhikkaji 2006) and FPFH (Rusu, Blodow, and Beetz 2009) to deep learning methods. CoFiNet (Yu et al. 2021) introduced a detector-free coarse-to-fine pipeline, and GeoTransformer (Qin et al. 2023) enhanced performance by using transformers to model global geometry and replacing RANSAC (Fischler and Bolles 1981) with a local-to-global registration framework.

Inter-Modality Registration

Compared to intra-modality registration, inter-modality registration presents greater challenges due to significant domain discrepancies. Traditional approaches typically follow a detect-then-match paradigm. For instance, 2D3DMatchNet (Feng et al. 2019) leverages SIFT (Ng and Henikoff 2003) for keypoint detection, while ISS (Sontag 1998) extracts local patches from point clouds and employs CNNs or PointNet (Qi et al. 2017) for descriptor learning

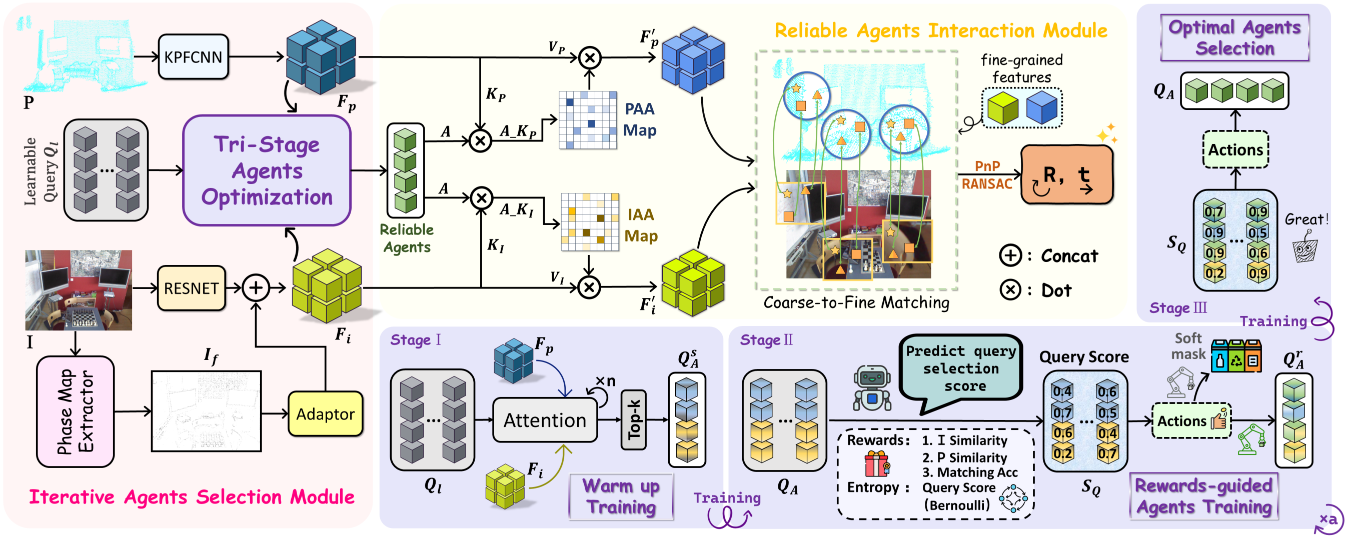

Figure 2: Overall pipeline of A2SI. A phase map extractor first enhances image feature sensitivity to edge structures. The Tri-Stage Agents Optimization strategy then selects informative reliable agents from redundant learnable queries. These agents guide cross-modal aggregation in the Reliable Agents Interaction Module, progressively establishing dense correspondences in a coarse-to-fine manner. Finally, a rigid transformation is estimated via PnP with RANSAC.

and matching. P2-Net (Wang et al. 2021) improves efficiency by jointly detecting and matching points in a single step. However, in cross-modal scenarios, keypoint detection accuracy often degrades (Simon et al. 2017; Barroso-Laguna et al. 2019), limiting performance and motivating the rise of detector-free methods. 2D3D-MATR (Li et al. 2023) employs a coarse-to-fine strategy, using a transformer for coarse matching and refining the results with PnP-RANSAC, thus eliminating keypoint detection and improving descriptor alignment, while B2-3D (Cheng et al. 2025b,a) further enhances accuracy through uncertainty modeling and domain adaptation. Building on 2D3D-MATR, our proposed A2SI also adopts a detector-free design, focusing on selecting features that capture region information for cross-modal aggregation, thereby setting a new benchmark for image-to-point cloud registration.

