Occlusion-Robust Relative Pose Estimation for Multi-Robot Systems via Geometric-Aware Diffusion Matching

Suyoung Kang[1†], Rishav Dutta[1†], Peng Gao[2], Maggie Wigness[3], John Rogers[3], Donghyun Kim[1], and Hao Zhang[1]

***Abstract*— Relative pose estimation is crucial for coordinated multi-robot navigation. However, robots in close proximity often face intra-team occlusions, where teammates partially block each other's field of view, while dynamic environments further introduce environmental occlusions. Classical relative pose estimation methods degrade under occlusion and texture scarcity, whereas learning-based methods often lack explicit geometric consistency, which limits their accuracy during real deployments. To address multi-robot relative pose estimation in complex 3D environments, we introduce *Geometric-Aware Diffusion Matching* (GADM), which enables a team of robots to estimate relative 6-DoF poses using only RGB-D sensors, even under occlusions. GADM uses a diffusion model to progressively exploit global and higher-order structural constraints encoded by a graph network, guiding smoother optimization and faster convergence to robust correspondence distributions under noise and occlusions. By integrating geometric consistency, GADM explicitly addresses occlusions by producing geometrically consistent matches suitable for real-time deployment on physical robots. The resulting correspondences are then used with geometry-based solvers to estimate 6-DoF relative poses, providing robustness even under partial view overlap and limited keypoint visibility. We conducted experiments using both robotics simulations and physical robot teams, and our results show that GADM achieves robust 6-DoF pose estimation performance in occluded scenarios.**

More details are provided on the project website: `https://gadm2026.github.io`.

I. Introduction

Multi-robot systems play a vital role in robotics research, enabling collaborative tasks across applications such as environmental monitoring, search and rescue, and autonomous driving [1]–[5]. One of the key capabilities for such collaboration is relative pose estimation, determining the full 6-DoF pose of one robot with respect to another using only onboard sensors, without any global reference such as GNSS or motion capture. While GNSS, LiDAR, and UWB sensing have advanced multi-robot localization, they face limitations in indoor settings, near reflective surfaces, or under challenging lighting condition. RGB-D sensors, by contrast, are low-cost, energy-efficient, and well-suited for real-time applications. Accurate real-time relative pose estimation with RGB-D is therefore critical for enabling coordinated behaviors such as formation control, collision avoidance, and collaborative mission execution.

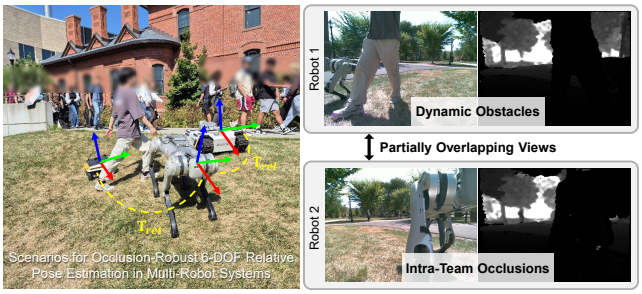


Fig. 1. A motivating scenario for occlusion-robust relative pose estimation. Multi-robot teams operating in formation frequently encounter occlusions caused not only by environmental obstacles, such as dynamic pedestrians, but also by other robots within the team (i.e., intra-team occlusions). These challenges motivate methods that can reliably estimate 3D relative poses even when only limited and partially overlapping visual information is available.

In real-world environments, which are often unstructured, cluttered, and dynamic, estimating the relative poses of robot teammates remains a significant challenge. For example, as illustrated in Fig. 1, when a team of three robots navigates an urban environment in a wedge formation, their fields of view (FOV) typically partially overlap and differ in viewpoint; these views may be occluded not only by environmental obstacles such as moving pedestrians but also by other robots within the team (i.e., intra-team occlusion). These challenges often reduce the amount of shared observable information between robots and exacerbate the difficulty of correctly associating observations across the team.

Given the importance of relative pose estimation, several methods have been proposed for multi-robot systems. Classical learning-free methods [6]–[8] employ geometric feature matching or the Iterative Closest Point (ICP) algorithm [9], [10] for data association. However, their accuracy degrades in dynamic scenes, on textureless surfaces, or under occlusion. Recently, learning-based approaches have shown promising performance in relative pose estimation [11]–[14]. However, they often do not explicitly consider geometric consistency among robots, leading to sub-optimal and unreliable estimates in real-world deployments. Crucially, neither classical nor learning-based methods effectively handle intra-team occlusions, which are especially severe in tight multi-robot formations and close-proximity operations.

*This work was partially supported by NSF CAREER Award IIS-2308492, DARPA Young Faculty Award (YFA) D21AP10114-00, DEVCOM ARL TBAM CRA W911NF2520024, and A2I2 CRA W911NF2320005.

[1]Suyoung Kang, Rishav Dutta, Donghyun Kim, and Hao Zhang are with the University of Massachusetts Amherst, Amherst, MA 01002, USA. Email: {suyoungkang, rishavdutta, donghyunkim, hao.zhang}@umass.edu.

[2]Peng Gao is with North Carolina State University, Raleigh, NC, 27695, USA. Email: pgao5@ncsu.edu.

[3]Maggie Wigness and John Rogers are with the U.S. Army DEVCOM Army Research Laboratory (ARL), Adelphi, MD 20783, USA. Email: {maggie.b.wigness, john.g.rogers59}.civ@army.mil.

[†]Authors contributed equally to this paper.

To address multi-robot relative pose estimation in complex 3D environments, we introduce *Geometric-Aware Diffusion Matching* (GADM), which enables a team of robots to estimate each other's 6-DoF relative poses using only RGB-D sensors, even under occlusions. GADM uses a diffusion process that progressively denoises the correspondence matrix between feature points extracted from each robot's observations, thereby generating a denser and more reliable set of keypoint matches. To further improve correspondence quality, we incorporate geometric constraints that suppress spurious matches caused by occlusions or sensor noise. By integrating a diffusion model with geometric consistency, GADM explicitly addresses occlusions by producing geometrically consistent matches suitable for real-time deployment on physical robots. The resulting correspondences are then used with geometry-based solvers to estimate 6-DoF relative poses, which provide robustness even under partial view overlap and limited keypoint visibility. This map-free formulation makes the system particularly well-suited for multi-robot teams operating in GPS-denied or cluttered outdoor environments, where building and maintaining maps is expensive or infeasible.

