A Multi-Modal 6DoF Pose Estimation Framework of Rebar Intersections Based on Fuzzy-Oriented FoundationPose

Bolong Shu[1], Lu Deng[1,2], Jichao Hou[3,4], and Jingjing Guo[1,2,*]

[1]College of Civil Engineering, Hunan University, China
[2]Hunan Provincial Key Laboratory for Damage Diagnosis for Engineering Structures, Hunan University, China
[3]State Grid Jiaxing Electric Power Supply Company
[4]Jiaxing Hengguang Electric Power Construction Co., Ltd
* Corresponding Author
shubolong@hnu.edu.cn, denglu@hnu.edu.cn, jchou@hnu.edu.cn, guojingjing@hnu.edu.cn

Abstract –

Rebar tying operations are highly repetitive and labor-intensive, presenting significant potential for automation. Although several rebar tying robots have already been deployed, in scenarios involving vertical or inclined rebar cages, conventional single-degree-of-freedom actuators prove inadequate, thereby necessitating the adoption of multi-axis robotic arm solutions. This introduces the novel problem of six-degree-of-freedom (6DoF) pose estimation for rebar intersections. However, due to the sensing noise of low-cost depth camera when capturing slender rebar structures, traditional pose estimation methods often struggle to deliver reliable results in complex construction environments. To address this challenge, this study leverages the robustness and contextual awareness advantages of deep learning to propose a pose estimation framework tailored for rebar cage intersections tying. The framework integrates object detection and pixel-level segmentation to enable tying sequence planning. Building upon FoundationPose, an observation-angle-guided efficient pose estimation method, termed Fuzzy-Oriented FoundationPose, is introduced, which enhances inference speed while preserving accuracy, and effectively resolves ambiguities in equivalent pose outputs for symmetric objects. This framework has been validated on diverse geometric body and rebar intersection datasets, demonstrating robust effectiveness and applicability in real-world rebar cage scenarios.

Keywords –

Rebar tying; Construction robot; 6DoF pose estimation; Symmetry suppression

1 Introduction

Rebar tying is a critical process in reinforced concrete construction, with its quality directly determining the spatial positioning accuracy of the reinforcement and the overall performance of the structure. Prolonged engagement in high-intensity, rebar tying operations is prone to inducing musculoskeletal injuries in workers [1], while such tasks align closely with the inherent advantages of robots in handling repetitive activities. Consequently, various rebar tying robots have emerged in recent years, including TyBOT [2], T-iROBO [3], and the crawling tying robot [4]. Although these systems have demonstrated certain automation capabilities for horizontal rebar grids, they generally rely on "flat-laying" construction conditions and struggle to adapt to the vertical, inclined, and complex posture rebar grids which are commonly encountered in engineering sites. To enable robots to perform stable tying operations in multi-posture rebar scenarios, the accurate acquisition of the six-degree-of-freedom (6DoF) pose of rebar intersections becomes an indispensable key capability.

Existing rebar tying robots are typically equipped with low-cost consumer-grade depth cameras. In construction sites, rebar grids often have slender geometries, multilayered arrangements, and highly reflective surfaces, resulting in substantial depth sensing noise. Traditional pose estimation methods that rely solely on depth information are highly sensitive to such noise and struggle to provide reliable estimates in complex environments. Conversely, methods that rely only on RGB images underutilize the available visual cues, failing to exploit the discriminative edge and texture features of rebar structures [5]. In contrast, rebar intersections combine distinctive geometric configurations with recognizable local visual features, offering favorable conditions for robust pose estimation that integrates both depth and visual information.

Motivated by the limitations of conventional methods in handling noisy depth data and extracting discriminative visual features, this paper proposes a 6DoF pose estimation framework for rebar intersections and introduces a novel algorithm based on

FoundationPose [6], i.e., Fuzzy-Oriented FoundationPose. The framework leverages object detection results and pixel-level data to generate intersection-region masks and facilitate convenient planning of the tying sequence. The proposed algorithm incorporates observation-angle-guided pose constraints, using coarsely inferred viewing directions to restrict the rotation initialization space. This approach not only enhances inference efficiency but also effectively suppresses equivalent pose ambiguities caused by symmetry, leading to improved processing speed.