Method

Let $I \in \mathbb{R}^{H\times W\times 3}$ represent an image and $P \in \mathbb{R}^{N\times 3}$ represent a point cloud from the same scene. Here, H and W denote the height and width of the image, respectively, while N is the number of 3D points in the point cloud. The objective of image-to-point cloud registration is to estimate a rigid transformation $[R, t]$, where the rotation matrix $R \in \mathrm{SO}(3)$ and the translation vector $t \in \mathbb{R}^3$.

We propose A2SI, a detection-free registration framework that efficiently aggregates cross-modal features. The Iterative Agents Selection module incorporates phase information to enhance edge sensitivity. Additionally, a lightweight Tri-Stage Agents Optimization strategy, inspired by reinforcement learning (Wang et al. 2022, 2023), is employed to select reliable agents. These are refined through a Reliable Agents Interaction Module, improving transformer-based coarse matching. Dense correspondences are further refined using high-resolution features, and the final rigid transformation is estimated with PnP + RANSAC.

Iterative Agents Selection Module

We adopt ResNet (He et al. 2016) with an FPN (Lin et al. 2017) and KPFCNN (Thomas et al. 2019) to extract 2D and 3D features with positional encoding, respectively. Images and point clouds use fundamentally different encoding methods based on texture and structure, respectively. This leads to feature discrepancies that hinder effective fusion and reduce registration accuracy.

Phase Map Extractor. In the frequency domain, previous studies (Mildenhall et al. 2021) have shown that the phase component of the Fourier spectrum retains high-level statistical information, which plays a critical role in preserving the structural and textural characteristics of images. Inspired by this, we propose to leverage Fourier phase to enhance the feature correlation between images and point clouds. Specifically, we design a phase map extractor.

Given an image $x \in \mathbb{R}^{H\times W\times 3}$, we apply a two-dimensional Fourier transform $\mathcal{F}(x)$ as:

$$\mathcal{F}(x)_{u,v} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}\cdot e^{-J2\pi\left(\frac{ui}{H}+\frac{vj}{W}\right)}, \tag{1}$$

where J refers to the imaginary unit, and u and v denote the frequency coordinates along the horizontal and vertical axes, respectively. Then we can acquire the corresponding amplitude $\mathcal{B}$ and phase Φ as:

$$\mathcal{B}(x)_{u,v} = \left|\mathcal{F}(x)_{u,v}\right|, \tag{2}$$

$$\Phi(x)_{u,v} = \arg\left(\mathcal{F}(x)_{u,v}\right) = \arctan\left(\frac{\mathrm{Im}(\mathcal{F}(x)_{u,v})}{\mathrm{Re}(\mathcal{F}(x)_{u,v})}\right), \tag{3}$$

where $\mathrm{Im}(\cdot)$ and $\mathrm{Re}(\cdot)$ denote the imaginary and real parts of $\mathcal{F}(x)$, respectively.

To generate the final phase map, we fix the amplitude to an average constant c^x and apply an inverse Fourier transform to obtain the phase texture map ${I_f}^x$:

$${I_f}^x = \mathcal{F}^{-1}\left(\Phi(x)_{u,v} \cdot e^{-Jc^x}\right), \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform.

Then, we extract the phase representation ϕ using a three-layer lightweight CNN adaptor and fuse it with the original image features to obtain the coarse-grained image representation $F_i \in \mathbb{R}^{H\times W\times C}$. The corresponding coarse-grained point cloud feature is denoted as $F_p \in \mathbb{R}^{N\times C}$.

Tri-Stage Agents Optimization Strategy. To address challenges such as repetitive structures, non-overlapping regions, and variations in lighting, we aim to effectively combine image and point cloud features by selecting reliable agents that capture correlations across modalities (see Figure 3). Previous studies (Han et al. 2024) typically rely on a fixed number of initialized agents, which limits adaptability and may hinder optimal feature representation. While reinforcement learning can facilitate the feedback-driven selection of representative and dependable features, an over-reliance on existing representations may impede model performance.

To overcome these limitations, we initialize redundant learnable queries and propose a Tri-Stage Agents Optimization Strategy. This approach allows us to adaptively select the most reliable agents.

Stage I: Warm up Training. We initialize a redundant set of learnable queries, denoted as Q_l. These queries interact with image and point cloud features through n layers of attention, and the resulting aggregated representations are referred to as Q_A (not explicitly shown in the Stage I illustration). Each query is associated with a learnable query score, which is maintained in a dedicated query evaluation block, allowing the model to weigh the importance of each query. We then select the top-k queries based on their query score and refer to them as Q_A^s for subsequent aggregation. The main objective of this stage is to enable the queries Q_l to quickly learn meaningful cross-modal representations, thereby facilitating convergence and establishing a solid foundation for the later stages.