The primary contribution of this paper is the introduction of GADM, a novel learning-based approach for relative pose estimation in multi-robot systems. This work presents two key specific novelties:

1) We introduce a novel learning-based method that integrates a diffusion process with geometric constraints within a unified learning framework, explicitly addressing occlusions, including dynamic and intra-team occlusions, under partial view overlap among robots.
2) We significantly advance multi-robot capabilities for relative pose estimation, with the new ability to estimate 6-DoF poses in 3D space, operate without GPS, run in real time, and execute in a decentralized manner on physical robot teams, all while maintaining robustness to occlusions and partial view overlap.

II. Related Works

A. Multi-robot SLAM and Map-free Relocalization

Multi-robot relative pose estimation methods generally fall into two paradigms: *SLAM-based* and *map-free* methods.

SLAM-based methods. Traditional multi-robot SLAM systems maintain a global map or pose graph by fusing data across multiple agents, enabling large-scale collaborative localization and mapping. Centralized frameworks such as LAMP 2.0 [15] and CCM-SLAM [16] aggregate multi-robot data at a central server and build unified global map. Distributed systems like Kimera-Multi [13] achieve globally consistent metric-semantic mapping by integrating multisensory information across robots. While effective in producing globally consistent maps from each robot's sensory information, both methods are not suitable for real-time multi-robot collaboration in dynamic environments due to heavy computational demands.

Map-free methods. Map-free relocalization approaches avoid building a global map during deployment by using compact scene representations. This strategy supports long-term operation with lower computational demands [17]. For example, CoViS-Net [11] encodes sensor images and performs uncertainty-aware relative pose regression, enabling estimates even under partially or non-overlapping fields of view. However, regression-based approaches often lack explicit geometric consistency, leading to reduced accuracy and robustness under occlusions.

B. Feature-Based Relative Pose Estimation

Vision-based relative pose estimation methods, both monocular and RGB-D, can be divided into *classical correspondence-based* approaches and *learning-based* methods.

Classical approaches. These methods establish correspondences using hand-crafted features and geometric validation. Keypoint-based visual association methods such as SIFT [8] and ORB [18] provide a foundation for geometric spatial understanding. They have been shown to provide accurate robot pose estimation when a sufficient number of inlier correspondences are available [7]. However, their performance degrades substantially in the presence of occlusions, textureless regions, or highly dynamic environments.

Learning-based approaches. Recent methods use neural networks to regress pose directly from images [19], [20], demonstrating robustness to wide baselines and large viewpoint changes. However, these methods often lack geometric awareness and have been limitedly validated in real-time multi-robot deployments. In parallel, learned geometric features and matchers such as SuperPoint [21], SuperGlue [22], LightGlue [23], LoFTR [24], and Dust3r [25] have shown strong correspondence performance across viewpoint and illumination changes. Yet even these methods struggle when overlap is sparse or heavily occluded.

C. Diffusion Models for Geometry and Correspondence

Diffusion models refine predictions through iterative denoising and have recently shown strong results across geometric vision tasks that benefit from multi-step refinement. For camera pose estimation and localization, PoseDiffusion [26] formulates bundle adjustment within a probabilistic diffusion framework for multi-view SfM. DiffusionSfM [27] predicts camera pose and dense scene structure through diffusion over ray origins/endpoints, demonstrating end-to-end improvements on multi-view benchmarks. These works show that iterative denoising can mirror and augment classical geometric solvers rather than relying on one-shot regression.

Diffusion has also been applied to feature correspondence and registration. DiffGlue [28] introduces a diffusion-aided sparse matcher that denoises an initial assignment matrix with assignment-guided attention inside a GNN, driving the solution toward a doubly-stochastic correspondence and improving downstream homography estimation. For dense matching, DiffMatch models both data and prior terms via a conditional diffusion process and achieves state-of-the-art results on dense correspondence benchmarks [29]. These diffusion-driven refinements help escape local minima, stabilize training, and produce more reliable poses and correspondences than one-shot estimators.

Despite this progress, existing diffusion methods in geometric vision predominantly address single-agent or pairwise settings. To our knowledge, diffusion has not yet been tailored for multi-robot relative pose estimation with explicit mechanisms for intra-team and environmental occlusions.

III. Approach

A. Problem Formulation

We denote RGB and depth images as I and D, respectively. Given paired sensory inputs from RGB-D cameras mounted on two robots, denoted as (I_A, D_A) and (I_B, D_B), the goal is to estimate the relative 6-DoF transformation $\mathbf{T}_{rel} = (\mathbf{R}, \mathbf{t}) \in SE(3)$, which defines the pose of robot B with respect to robot A, where $\mathbf{R} \in SO(3)$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector. Given each image I, we use a feature extraction network to produce a set of candidate visual keypoints with coordinates $\mathbf{C}$ and their corresponding descriptors $\mathbf{D}$.

Classical keypoint matching for relative pose estimation is typically formulated as selecting the nearest descriptor between $\mathbf{D}_A$ and $\mathbf{D}_B$, using similarity metrics such as Euclidean or Hamming distance. Matching is performed via brute-force or nearest-neighbor search, where the keypoint with the closest descriptor is considered the correspondence.

Once keypoint correspondences are obtained, 2D keypoints in the RGB images are back-projected into 3D space using the corresponding depth measurements. Specifically, let $\mathbf{X}_j \in \mathbb{R}^3$ denote the 3D point from depth D_B corresponding to keypoint j, and $\mathbf{x}_i$ the associated 2D observation in I_A. Relative pose estimation between a pair of robots is then formulated as an optimization problem that minimizes reprojection error [30]:

$$(\mathbf{R}^*, \mathbf{t}^*) = \arg\min_{\mathbf{R},\mathbf{t}} \sum_{(i,j)\in\mathcal{M}} \left\|\Pi(\mathbf{R}\mathbf{X}_j + \mathbf{t}) - \mathbf{x}_i\right\|^2, \quad (1)$$

where $\Pi(\cdot)$ is the camera projection function that maps 3D points in the robot's coordinate frame to 2D pixel coordinates in the camera frame mounted on the robot, and $\mathcal{M}$ denotes the set of matched features. The solution $(\mathbf{R}^*, \mathbf{t}^*)$ represents the relative pose transformation between robots A and B.