2 Related Works

Early pose estimation methods primarily rely on detecting local feature points on object surfaces and establishing correspondences between these features and the object model. Commonly used local feature extractors include SIFT [7], SURF [8], and ORB [9]. However, due to the highly repetitive textures present in rebar cage images, such methods often fail to produce reliable feature correspondences, leading to inaccurate or failed pose estimation.

Beyond feature-based methods, template-based point cloud registration methods estimate the 6DoF pose of an object by solving for the optimal rigid transformation between the source and target point clouds. Among these methods, the Iterative Closest Point algorithm is one of the most representative approaches, whose main idea is to iteratively refine the transformation parameters under the current alignment until the two point sets converge to an optimal registration state. However, such methods require a reasonably accurate initial pose; in complex industrial scenarios with irregular data distributions, an inaccurate initialization often leads to a local minimum. In addition, existing studies have shown that consumer-grade active stereo depth cameras (e.g., Intel RealSense D435i) perform poorly on non-Lambertian surfaces like reinforcing steel bars [10,11], which manifests as substantial noise and millimeter-scale errors along the optical axis in the captured point clouds, presenting a major challenge for template-based point cloud registration methods.

In addition, image template-based 6DoF pose estimation methods determine an object's pose by comparing the observed image against a set of prerendered templates depicting the object under various poses [10]. The pose associated with the most similar template is selected. However, for targets with high repetition and symmetry, such as rebar intersections, this approach is prone to ambiguity. Furthermore, the pose accuracy is inherently limited by the discrete sampling of the pose space, which restricts its practical application in engineering scenarios.

With the rapid advancement of deep learning, 6DoF pose estimation methods based on single RGB images have emerged, typically relying on keypoint detection or dense correspondence prediction followed by PnP for pose computation. However, their accuracy degrades significantly in scenarios involving weak textures, repetitive patterns, or reflective surfaces [11]. With the integration of depth information and the increase in computational capacity, RGB-D multimodal fusion networks have demonstrated improved accuracy and robustness by jointly leveraging appearance and geometric cues. Representative approaches include DenseFusion [12], which performs pixel-wise fusion of RGB and point cloud features for pose regression; PVN3D [13], which predicts keypoints directly through 3D convolutions; and FFB6D [14], which encodes local geometric neighborhoods using graph convolutions for end-to-end estimation. Compared with RGB-only methods, these models exhibit superior robustness under occlusion and in complex industrial environments, yet they remain costly to train and highly dependent on large annotated datasets. To improve generalization, FoundationPose [6] aligns pretrained universal feature templates with observed data at the pixel level, enabling category-level pose estimation without object-specific retraining while maintaining high robustness and accuracy. Nevertheless, its inference speed remains relatively slow, and its output poses for symmetric objects are not guaranteed to be unique, which may introduce interference risks in practical engineering applications.

3 Method

The proposed pose prediction framework consists of three modules as shown in Figure 1: (1) Image preprocessing; (2) Mask generation; (3) 6DoF pose estimation.

3.1 Image Preprocessing

This section presents the preprocessing pipeline applied to the raw data. Since the proposed pose estimation framework requires the segmentation mask of each individual rebar intersection as input, additional processing is necessary after data acquisition. In typical rebar tying scenarios, the rebar cage consists of two layers, and the captured data inevitably contains structural overlap between them. Therefore, the raw data must be separated and filtered to extract the tying layer and eliminate interference from non-target layers.