Stage II: Rewards-guided Agents Training. We aim to identify a subset of reliable agents that contribute most to performance improvement. To this end, we analyze the quality of each query from both local and global perspectives, rather than selecting them based on a top-k criterion. At the local level, we evaluate each query based on its similarity to image and point cloud features. Let F_q be the feature of a query, and F_i, F_p be the image and point cloud features, respectively. The local reward is defined as:

$$\mathit{Local\_Reward} = \frac{1}{2}\left[\cos(F_q, F_i) + \cos(F_q, F_p)\right]. \tag{5}$$

At the global level, we assess the contribution of each query by how much it helps reduce the overall task loss. We derive the L_{rl} from the the main task loss and define the global reward as the inverse of L_{rl}:

$$\mathit{Global\_Reward} = 1.0/\mathcal{L}_{rl}, \tag{6}$$

where lower task loss indicates better matching and thus higher global reward. To balance these two types of rewards, we introduce a dynamic weighting strategy controlled by a fusion coefficient α that evolves over time. Specifically, in the early period of Stage II, the model has not yet converged, and global signals offer more informative gradients. As training progresses, we gradually increase the importance of the local reward to encourage convergence and prevent overfitting. We define the fusion coefficient as $\alpha = 1 - e^{-\mathit{epoch}/\tau}$, where *epoch* is the current training iteration and τ is a predefined time constant that controls the growth rate of α. To enhance adaptability and avoid premature convergence, τ is progressively decayed during training with a lower bound, allowing α to gradually increase and adjust the balance between local and global rewards. The final reward is computed as a weighted combination of the local and global rewards using this fusion coefficient.

$$\mathit{reward} = \alpha \cdot \mathit{Local\_Reward} + (1-\alpha) \cdot \mathit{Global\_Reward}. \tag{7}$$

Based on the learned score S_q for each query, we apply a sigmoid function to normalize it to the range $[0, 1]$. We then perform Bernoulli sampling to determine which k queries are selected. The queries that are positively sampled form the set Q_A^r, which will engage in cross-modal interaction and forward propagation.

To stabilize training and maintain gradient flow from the unselected queries, we employ a soft masking strategy. In this approach, each query is assigned a *soft_mask* defined as: $\{\mathit{soft\_mask} = \beta + (1-\beta)\cdot a\}$, where $a \in \{0, 1\}$ is the Bernoulli action, and β is a predefined minimum weight ensuring unselected queries retain a small influence. We then enhance query selection using the reinforce algorithm, with the log-probability for each query computed as:

$$\log P(a) = a \cdot \log(p) + (1-a)\cdot \log(1-p), \tag{8}$$

where $p = \sigma(S_q)$ is the sampling probability. The reinforcement loss is:

$$\mathcal{L}_g = -\left(\mathit{reward} - \overline{\mathit{reward}}\right)\cdot \log P(a), \tag{9}$$

where $\overline{\mathit{reward}}$ denotes the mean of the rewards across the sampled queries, which serves as a baseline to reduce variance. The resulting loss is used to update the query scores S_q through backpropagation.

To encourage exploration and avoid premature convergence of the query selection policy, we incorporate an entropy regularization term on the query sampling probabilities. This entropy bonus promotes diversity by penalizing over-confident selections. The entropy for each query is defined as:

$$\mathit{Entropy} = -\left[p\cdot\log(p) + (1-p)\cdot\log(1-p)\right], \tag{10}$$

the final Stage II loss with entropy regularization becomes:

$$\mathcal{L}_{full} = \mathcal{L}_g - \mu \cdot \sum \mathit{Entropy}, \tag{11}$$

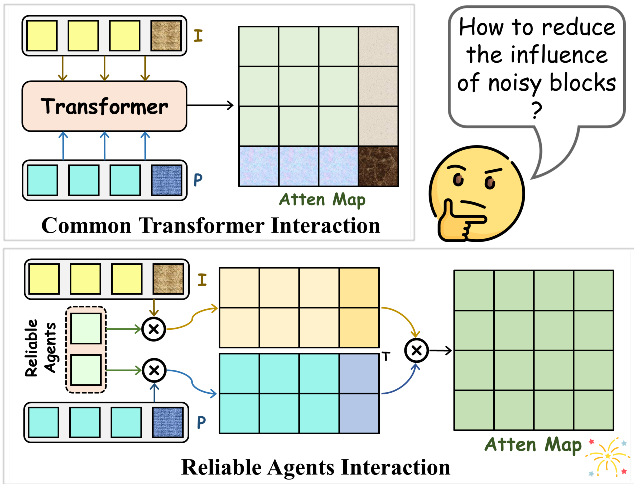

Figure 3: Visualization of difference between common transformer interaction and reliable agents interaction.

where μ is a small weight applied to entropy regularization. By maximizing entropy, the model encourages a balanced exploration–exploitation trade-off. To balance representation learning and query selection, we enable Stage II every 5 epochs after the 15-epoch Stage I, with other epochs continuing as Stage I. This allows periodic refinement of query selection throughout training.