However, the classical matching approaches have several limitations. First, because it relies solely on local descriptor similarity, it often produces many outliers and fails to enforce global structural consistency. Second, it is particularly vulnerable to occlusions, as large portions of descriptors may be missing or corrupted, resulting in unreliable matches and degraded pose estimation.

B. Diffusion-based Graph Matching

To enable relative pose estimation in multi-robot systems, we introduce a novel learning-based GADM approach, as illustrated in Fig. 2, which aims to learn a distribution over correspondence assignment matrices $\mathbf{P} = \{P_{ij}\}$, where $P_{ij} \in [0, 1]$ is the probability of matching between the i-th descriptor in $\mathbf{D}_A$, and the j-th descriptor in $\mathbf{D}_B$, computed from observations acquired by a pair of robots. Instead of making a single-shot prediction of correspondences, we formulate the correspondence estimation problem as a diffusion process over a graph-based denoising network guided by a multi-objective learning regime.

Diffusion Process: We adopt a denoising diffusion probabilistic model (DDPM) [31] to refine the assignment matrix. Starting from a clean ground-truth assignment $\mathbf{P}_0$, the forward process corrupts it over T steps by progressively adding Gaussian noise, producing a sequence $\{\mathbf{P}_t\}_{t=1}^T$. Given a variance schedule $\{\beta_t\}_{t=1}^T$, which controls the rate of noise injection, the forward diffusion step is defined as:

$$p(\mathbf{P}_t \mid \mathbf{P}_{t-1}) = \mathcal{N}\Big(\mathbf{P}_t; \sqrt{1-\beta_t}\,\mathbf{P}_{t-1},\, \beta_t\mathbf{I}\Big). \quad (2)$$

The variance schedule ensures that noise is introduced smoothly, enabling stable training and accurate reversal.

The reverse process learns to iteratively recover a cleaner assignment matrix $\mathbf{P}_{t-1}$ from its noisier counterpart $\mathbf{P}_t$. This step is parameterized by a denoiser network $\mathcal{D}_\theta(\cdot)$. We can derive the input of next denoising step $\mathbf{P}_{t-1}$ by substituting $\mathbf{P}_0$ with $\hat{\mathbf{P}}_0 = \mathcal{D}_\theta(\mathbf{P}_t)$:

$$p(\mathbf{P}_{t-1} \mid \mathbf{P}_t, \hat{\mathbf{P}}_0) \approx \mathcal{N}\Big(\mathbf{P}_{t-1}; \mu_t(\mathbf{P}_t, \mathcal{D}_\theta(\mathbf{P}_t)), \tilde{\beta}_t\mathbf{I}\Big). \quad (3)$$

During inference, this iterative denoising transforms a noisy prior $\mathbf{P}_T$ into a refined assignment $\hat{\mathbf{P}}_0$, which encodes the predicted correspondences.

During training, the denoiser is supervised with a mean-squared error loss:

$$\mathcal{L}_{diff} = \mathbb{E}_{t\sim[1,T],\mathbf{P}_t\sim p(\mathbf{P}_t|\mathbf{P}_0)} \left\|\mathcal{D}_\theta(\mathbf{P}_t) - \mathbf{P}_0\right\|^2, \quad (4)$$

where $\mathbf{P}_0$ is the ground-truth assignment matrix. At each step, $\mathbf{P}_t$ is sampled along the denoising trajectory with added Gaussian noise, encouraging broader exploration of the solution space. This progressive refinement allows the network to exploit global pairwise and higher-order structural constraints, guiding it along a smoother optimization trajectory. As a result, the model converges faster to confident correspondence distributions that remain robust under noise and occlusions.

Graph Neural Network Denoiser: The denoiser network $\mathcal{D}_\theta$ is implemented as a graph neural network (GNN) with attention layers, which is designed to leverage both intra-image structure and inter-image relationships. Specifically, to make the diffusion model guided by the assignment matrix score for better connectivity, we use an attention-like module that uses denoised P_t instead of the attention weight [28]. This GNN denoiser layer is stacked L times, with l denoting the layer index.

First, each timestep t is encoded with an embedding function $\mathcal{E}(t)$ [32], producing $\tau_t \in \mathbb{R}$ that conditions the denoising process:

$$\tau_t = \mathcal{E}(t), \quad t \in [1, T]. \quad (5)$$

To model intra-graph structure, descriptors from each image are projected into the embedding space, e.g., $\mathbf{F}_A^0 = \mathcal{E}_{desc}(\mathbf{D}_A)$, and refined using self-attention layers that capture spatial and contextual dependencies:

$$\mathbf{F}_{A,self}^l = \mathcal{G}_{self}^l(\mathbf{F}_A^{l-1}, \mathbf{F}_A^{l-1}). \quad (6)$$

Then, to explicitly incorporate correspondence priors $\mathbf{P}_t$, we use a layer similar to cross-attention to inject assignment priors

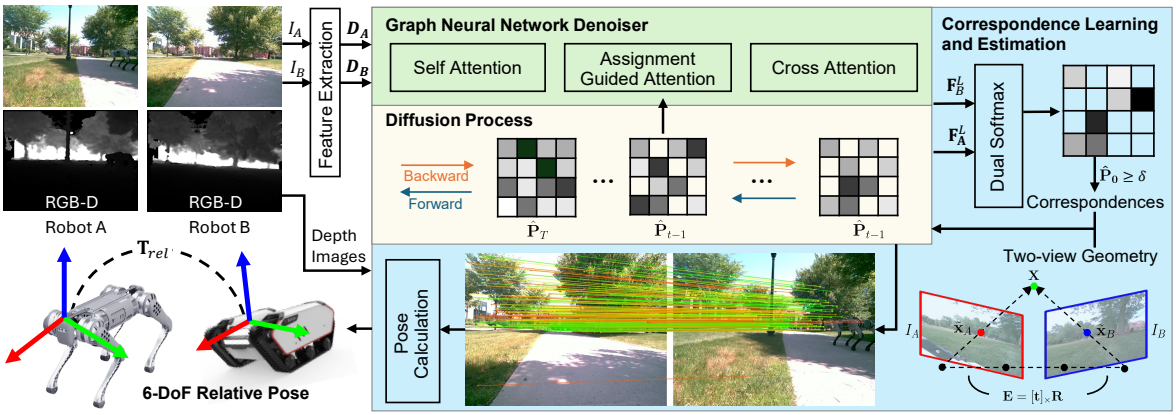


Fig. 2. Overview of our novel GADM approach for enabling occlusion-robust 6-DoF relative pose estimation in multi-robot systems through geometric-aware diffusion matching.