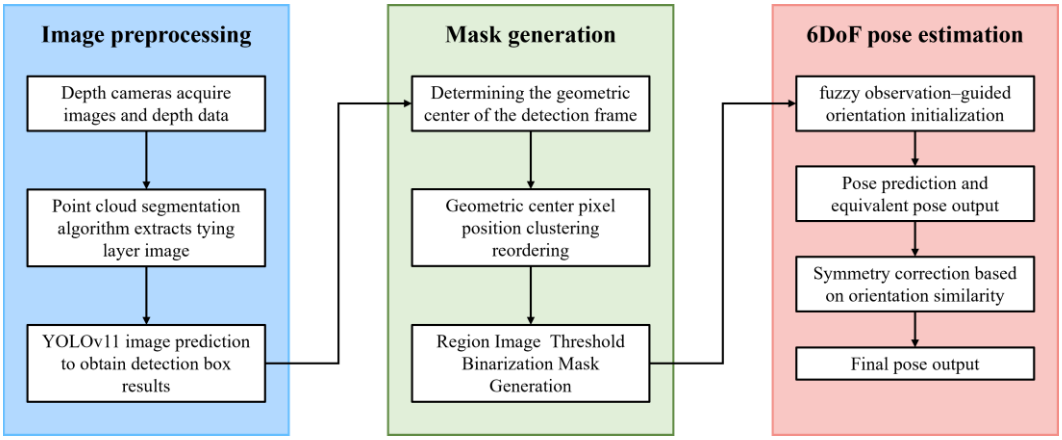

Figure 1. The workflow of the proposed method

3.1.1 Point Cloud Segmentation Algorithm

As illustrated in Figure 2, the dual-layer rebar cage point clouds exhibit a clustered distribution along the optical axis. To isolate the tying layer, the point cloud is first analyzed by computing a height histogram along the z-axis, where peak detection is used to automatically identify the primary dense layer closest to the sensor. Subsequently, candidate point clouds within the depth interval corresponding to the detected peak are extracted, and a best-fit plane is estimated using RANSAC [15]. By indexing the point cloud, the associated image pixels are retrieved to obtain the rebar grid image of the tying layer.

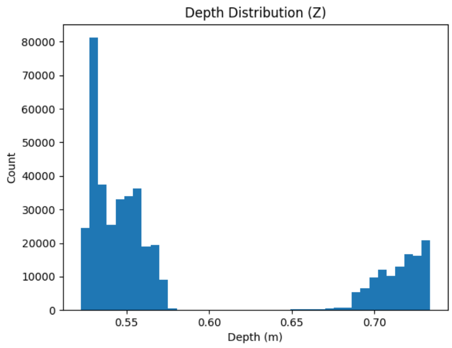

Figure 2. Z-axis height histogram of the dual-layer rebar mesh point cloud

3.1.2 Dataset Construction

Currently, publicly available datasets for intersection detection are scarce. Since the proposed method operates on preprocessed rebar grid images, manual annotation of bounding boxes for rebar intersections is required. As shown in Figure 3, given the critical influence of dataset quality on model performance, three key strategies were adopted during data acquisition and construction: (1) scene complexity, covering variations in lighting conditions and viewing angles; (2) quality diversification, including differences in image sharpness and noise levels; (3) variation in rebar diameters, to enhance the model's generalization across different rebar specifications. In total, an annotated dataset comprising 800 images of rebar intersections was constructed.

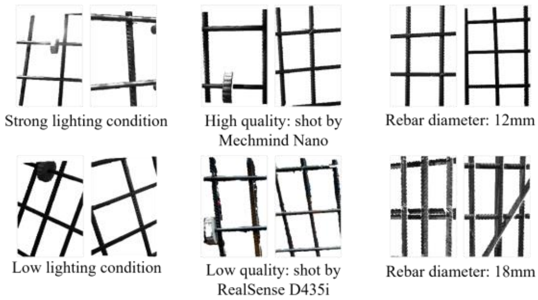

Figure 3. Selected images from the rebar grids dataset

3.1.3 Rebar Intersection Target Detection

A YOLO11n model was trained on the constructed rebar intersection dataset. The training was conducted using the official Ultralytics [16] implementation, with all input images resized to 640×640. To enhance robustness against illumination variations and reflective surfaces commonly encountered in construction environments, several data augmentation strategies were applied, including random scaling, rotation, color jittering, and noise injection. The model was trained for 200 epochs with a batch size of 8. The training hyperparameters were set using an initial learning rate of 0.01 decayed via cosine annealing schedule, momentum of 0.937, and weight decay of 0.005. The training results indicate stable convergence, and the obtained model serves as a reliable detection module within the proposed framework.

3.2 Mask generation

In automated rebar tying tasks, an S-shaped tying path can significantly reduce robot motion compared with the conventional Z-shaped pattern. To achieve this, an intersection reordering method is illustrated in Figure 4.