Stage III: Optimal Agents Selection. After training in the first two stages, the model can effectively identify optimal agents. We then use S_Q to select agents that maximally enhance the model's matching for aggregation.

Discussion. *1. Why not train a separate reward model solely for agent selection?* Because integrating the reward mechanism with the main model allows for more efficient training that aligns closely with the task objective, avoiding extra computational overhead and training complexity. Joint training also reduces reward bias, improves the stability and accuracy of agent selection, and enables the selection strategy to adapt and generalize better, thereby enhancing the overall model's matching performance.

2.Why not simply use top-k selection all the time? Top-k selection based on scores is rigid and may miss agents with long-term potential but low initial scores. Using Bernoulli sampling increases exploration of such agents, while a soft mask allows flexible contribution weighting, improving overall matching performance. We allow the model to gradually learn an adaptive selection strategy during training, resulting in better cross-modal matching performance.

3. Why can the Tri-Stage Agents Optimization Strategy select high-quality agents? The Tri-Stage Agents Optimization Strategy improves agent quality by focusing on different goals at each training stage: global matching early, local details mid-stage, and a balanced optimization later. Reinforcement learning with joint training enables diverse query exploration and avoids local optima. Dynamic rewards uncover potential queries, while multi-stage, multi-view training improves stability, diversity, and generalization.

Reliable Agents Interaction Module

After obtaining reliable agents carrying image–point cloud correlation information through the Iterative Agents Selection Module, we leverage these agents as a bridge to refine the transformer aggregation mechanism. As illustrated in Figure 3, standard transformer interaction treats all features equally, often allowing noisy blocks to dominate. In contrast, our reliable agents interaction selects informative features in advance, reducing attention noise and yielding cleaner attention maps with improved feature alignment. By filtering information sources before cross-modal interaction and retaining only high-quality queries for attention computation, the attention becomes more focused, enhancing both the robustness and discriminative power of the learned representations. This strategy effectively suppresses noise from repetitive patterns, non-overlapping regions, and illumination changes, resulting in more accurate image–point cloud matching.

As shown in Figure 2, the reliable agents selected at different stages are denoted as $A \in \mathbb{R}^{k \times C}$, which are used as queries Q, while the image and point cloud features F_i and F_p serve as keys and values. Formally,

$$Q = W^Q \cdot A, K_p = W^P \cdot F_p, V_p = W^P \cdot F_p, \quad (12)$$

$$K_i = W^I \cdot F_i, \quad V_i = W^I \cdot F_i. \quad (13)$$

Here, W^Q, W^P, and W^I are learnable linear projection matrices that map the input features into a shared embedding space for attention computation. We then compute the Image-to-Agent Attention (IAA) and Point-to-Agent Attention (PAA) maps as:

$$\mathrm{IAA} = \sigma\left(QK_p^\top/\sqrt{C}\right), \mathrm{PAA} = \sigma\left(QK_i^\top/\sqrt{C}\right), \quad (14)$$

$\sigma(\cdot)$ denote the softmax function. These attention maps are then used to aggregate V_p and V_i, resulting in the fused features F_p' and F_i'. Patch-level correspondences are established based on feature similarity, followed by dense matching using fine-grained features. Finally, the PnP-RANSAC algorithm is applied to estimate a reliable rigid transformation.

Model Training & Post-Processing Details

Let us examine the loss functions for the coarse and fine-matching networks. Both $\mathcal{L}_{\text{coarse}}$ and $\mathcal{L}_{\text{fine}}$ utilize a general circle loss (Sun et al. 2020; Qin et al. 2022). For a given anchor descriptor d_t, the descriptors of its positive and negative pairs are represented as $\mathcal{D}_t^P$ and $\mathcal{D}_t^N$, respectively. The loss function is defined as follows:

$$\mathcal{L}_t = \frac{1}{\gamma}\log\Bigg[1 + \Big(\sum_{d^j \in \mathcal{D}_t^P} e^{\beta_p^{t,j}(d_t^j - \Delta_p)}\Big) \cdot \Big(\sum_{d^k \in \mathcal{D}_t^N} e^{\beta_n^{t,k}(\Delta_n - d_t^k)}\Big)\Bigg]. \quad (15)$$

The Tri-Stage Agents Optimization strategy is applied only during training. The overall loss function is defined as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_t + \lambda_2 \cdot \mathcal{L}_{full}, \quad (16)$$

where $\mathcal{L}_{full}$ is activated every 5 epochs during Stage II.