to guide the diffusion process [28]. Instead of relying solely on standard attention weights, the current assignment estimate $\mathbf{P}_t$ used as the correspondence prior is injected into the attention mechanism, yielding assignment-guided embeddings:

$$\mathbf{F}^l_{A,assign} = \mathcal{G}^l_{assign}(\mathbf{F}^l_{A,self} + \tau_t, \mathbf{F}^l_{B,self} + \tau_t, \mathbf{P}_t). \quad (7)$$

Finally, cross-attention layers are used to enable inter-graph communication, allowing features from the two graphs to interact. Merged embeddings are first constructed by:

$$\mathbf{F}^l_{A,merge} = \mathcal{E}_{merge}(\mathbf{F}^l_{A,assign} || \mathbf{F}^{l-1}_A), \quad (8)$$

where $||$ denotes concatenation, and then cross-attention updates are applied:

$$\mathbf{F}^l_{A,cross} = \mathcal{G}^l_{cross}(\mathbf{F}^l_{A,merge}, \mathbf{F}^l_{B,merge}). \quad (9)$$

This process is repeated across L layers, which produces the final embedding $\mathbf{F}_A \in \mathbb{R}^{N_A \times d}$, where N_A is the number of keypoints and d is the dimensionality of the embedding.

Correspondence Estimation and Learning: Given $\mathbf{F}_A$ from robot A and $\mathbf{F}_B$ from robot B, we compute the correspondence assignment matrix $\mathbf{P}$ using a dual-softmax projection by:

$$\mathbf{P} = \text{Softmax}(\mathbf{F}_A \mathbf{F}_B^\top) \odot \text{Softmax}(\mathbf{F}_B \mathbf{F}_A^\top)^\top, \quad (10)$$

where $\odot$ is the Hadamard product. This operation captures both local and global similarity scores while enforcing mutual consistency of correspondences, effectively representing all possible pairings between feature points in the two images. It further normalizes the similarity scores along both rows and columns, producing a doubly stochastic matrix.

To enhance robustness, we incorporate a matchability prediction for the keypoints. Specifically, we compute a matchability score as $\boldsymbol{\sigma} = \text{Sigmoid}(\mathcal{E}(\mathbf{F})) \in [0,1]^N$, where $\mathcal{E}(\cdot)$ denotes an MLP that projects a vector to a single dimension. This score serves as a weight, assigning higher values to better-matched keypoints. We then compute an augmented assignment matrix $\widetilde{\mathbf{P}}$ by weighting $\mathbf{P}$ in Eq. (10) with matchability scores through $\widetilde{\mathbf{P}} = \boldsymbol{\sigma}_A^\top \boldsymbol{\sigma}_B \odot \mathbf{P}$. This suppresses poor matches by down-weighting low-confidence keypoints while reinforcing high-confidence ones, resulting in a more accurate and discriminative correspondence distribution. During training, given a set of ground-truth correspondences $\mathcal{M}_{gt}$, we use the augmented matrix $\widetilde{\mathbf{P}}$ and the matchability scores $\boldsymbol{\sigma}_A$ and $\boldsymbol{\sigma}_B$ to define the matching loss:

$$\mathcal{L}_{match} = -\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{1}{|\mathcal{M}_{gt}|}\sum_{(i,j)\in\mathcal{M}_{gt}} \log \widetilde{P}_{ij}^{(\ell)} + \frac{1}{2|\bar{I}|}\sum_{i\in\bar{I}}\log\left(1-\sigma_{A,i}^{(\ell)}\right) + \frac{1}{2|\bar{J}|}\sum_{j\in\bar{J}}\log\left(1-\sigma_{B,j}^{(\ell)}\right)\right), \quad (11)$$

where $\bar{I}, \bar{J}$ denote unmatched points in A and B, respectively. The first term maximizes the likelihood of correct matches, while the second and third terms encourage unmatched points to receive low matchability scores.

C. Geometric Consistency Loss

While diffusion refines the assignment matrix distribution, correspondence quality must also meet the geometric constraints inherent to two-view geometry [33]. We incorporate a geometric consistency loss into our GADM training, which utilizes the epipolar constraint between two different views of a 3D scene from RGB-D images acquired by a pair of robots.

Let $\mathbf{x}_i \in \mathbf{C}_A$ and $\mathbf{x}_j \in \mathbf{C}_B$ be the coordinates of a pair of keypoints in the 2D image space that are matched according to the correspondence assignment matrix $\mathbf{P}$. We calculate the essential matrix $\mathbf{E} \in \mathbb{R}^{3\times3}$ to encode the epipolar geometry between two calibrated cameras mounted on a pair of robots, as follows:

$$\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}, \quad (12)$$

where $(\mathbf{R}, \mathbf{t}) \in SE(3)$ denotes the ground-truth relative pose used during the training phase, and $[\mathbf{t}]_\times$ is the skew-symmetric matrix of the translation vector $\mathbf{t}$, which is a square matrix whose transpose is equal to its negative, i.e., $[\mathbf{t}]_\times^\top = -[\mathbf{t}]_\times$.