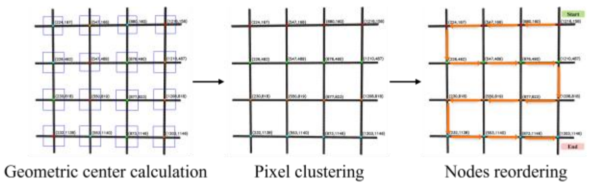

Figure 4. Intersection reordering methods and tying order illustration

3.2.1 Bounding Box Ordering Based on Geometric Centers

The process begins by computing the geometric centers of all detected intersections from the YOLO bounding boxes and obtaining their pixel coordinates. A clustering step is then applied, in which centers whose y-coordinate differences are smaller than 50 pixels are grouped into the same cluster. This procedure effectively aggregates intersections belonging to the same horizontal row. Within each row, an alternating ordering strategy is used: for odd-indexed rows, intersections are sorted in ascending order of x-coordinate (corresponding to a left-to-right tying direction), whereas for even-indexed rows, intersections are sorted in descending order of x-coordinate (corresponding to a right-to-left tying direction). Such an alternating strategy naturally forms a continuous S-shaped traversal path. The underlying rationale is that the ordering of these detection boxes directly determines the subsequent mask generation sequence, which in turn governs the final tying execution order and thus improves overall path efficiency.

3.2.2 Binary Mask Generation

For each detected bounding box, the background has already been removed, allowing direct mask generation for the intersection region. To balance processing efficiency and segmentation accuracy, an adaptive thresholding method is employed, with the threshold computed as shown in Equation (1):

$$T(x,y) = \frac{1}{\lfloor\Omega\rfloor} \sum_{(i,j)\epsilon\Omega} I(i,j) - C \tag{1}$$

Where: $T(x,y)$ denotes the grayscale threshold, Ω represents the local neighborhood, $\lfloor\Omega\rfloor$ is the number of pixels within the neighborhood, and C is a tuning constant.

The binarized results of the rebar intersection regions are shown in Figure 5. For each intersection, the generated mask, together with the corresponding image and depth data, is used as the input to the proposed Fuzzy-Oriented FoundationPose method.

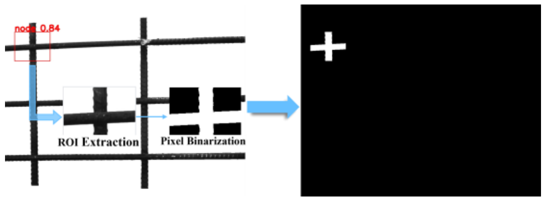

Figure 5. Mask Generation Pipeline for a Single Intersection

3.3 6DoF pose estimation

This section first introduces an improved FoundationPose-based pose estimation method, in which an observation-angle-guided optimization strategy and an active front-view rendering scheme are incorporated into the rotation initialization process. These enhancements accelerate network inference and suppress symmetry-induced equivalent pose outputs. Subsequently, a similarity-based orientation criterion is used to constrain the solution and ensure a unique pose output.

3.3.1 Fuzzy-Oriented FoundationPose

The original FoundationPose network performs pose refinement by rendering and matching a large set of candidate rotations around the initial pose estimate. Concretely, the rotations are initialized by uniformly sampling N_s viewpoints from an icosphere centered on the object with the camera facing the object center. Each viewpoint is further augmented with N_i discretized in-plane rotations, resulting $N_s \cdot N_i$ global pose initializations that are sent as input to the pose refiner. While this global search strategy improves robustness, it introduces many redundant pose hypotheses in rebar intersection scenarios, significantly increasing the computational cost of refinement. To address this limitation, the proposed method concentrates on a streamlined pose hypothesis generation stage designed for targeted search.

A camera position search strategy guided by the observation angle is first employed. When the depth camera captures the double-layer rebar grid, the relative pose between the camera coordinate system and the rebar intersection's coordinate system is shown in Figure 6, where the optical axis of the camera is directed toward the intersection.