Dataset	RGB-D Scenes v2					7-Scenes							
Model	Scene-11	Scene-12	Scene-13	Scene-14	Mean	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Mean
Mdpt(m)	1.74	1.66	1.18	1.39	1.49	1.78	1.55	0.80	2.03	2.25	2.13	1.84	1.49
Inlier Ratio ↑													
FCGF-2D3D	6.8	8.5	11.8	5.4	8.1	34.2	32.8	14.8	26	23.3	22.5	6.0	22.8
P2-Net	9.7	12.8	17.0	9.3	12.2	55.2	46.7	13.0	36.2	32.0	32.8	5.8	31.7
Predator-2D3D	17.7	19.4	17.2	8.4	15.7	34.7	33.8	16.6	25.9	23.1	22.2	7.5	23.4
2D3D-MATR	32.8	**34.4**	39.2	23.3	32.4	72.1	66.0	31.3	60.7	50.2	**52.5**	**18.1**	50.1
B2-3Dnet	**36.4**	32.7	**43.8**	**27.4**	**35.1**	**73.8**	**66.7**	**33.1**	**61.7**	**50.8**	52.3	**18.1**	**50.9**
A2SI	**40.1**	**41.1**	**44.8**	**28.5**	**38.6**	**75.5**	**68.9**	**40.0**	**66.6**	**53.6**	**55.6**	**18.2**	**54.1**
Feature Matching Recall ↑													
FCGF-2D3D	11.1	30.4	51.5	15.5	27.1	**99.7**	98.2	69.9	97.1	83.0	87.7	16.2	78.8
P2-Net	48.6	65.7	82.5	41.6	59.6	**100.0**	99.3	58.9	**99.1**	87.2	92.2	16.2	79
Predator-2D3D	86.1	89.2	63.9	24.3	65.9	91.3	95.1	76.6	88.6	79.2	80.6	31.1	77.5
2D3D-MATR	**98.6**	98.0	**88.7**	77.9	90.8	**100.0**	99.6	**98.6**	**100.0**	**92.4**	**95.9**	58.2	**92.1**
B2-3Dnet	**100.0**	**99.0**	**92.8**	**85.8**	**94.4**	**100.0**	**100.0**	**98.6**	**100.0**	**92.7**	95.6	**64.9**	**93.1**
A2SI	**98.6**	**100.0**	**92.8**	**85.6**	**94.3**	**100.0**	**99.8**	**100.0**	**100.0**	90.6	**96.0**	**65.1**	**93.1**
Registration Recall ↑													
FCGF-2D3D	26.5	41.2	37.1	16.8	30.4	89.5	79.7	19.2	85.9	69.4	79.0	6.8	61.4
P2-Net	40.3	40.2	41.2	31.9	38.4	96.9	86.5	20.5	91.7	75.3	85.2	4.1	65.7
Predator-2D3D	44.4	41.2	21.6	13.7	30.2	69.6	60.7	17.8	62.9	56.2	62.6	9.5	48.5
2D3D-MATR	**63.9**	53.9	58.8	49.1	56.4	96.9	90.7	52.1	95.5	80.9	86.1	28.4	75.8
B2-3Dnet	58.3	**60.8**	**74.2**	**60.2**	**63.4**	**98.3**	**90.5**	**56.2**	**96.4**	**84.0**	**86.1**	**32.4**	**77.7**
A2SI	**69.4**	**76.5**	**87.6**	**58.8**	**73.1**	**97.6**	**95.4**	**68.5**	**98.0**	**81.2**	**90.5**	**28.5**	**79.9**

Table 1: Evaluation results on RGB-D Scenes v2 and 7-Scenes. **Orange** and **Blue** numbers highlight the best, the second best are **Boldfaced** and the baseline are underlined.

Experiments

Datasets and Implementation Details

Based on the 2D3D-MATR (Li et al. 2023) benchmark, we conduct extensive experiments and ablation studies on two challenging benchmarks: RGB-D Scenes v2 (Lai, Bo, and Fox 2014) and 7Scenes (Glocker et al. 2013).

Implementation Details. We used an NVIDIA Geforce RTX 3090 GPU for training. We implement our model using PyTorch 1.13.1. We set the number of attention layers to $n = 3$ by default. The initial τ is set to 20.0 and decayed by a factor of 0.9 every 10 epochs until reaching a minimum of 5.0. We use an entropy regularization weight of $\mu = 0.01$ and set the loss balancing factors to $\lambda_1 = \lambda_2 = 1$. The number of agents is fixed at 12, which also matches the number of queries selected by the top-k strategy. Additionally, we set the soft masking coefficient to $\beta = 0.3$ to stabilize gradient propagation.