For a valid correspondence between $\mathbf{x}_i$ and $\mathbf{x}_j$, the epipolar constraint requires:

$$\hat{\mathbf{x}}_j^\top \mathbf{E} \hat{\mathbf{x}}_i = 0, \quad (13)$$

where $\hat{\mathbf{x}} = \mathbf{K}^{-1}\tilde{\mathbf{x}}$ are normalized homogeneous coordinates, $\mathbf{K}$ is the camera intrinsic matrix, and $\tilde{\mathbf{x}} = [\mathbf{x}; 1]$. Intuitively, this equation states that if a 3D point is observed at pixel $\mathbf{x}_i$ in image I_A, its corresponding point $\mathbf{x}_j$ in image I_B must lie on the epipolar line defined by $\mathbf{E}\hat{\mathbf{x}}_i$. This epipolar constraint ensures that two matched keypoints are geometrically consistent with the relative pose of the robots.

However, directly optimizing the epipolar constraint in Eq. (13) is generally suboptimal due to its sensitivity to noise and scale. To address this, we employ a Sampson Epipolar Distance [26], which provides a first-order approximation of the geometric reprojection error by linearizing the epipolar constraint around each correspondence. Given a pair of corresponding points $\mathbf{x}_i$ and $\mathbf{x}_j$, Sampson Epipolar Distance is defined as below:

$$\mathcal{L}_{geom} = \sum_{(i,j)\in\mathcal{M}} \frac{(\hat{\mathbf{x}}_j^\top \mathbf{E}\hat{\mathbf{x}}_i)^2}{(\mathbf{E}\hat{\mathbf{x}}_i)_1^2 + (\mathbf{E}\hat{\mathbf{x}}_i)_2^2 + (\mathbf{E}^\top\hat{\mathbf{x}}_j)_1^2 + (\mathbf{E}^\top\hat{\mathbf{x}}_j)_2^2} \tag{14}$$

where $(\cdot)_1$ and $(\cdot)_2$ denote the first and second components of a vector. This suppresses geometrically invalid matches, thus improving the robustness of 6-DoF relative pose estimation, particularly under occlusion conditions where many features lack valid correspondences. Moreover, it is numerically more stable and provides a more accurate reflection of the geometric error compared to direct epipolar constraint optimization.

By integrating the loss $\mathcal{L}_{geom}$ modeling geometric constraints with the diffusion and matching losses, we obtain the final loss function used to train GADM for 6-DoF relative pose estimation between a pair of robots:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \lambda_1\mathcal{L}_{match} + \lambda_2\mathcal{L}_{geom}, \tag{15}$$

where λ_1 and λ_2 are the weights used to balance the losses.

During training, we optimize $\mathcal{L}_{total}$ using Adam [34], an adaptive gradient-based optimizer that computes parameter-specific learning rates by maintaining running estimates of the first- and second-order moments of the gradients. Adam offers faster convergence, improved stability in the presence of noisy gradient signals, and better handling of the non-stationary loss landscape typical of diffusion models compared to plain stochastic gradient descent. After learning using $\mathcal{L}_{total}$, our approach outputs the final correspondence assignment matrix $\mathbf{P}$. Consequently, the set of correspondences $\mathcal{M}$ used for relative pose calculation is obtained as follows:

$$\mathcal{M} = \{(i,j) \mid \mathbf{P}_{ij} \geq \delta\}, \tag{16}$$

where $\delta \in [0, 1]$ denotes a confidence threshold that controls the trade-off between accepting more matches and rejecting uncertain ones.

IV. Experimental Results

A. Experimental Setup

For feature extraction, we use the state-of-the-art SuperPoint network [21] due to its robustness to viewpoint and lighting variations as well as its efficiency in real-time robotic applications. We train our model using synthetic image pairs from HPatches [35], a dataset containing ground-truth correspondences, and the model learns to denoise corrupted assignment matrices with a diffusion loss and matching loss. The model is then fine-tuned on MegaDepth [36], which consists of real-world images, allowing the model to learn geometrically consistent matches using the geometric consistency loss. We configure the hyperparameters of our approach as $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\delta = 0.1$. For inference, we employ DDIM [37] with two denoising steps to reduce computation while maintaining accuracy. The relative pose is then computed by iteratively solving Eq. (1) using the Levenberg–Marquardt algorithm, with a RANSAC inlier threshold of 3.0 pixels [30].

We evaluate GADM in a multi-robot deployment in both Gazebo simulations and a team of real physical robots. To evaluate performance under varying occlusion scenarios, we consider two representative scenarios: (1) intra-team occlusions and (2) manually introduced occlusions. We compare GADM against two representative baselines:

- **CoViS-Net** [11], which is an end-to-end pose regression network for multi-robot relative pose estimation,
- **ORB-BF** [18], a classical pipeline combining ORB feature extraction, brute-force matching, and a geometry-based solver for pose recovery.

Performance is quantitatively evaluated using ground-truth relative poses aligned with the estimated poses at corresponding timestamps. We compute the absolute trajectory error (**ATE**) [38], and absolute rotation error (**ARE**) [39]. ATE is the absolute distance between the estimated translation and the true translation vector. ARE is the absolute angle difference between the estimated rotation and the true rotation matrix. Both errors are summarized in terms of root mean square errors (**RMSE**) to provide a total measure of accuracy. For fair comparison, the accuracy is calculated only on synchronized frame pairs for which all three methods successfully produce pose estimates.

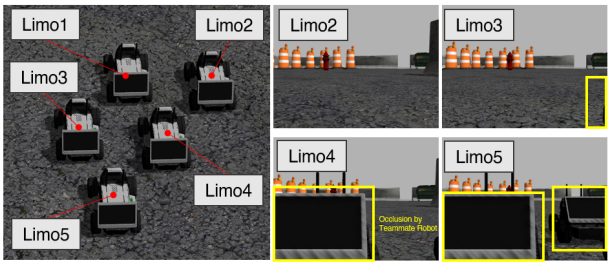


Fig. 3. Examples of intra-team occlusions. The left image illustrates a multi-robot formation control scenario, while the right images show ego-views from four robots where teammates partially obstruct the field of view with varying degrees of occlusion.

B. Results under Occlusions in Gazebo Simulations

We evaluate our approach in Gazebo to demonstrate robustness under two representative occlusion scenarios: (1) intra-team occlusions and (2) manually introduced occlusions.