Method	Phase	RAI	Tri	Topk	PIR	IR	FMR	RR
M1					48.5	32.5	91.0	56.4
M2	✓				53.4	34.5	91.6	57.5
M3		✓			56.8	37.6	92.8	68.4
M4		✓		✓	58.9	38.2	93.4	70.1
M5		✓	✓		60.2	38.5	94.2	72.8
M6	✓	✓			56.1	37.8	93.1	69.1
M7	✓	✓		✓	59.8	38.3	93.7	70.2
M8	✓	✓	✓		**60.8**	**38.6**	**94.3**	**73.1**

Table 2: Ablation study results on RGB-D Scenes v2. Phase indicates the use of phase map enhancement, RAI denotes the Reliable Agents Interaction module, Tri refers to the Tri-Stage Agents Optimization, and Topk represents the Top-k selection strategy. **Orange** numbers highlight the best, the second best are underlined.

Metrics. We evaluate models with three metrics: Inlier Ratio (IR) — the percentage of pixel-point matches with a 3D distance below 5 cm; Feature Matching Recall (FMR) — the percentage of image–point-cloud pairs with an inlier ratio above 10%; and Registration Recall (RR) — the percentage of image–point-cloud pairs with an RMSE below 10 cm. Patch Inlier Ratio (PIR) evaluates coarse-level alignment by measuring the fraction of patch correspondences with sufficient overlap under GT transformation.

Evaluations on Dataset

We compare our approach with 2D3D-MATR (Li et al. 2023) and other baselines (Choy, Park, and Koltun 2019; Wang et al. 2021; Huang et al. 2021; Cheng et al. 2025b) on the RGB-D Scenes v2 and 7-Scenes datasets (Table 1).

On the RGB-D Scenes v2 dataset, our method achieves

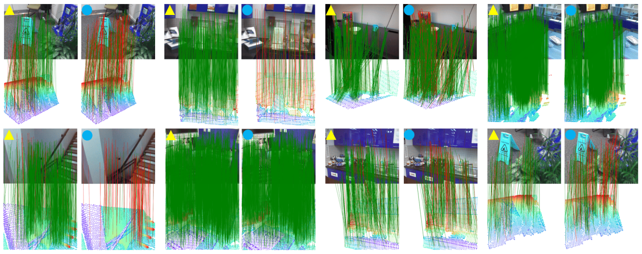

Figure 4: Visualization of matching accuracy: correct (green) if error < 60 px; otherwise red. Triangles represent A2SI, and circles represent Baseline.

notable gains in narrow indoor environments such as Scene-11, Scene-12, and Scene-13, where geometric ambiguities often cause mismatches. By leveraging edge information from the IAS phase map and emphasizing critical regions, the RAI module enhances correspondence reliability, leading to a 28.8 pp increase in registration recall on Scene-12 (from 58.8% to 87.6%). Overall, our method improves the mean inlier ratio by 6.2 pp, feature matching recall by 3.5 pp, and registration recall by 16.7 pp.

On the 7-Scenes dataset, which exhibits significant scale variations, our method consistently outperforms previous approaches, achieving a mean registration recall of 79.9%, surpassing 2D3D-MATR by 4.4 percentage points. While performance improvements in the *Stairs* scene are limited due to repetitive textures, notable gains are observed in the *Heads* and *Kitchen* scenes, with recall increases of +16.6 pp and +4.4 pp, respectively. These results highlight the robustness and strong generalization of our approach across diverse indoor environments.

Ablation Studies

To assess the contribution of each module, we conduct an ablation study on the RGB-D Scenes v2 dataset (Table 2). Using the phase map (M2) improves registration recall by 1.1% over the baseline. Adding the RAI module (M3) further enhances performance, while the Tri-Stage Agents Optimization (M4) achieves the largest gains by progressively refining correspondences through multi-stage selection and optimization. Compared with M4, the Top-k strategy (M5) yields smaller improvements due to its lack of refinement across stages. M6, which replaces adaptive selection with fixed queries, performs worse, indicating the advantage of dynamic selection. M7, adopting a simple Top-k scheme on redundant queries, outperforms M6 but remains inferior to M4. Finally, the full model (M8) integrates all modules, achieving the best overall performance and confirming their complementary effects.

Figure 5: Visualization of A2SI.

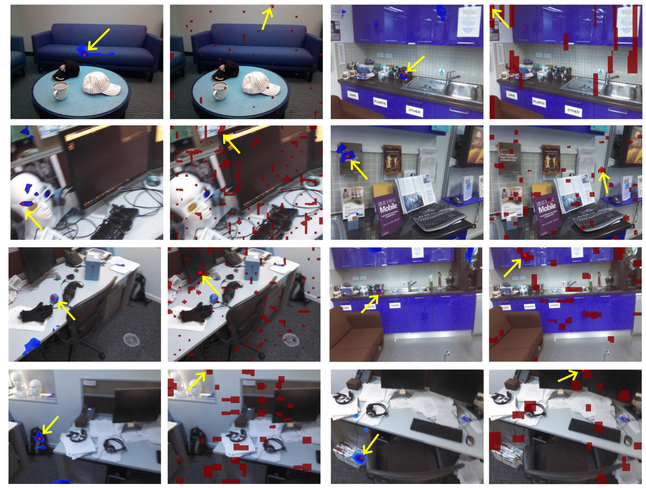

Figure 6: Visualization of differences in attention focus.

Qualitative Results

We further conduct visualization analyses to intuitively demonstrate the model's performance. In Figure 4, a match is considered correct if the projected point cloud keypoint lies within 60 pixels of its image correspondence. The yellow triangles denote A2SI, while the blue circles represent 2D3D-MATR. Benefiting from improved feature aggregation and edge information, our method shows more correct (green) and fewer incorrect (red) matches, indicating superior robustness. In Figure 5, projecting the point cloud onto the image reveals accurate matches on structural objects such as tables, cabinets, and chairs, demonstrating the model's ability to capture reliable geometric correspondences. In Figure 6, attention maps from the last transformer layer show that our model focuses on key regions, emphasizing salient objects rather than distributing attention uniformly. This focused attention yields more meaningful features, enhancing both matching accuracy and robustness in complex indoor scenes.

Conclusion

In this paper, we introduce a novel Adaptive Agent Selection and Interaction (A2SI) Network for image-to-point cloud registration. Our method enhances features by incorporating edge information from phase maps and adopts reinforcement learning-inspired strategies to select informative agents. We further aggregate image and point cloud features through reliable agents, which improves registration accuracy and reduces computational complexity. Extensive experiments on the RGB-D Scenes v2 and 7-Scenes datasets demonstrate that our A2SI approach surpasses existing methods in image-to-point cloud registration.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFB3909902) and the Youth Innovation Promotion Association of the Chinese Academy of Sciences. We gratefully acknowledge their financial support, which made this research possible. I also deeply appreciate the emotional and mental support from the people around me, whose encouragement and understanding have been indispensable throughout this journey.

References

Barroso-Laguna, A.; Riba, E.; Ponsa, D.; et al. 2019. Key.net: Keypoint detection by handcrafted and learned CNN filters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5836–5844.

Chang, H.; Zhu, R.; Chang, W.; Yu, M.; Liang, Y.; Lu, J.; Li, Z.; and Zhang, T. 2025. MeshSplat: Generalizable Sparse-View Surface Reconstruction via Gaussian Splatting. arXiv:2508.17811.

Chen, Y.; Sun, R.; Li, W.; Mai, H.; Luo, N.; Pan, Y.; and Zhang, T. 2025. Alleviate and mining: Rethinking unsupervised domain adaptation for mitochondria segmentation from pseudo-label perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2339–2347.

Cheng, Z.; Deng, J.; Li, X.; Yin, B.; and Zhang, T. 2025a. B2-3D++: Uncertainty-aware Hierarchical Registration Network with Domain Alignment. TechRxiv. Preprint.

Cheng, Z.; Deng, J.; Li, X.; Yin, B.; and Zhang, T. 2025b. Bridge 2D-3D: Uncertainty-aware Hierarchical Registration Network with Domain Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2491–2499.

Cheng, Z.; Deng, J.; Li, X.; Yin, X.; Liao, B.; Yin, B.; Yang, W.; and Zhang, T. 2025c. CA-I2P: Channel-Adaptive Registration Network with Global Optimal Selection. *arXiv preprint arXiv:2506.21364*.

Choy, C.; Park, J.; and Koltun, V. 2019. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8958–8966.

Deng, J.; Lu, J.; and Zhang, T. 2024. Unsupervised Template-assisted Point Cloud Shape Correspondence Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5250–5259.

Durrant-Whyte, H.; and Bailey, T. 2006. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine*, 13(2): 99–110.

Feng, M.; Hu, S.; Ang, M. H.; and Lee, G. H. 2019. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, 4790–4796. IEEE.

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

Glocker, B.; Izadi, S.; Shotton, J.; and Criminisi, A. 2013. Real-time RGB-D camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 173–179. IEEE.

Han, D.; Ye, T.; Han, Y.; Xia, Z.; Pan, S.; Wan, P.; Song, S.; and Huang, G. 2024. Agent attention: On the integration of softmax and linear attention. In *European Conference on Computer Vision*, 124–140. Springer.

He, J.; Deng, J.; Zhang, T.; Zhang, Z.; and Zhang, Y. 2023. Hierarchical shape-consistent transformer for unsupervised point cloud shape correspondence. *IEEE Transactions on Image Processing*, 32: 2734–2748.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, S.; Gojcic, Z.; Usvyatsov, M.; Wieser, A.; and Schindler, K. 2021. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 4267–4276.

Kang, S.; Liao, Y.; Li, J.; Liang, F.; Li, Y.; Li, F.; Dong, Z.; and Yang, B. 2023. CoFiI2P: Coarse-to-Fine Correspondences for Image-to-Point Cloud Registration. *arXiv preprint arXiv:2309.14660*.