For intra-team occlusions (Fig. 3), a team of five Limo robots follows predefined trajectories. *Limo1* serves as the

leader, broadcasting query images, while *Limo2–Limo5* estimate relative poses under increasing occlusion rates (approximately $0\%, 22.5\%, 25\%$, and 37.5%). We test two motion patterns: *Team_translation* (forward motion with formation) and *Team_rotation* (forward motion with yaw).

For manually introduced occlusions, two robots execute trajectories (*Circleloop*, *Curvyloop*, *Farminspection*), where occlusions are synthetically applied as black masks on RGB images (Fig. 4).

Table I shows that GADM achieves the lowest translation (ATE) and rotation (ARE) errors across occlusion rates. While all methods degrade under occlusion, GADM remains stable. ORB-BF matches our accuracy in rare cases but suffers from a low success rate (27% vs. 99%). CoViS-Net achieves 100% success but incurs larger errors due to a lack of geometric consistency.

Trajectory visualizations in Fig. 6 further highlight robustness under occlusions: GADM (green) produces stable estimates near ground truth, CoViS-Net (blue) shows drift, and ORB-BF (red) is intermittently accurate but frequently fails, evidenced by missing poses. As occlusion rates increase, GADM maintains accurate pose estimation while other methods deteriorate. Additionally, trajectory visualizations in Fig. 7 confirm that GADM delivers superior results even in scenarios with 0% occlusion, where robots follow different trajectories during navigation.

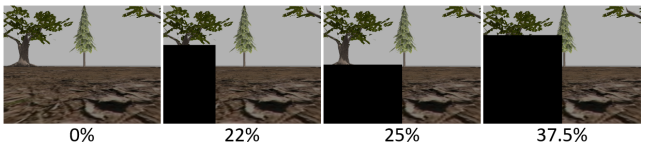


Fig. 4. Examples of manually introduced occlusions with varying degrees of severity in a simulation, introduced by synthetic black masks.

TABLE I

QUANTITATIVE RESULTS OF ATE RMSE (UNIT: M) AND ARE RMSE (UNIT: DEG) ON GAZEBO PER OCCLUSION RATE.

Occ Rate	Method	circleloop		curvyloop		farminspection		team_translation		team_rotation	
		ATE	ARE	ATE	ARE	ATE	ARE	ATE	ARE	ATE	ARE
0%	CoViS-Net	0.55	7.64	0.70	12.51	1.01	13.48	0.27	2.56	0.42	4.19
	ORB-BF	0.35	2.65	0.17	1.09	0.50	2.93	0.27	3.15	0.20	1.58
	Ours	**0.13**	**1.03**	**0.11**	**0.86**	**0.37**	**2.10**	**0.15**	**0.91**	**0.09**	**0.59**
22%	CoViS-Net	0.65	11.86	0.57	12.05	1.18	19.03	0.41	2.16	0.42	7.82
	ORB-BF	0.94	12.59	0.27	1.76	0.94	4.77	0.30	2.82	0.72	5.06
	Ours	**0.18**	**1.59**	**0.15**	**1.35**	**0.33**	**1.84**	**0.11**	**0.63**	**0.11**	**0.87**
25%	CoViS-Net	0.77	16.47	0.70	12.48	1.23	16.53	0.49	5.89	0.62	10.28
	ORB-BF	0.42	7.71	0.17	1.09	0.51	2.87	**0.07**	0.61	0.43	2.84
	Ours	**0.15**	**1.31**	**0.11**	**0.85**	**0.39**	**2.24**	**0.07**	**0.47**	**0.09**	**0.73**
37.5%	CoViS-Net	1.40	37.06	0.92	23.64	1.49	18.89	0.97	4.15	0.81	4.10
	ORB-BF	1.04	19.48	0.47	3.31	0.46	2.26	0.93	7.98	0.36	2.15
	Ours	**0.24**	**3.22**	**0.10**	**0.92**	**0.40**	**2.21**	**0.19**	**1.27**	**0.15**	**1.01**

C. Results under Occlusions in Indoor Environments

We further validate the proposed framework on hardware using a team of two robots. We also generate the manual occlusion as in Fig. 5. Quantitative results are summarized in Table II. We evaluate three motion scenarios: *Indoorslope*, in which one robot ascends a slope while the second remains stationary to expose the full 6-DoF relative motion; and *Indoorloop1* and *Indoorloop2*, in which the robots traverse looped trajectories within a motion-capture room. While all methods degrade with increasing occlusion, our approach (GADM) consistently attains the lowest errors and maintains accuracy across occlusion levels. CoViS-Net and ORB-BF incur large, high-variance errors under occlusion, whereas GADM preserves stably low absolute errors.

TABLE II

QUANTITATIVE RESULTS OF ATE RMSE (UNIT: M) AND ARE RMSE (UNIT: DEG) ON REAL-WORLD INDOOR SCENARIO PER OCCLUSION RATE.

Occ Rate	Method	indoorslope		indoorloop1		indoorloop2	
		ATE	ARE	ATE	ARE	ATE	ARE
0%	CoViS-Net	1.39	15.32	1.16	24.55	1.30	36.53
	ORB-BF	**1.00**	15.60	**0.98**	10.18	0.56	9.84
	Ours	1.01	**14.92**	**0.98**	**10.17**	**0.51**	**9.20**
22%	CoViS-Net	1.54	16.80	1.18	22.50	1.20	31.36
	ORB-BF	**1.02**	15.27	**0.97**	10.46	0.55	9.69
	Ours	1.05	**14.90**	0.99	**10.05**	**0.50**	**8.94**
25%	CoViS-Net	1.52	16.44	1.24	20.54	1.25	39.72
	ORB-BF	**1.03**	15.60	0.99	10.17	0.54	9.67
	Ours	1.04	**14.82**	**0.98**	**10.12**	**0.52**	**9.21**
37.5%	CoViS-Net	1.73	17.08	1.65	22.66	1.22	32.09
	ORB-BF	**1.03**	15.60	**0.99**	10.26	**0.51**	9.33
	Ours	1.05	**14.87**	**0.99**	**9.97**	**0.49**	**8.82**

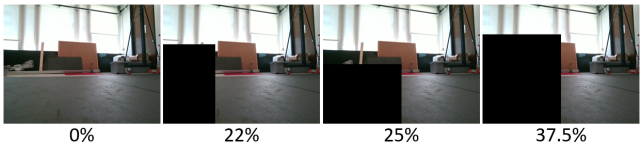
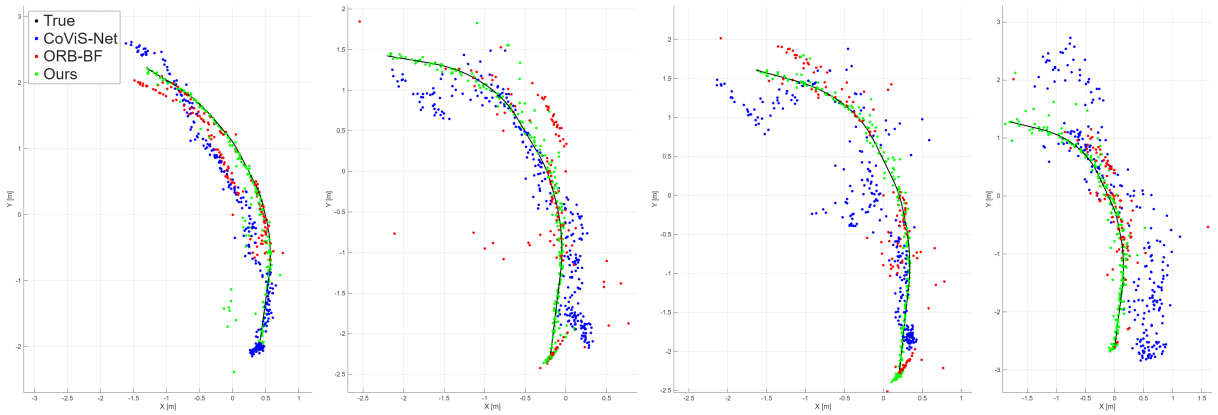


Fig. 5. Examples of manually introduced occlusions with varying degrees of severity in a real-world environment.

D. Results on Physical Robots in Real Outdoor Scenarios

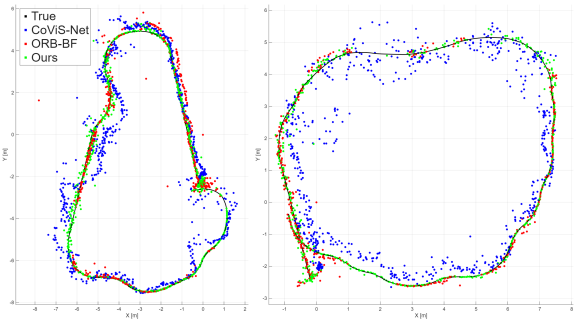
We further validate the proposed system in a GPS-denied outdoor deployment that emphasizes platform-agnostic operation and closed-loop control. The environment poses several practical challenges: robots traverse uneven, mixed-compliance terrain (grass, compacted soil, and pavement), adopt wedge formations that induce frequent mutual occlusions, and share space with pedestrians and moving objects, which introduce additional, time-varying environmental occlusions (Fig. 1). Unlike controlled indoor settings, illumination varies throughout the runs and scene geometry is cluttered, stressing both appearance- and geometry-based components of the pipeline.

During physical deployment, the trained GADM model can run in real time at $\sim$10 Hz (trained model runs at $\sim$10 Hz standalone, but the full perception–control loop runs at $\sim$5 Hz) on an Intel i7-13620H CPU with an NVIDIA RTX 4050 GPU mounted on the B1 legged robot. To implement a leader-follower wedge formation, we set the B1 robot to serve as the leader, while the tracked Bunker and wheeled Jackal robots act as followers. Each follower robot (Jackal and Bunker) broadcasts its observations to the leader (B1), which estimates the relative poses using GADM. With synchronized RGB-D observations streamed to the remote node, B1 performs pairwise relative pose estimation via GADM and broadcasts the resulting estimates back to the follower robots. This configuration preserves a map-free design: no pre-built map, GPS, or infrastructure is required beyond standard time synchronization and sensor-to-base extrinsics calibration. By avoiding global mapping and relying solely on inter-robot observations, the system remains decentralized and agnostic to platform morphology, as long as extrinsics are known.



(a) limo2 (Occ Rate 0%) (b) limo3 (Occ Rate 22%) (c) limo4 (Occ Rate 25%) (d) limo5 (Occ Rate 37.5%)

Fig. 6. Visualization of map-free pose estimation using a team of five Limo robots under varying levels of intra-team occlusion in the Gazebo simulation, shown in the global coordinate frame. Black indicates ground truth, while colored dots represent per-method estimates.



(a) curvyloop (b) circleloop

Fig. 7. Visualization of map-free pose estimation in Gazebo simulations, where robots follow different trajectories during navigation.

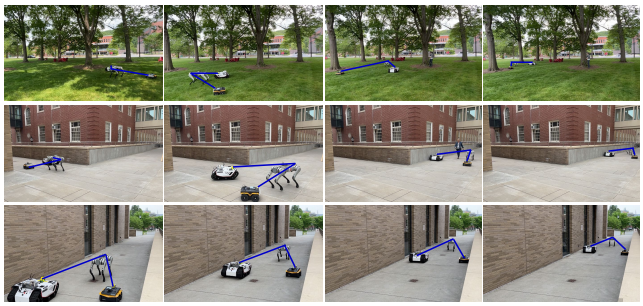


Fig. 8. Qualitative results from real-world deployments of a physical multi-robot system executing wedge formation across different scenarios. Blue lines denote the robots' relative positions within the formation.

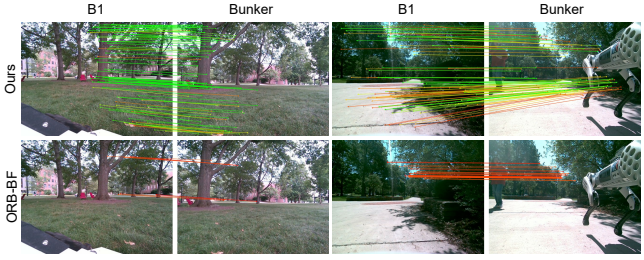


Fig. 9. Qualitative comparison of feature matching. Each column shows the view of each robot. The top row shows ORB features with BF matching results, while the bottom row shows our GADM feature matching results. Green lines indicate inliers or high-confidence matches, and red lines indicate outliers or low-confidence matches.