Lai, K.; Bo, L.; and Fox, D. 2014. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 3050–3057. IEEE.

Lepetit, V.; Moreno-Noguer, F.; and Fua, P. 2009. EP n P: An accurate O (n) solution to the P n P problem. *International journal of computer vision*, 81: 155–166.

Li, M.; Qin, Z.; Gao, Z.; Yi, R.; Zhu, C.; Guo, Y.; and Xu, K. 2023. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14128–14138.

Liao, B.; Zhai, W.; Wan, Z.; Cheng, Z.; Yang, W.; Zhang, T.; Cao, Y.; and Zha, Z.-J. 2024. Ef-3dgs: Event-aided free-trajectory 3d gaussian splatting. *arXiv preprint arXiv:2410.15392*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Moheimani, S. R.; Vautier, B. J.; and Bhikkaji, B. 2006. Experimental implementation of extended multivariable PPF control on an active structure. *IEEE Transactions on Control Systems Technology*, 14(3): 443–455.

Ng, P. C.; and Henikoff, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13): 3812–3814.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; Ilic, S.; Hu, D.; and Xu, K. 2023. Geotransformer: Fast and robust point cloud registration with geometric transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9806–9821.

Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11143–11152.

Ren, S.; Zeng, Y.; Hou, J.; and Chen, X. 2022. CorrI2P: Deep image-to-point cloud registration via dense correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3): 1198–1208.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, 2564–2571. Ieee.

Rusu, R. B.; Blodow, N.; and Beetz, M. 2009. Fast point feature histograms (FPFH) for 3D registration. In *2009 IEEE international conference on robotics and automation*, 3212–3217. IEEE.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938–4947.

Simon, T.; Joo, H.; Matthews, I.; et al. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1145–1153.

Sontag, E. D. 1998. Comments on integral variants of ISS. *Systems & Control Letters*, 34(1-2): 93–100.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; and Zhou, X. 2021. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922–8931.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6398–6407.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Torii, A.; Sivic, J.; Pajdla, T.; and Okutomi, M. 2013. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 883–890.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Chen, C.; Cui, Z.; Qin, J.; Lu, C. X.; Yu, Z.; Zhao, P.; Dong, Z.; Zhu, F.; Trigoni, N.; et al. 2021. P2-net: Joint description and detection of local features for pixel and point matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16004–16013.

Wang, C.; Yang, W.; Liu, X.; and Zhang, T. 2025a. State Space Model Meets Transformer: A New Paradigm for 3D Object Detection. In *The Thirteenth International Conference on Learning Representations*.

Wang, C.; Yang, W.; and Zhang, T. 2023. Not every side is equal: Localization uncertainty estimation for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3814–3824.

Wang, C.; Zha, Y.; Yang, W.; and Zhang, T. 2025b. StruMamba3D: Exploring Structural Mamba for Self-supervised Point Cloud Representation Learning. *arXiv preprint arXiv:2506.21541*.

Wang, Y.; He, X.; Peng, S.; Tan, D.; and Zhou, X. 2024a. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21666–21675.

Wang, Z.; Pan, T.; Zhou, Q.; and Wang, J. 2023. Efficient Exploration in Resource-Restricted Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 10279–10287.

Wang, Z.; Wang, J.; Zhou, Q.; Li, B.; and Li, H. 2022. Sample-Efficient Reinforcement Learning via Conservative Model-Based Actor-Critic. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 8612–8620.

Wang, Z.; Wang, J.; Zuo, D.; Ji, Y.; Xia, X.; Ma, Y.; Hao, J.; Yuan, M.; Zhang, Y.; and Wu, F. 2024b. A Hierarchical Adaptive Multi-Task Reinforcement Learning Framework for Multiplier Circuit Design. In *Forty-first International Conference on Machine Learning*. PMLR.

Wu, Q.; Jiang, H.; Luo, L.; Li, J.; Ding, Y.; Xie, J.; and Yang, J. 2024. Diff-Reg v1: Diffusion Matching Model for Registration Problem. *arXiv preprint arXiv:2403.19919*.

Yu, H.; Li, F.; Saleh, M.; Busam, B.; and Ilic, S. 2021. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34: 23872–23884.

Zha, Y.; Wang, C.; Yang, W.; and Zhang, T. 2025a. Exploring Semantic Masked Autoencoder for Self-supervised Point Cloud Understanding. *arXiv preprint arXiv:2506.21957*.

Zha, Y.; Wang, C.; Yang, W.; Zhang, T.; and Wu, F. 2025b. Exploring Vision Semantic Prompt for Efficient Point Cloud Understanding. In *Forty-second International Conference on Machine Learning*.