We implement a leader-follower strategy to demonstrate downstream utility. The leader (B1) tracks a predefined trajectory specified in its local frame. The followers (Jackal and Bunker) maintain a fixed formation by regulating a desired relative pose with respect to the leader, parameterized by a position offset and a yaw (heading) offset. At each update, a pose error between the desired offset and the GADM-estimated relative pose is fed to a PID controller to generate linear and angular velocity commands. To ensure robust behavior under transient occlusions and dynamic clutter, the controller consumes estimates only when the associated correspondence quality is adequate (as reflected by GADM's matching consistency/inlier support); otherwise, the follower temporarily holds the previous command until the estimate quality recovers. This gating prevents abrupt control reactions to low-confidence measurements while retaining responsiveness when visibility improves.

As illustrated in Fig. 8, the robots consistently maintain stable wedge formations across diverse scenes, despite uneven terrain and intermittent mutual and environmental occlusions. The sustained formation indicates that GADM's relative pose estimates are both accurate and sufficiently stable for real-time control. Fig. 9 further illustrates correspondence quality: GADM (top row) yields dense, geometrically consistent matches (green inliers) that translate into reliable pose updates for the controller, whereas ORB-BF (bottom row) produces sparse and error-prone matches in the presence of occlusions and moving distractors, often precipitating pose-estimation failures. Together, these results substantiate that a map-free, GPS-denied, platform-agnostic localization pipeline can support robust multi-robot formation control under realistic outdoor conditions.

V. Conclusion

In this paper, we propose GADM to enable occlusion-robust 6-DoF relative pose estimation for multi-robot systems. We formulate the correspondence estimation problem as a diffusion process over a graph-based denoising network guided by a multi-objective learning regime. By integrating

geometric consistency, GADM explicitly addresses occlusions by producing geometrically consistent matches suitable for real-time deployment on physical robots. The resulting correspondences are then used with geometry-based solvers to estimate 6-DoF relative poses, providing robustness even under partial view overlap and limited keypoint visibility. We validated GADM under both intra-team and synthetic occlusions, demonstrating strong performance across diverse multi-robot scenarios. Outdoor experiments further show its capability to support closed-loop formation control in GPS-denied, cluttered environments without relying on a global map. In future work, we aim to extend GADM beyond pairwise interactions to larger teams with joint consistency constraints and improved perception–control for multi-robot coordination.

REFERENCES

[1] Z. Deng, P. Gao, W. J. Jose, C. Reardon, M. Wigness, J. Rogers, and H. Zhang, "Coordinated multi-robot navigation with formation adaptation," in *International Conference on Robotics and Automation (ICRA)*, 2025.

[2] W. J. Jose and H. Zhang, "Learning for dynamic subteaming and voluntary waiting in heterogeneous multi-robot collaborative scheduling," in *International Conference on Robotics and Automation (ICRA)*, 2024.

[3] P. Gao, R. Guo, H. Lu, and H. Zhang, "Correspondence identification for collaborative multi-robot perception under uncertainty," *Autonomous Robots*, vol. 46, no. 1, pp. 5–20, 2022.

[4] P. Gao, S. Siva, A. Micciche, and H. Zhang, "Collaborative scheduling with adaptation to failure for heterogeneous robot teams," in *International Conference on Robotics and Automation (ICRA)*, 2023.

[5] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *IEEE Access*, vol. 8, pp. 191 617–191 643, 2020.

[6] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[7] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] M. J. Schuster, C. Brand, H. Hirschmüller, M. Suppa, and M. Beetz, "Multi-robot 6d graph slam connecting decoupled local reference filters," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[10] S. Li and D. Lee, "Rgb-d slam in dynamic environments using static point weighting," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2263–2270, 2017.

[11] J. Blumenkamp, S. Morad, J. Gielis, and A. Prorok, "Covis-net: A cooperative visual spatial foundation model for multi-robot applications," in *Conference on Robot Learning (CoRL)*, 2025.

[12] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[13] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 1067–1086, 2022.

[14] R. Scona, M. Jaimez, Y. R. Petillot, M. Fallon, and D. Cremers, "Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[15] Y. Chang, K. Ebadi, C. E. Denniston, M. F. Ginting, A. Rosinol, A. Reinke, M. Palieri, J. Shi, A. Chatterjee, B. Morrell, *et al.*, "Lamp 2.0: A robust multi-robot slam system for operation in challenging large-scale underground environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9175–9182, 2022.

[16] P. Schmuck and M. Chli, "Ccm-slam: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams," *Journal of Field Robotics*, vol. 36, no. 4, pp. 763–781, 2019.

[17] R. G. Goswami, N. Patel, P. Krishnamurthy, and F. Khorrami, "Flashmix: Fast map-free lidar localization via feature mixing and contrastive-constrained accelerated training," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[19] C. Rockwell, N. Kulkarni, L. Jin, J. J. Park, J. Johnson, and D. F. Fouhey, "Far: Flexible, accurate and robust 6dof relative camera pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] R. Cai, B. Hariharan, N. Snavely, and H. Averbuch-Elor, "Extreme rotation estimation using dense correlation volumes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[21] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[22] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[24] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] J. Wang, C. Rupprecht, and D. Novotny, "Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[27] Q. Zhao, A. Lin, J. Tan, J. Y. Zhang, D. Ramanan, and S. Tulsiani, "Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[28] S. Zhang and J. Ma, "Diffglue: Diffusion-aided image feature matching," in *ACM International Conference on Multimedia (ACM MM)*, 2024.

[29] J. Nam, G. Lee, S. Kim, H. Kim, H. Cho, S. Kim, and S. Kim, "Diffusion model for dense matching," in *International Conference on Learning Representations (ICLR)*, 2024.

[30] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[32] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[33] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[36] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.

[38] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[39] Y. Zhou, L. Kneip, C. Rodriguez, and H. Li, "Divide and conquer: Efficient density-based tracking of 3d sensors in manhattan worlds," in *Asian Conference on Computer Vision (ACCV)*, 2016.