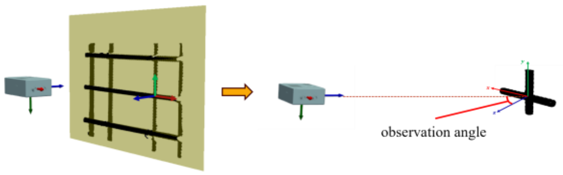

Figure 6. Location distribution diagram of depth camera, tying layer point clouds and fitted plane

The fitted plane of the rebar layer obtained from the point cloud segmentation algorithm, $Ax + By + Cz + D = 0$, can be approximately considered perpendicular to the z-axis of each rebar intersection's object coordinate frame. Accordingly, the normal vector of the plane facing the camera, (A, B, C), is regarded as the representation of the intersection's z-axis in the camera coordinate system. By considering only the direction from the camera to the object center and ignoring the distance magnitude, the camera optical axis can be regarded as passing through the center of the object frame, enabling the observation angle θ to be computed according to Equations (2) to Equation (3).

$$cos\theta = \frac{\vec{n} \cdot \vec{a}}{|\vec{n}| \times |\vec{a}|} = \frac{C}{\sqrt{A^2 + B^2 + C^2} \times 1} \qquad (2)$$

$$\theta = arccos\left(\frac{C}{\sqrt{A^2 + B^2 + C^2} \times 1}\right) \qquad (3)$$

Where $\vec{a}$ represents the optical axis vector expressed in the camera coordinate system, given by (0, 0, 1). It is worth noting that the angle θ must satisfy $0 < \theta < 180°$.

The icosphere, commonly employed as a rotation sampling tool in computer vision, provides a uniform coverage of the 3D orientation space, thereby enhancing the completeness of pose exploration [17]. However, in practical engineering applications, performing a global search incurs substantial computational overhead and fails to satisfy real-time requirements. To address this limitation, a rotation-initialization optimization strategy is proposed: leveraging coarse observational angle cues to constrain the search space and improve inference efficiency.

Specifically, candidate viewpoints are sampled from an icosphere centered on the target object, with each viewpoint facing the center. These viewpoints are further constrained to match the observational angle during real image acquisition. The corresponding constraint conditions are formulated in Equation (4) to Equation (5).

$$cos\theta' = \frac{\overrightarrow{n'} \cdot \overrightarrow{a'}}{|\overrightarrow{n'}| \times |\overrightarrow{a'}|} = \frac{-\varepsilon_z}{\sqrt{{\varepsilon_x}^2 + {\varepsilon_y}^2 + {\varepsilon_z}^2} \times 1} \qquad (4)$$

$$cos\theta' = cos\theta \qquad (5)$$

Where $\overrightarrow{n'} = -(\varepsilon_x, \varepsilon_y, \varepsilon_z)$ represents the camera optical vector, θ' denotes the observational angle associated with a candidate viewpoint. Similarly, θ' is required to satisfy $0 < \theta' < 180°$.

Another challenge arises from the rebar intersection, whose symmetry is considerably more complex. As illustrated in Figure 7, the intersection exhibits not only up–down symmetry but also two additional back-facing symmetric viewing configurations.

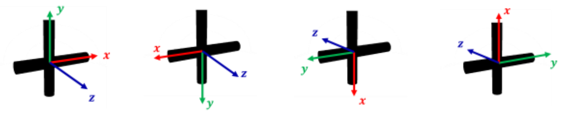

Figure 7. Diagram of rebar intersection symmetry

To suppress these symmetry-induced equivalent poses, the coarse pose sampling is further constrained on top of the icosphere-based optimization. When the camera position corresponds to quadrants I, II, VII or VIII of the object frame, the rotation angle is limited to the range $(-\beta, \beta)$ so that only upright rendering poses are retained. In contrast, for quadrants III, IV, V and VI, the constraint is set to the range $(180° - \beta, 180° + \beta)$, ensuring that only upright rendering poses are retained.

3.3.2 Orientation-Based Symmetry Resolution

Although the aforementioned strategy largely restricts the output of symmetric poses, the pose estimation network may still produce non-standard poses with a certain probability, as it encourages large updates of coarse poses during iterative optimization. To address this, a symmetry constraint based on orientation similarity is applied. A set of equivalent transformation matrices $S = \{S_i\}$ is generated, and all corresponding equivalent rotation matrices of the network predicted pose R_{est} are computed:

$$R_i = R_{est} S_i \qquad (6)$$

$$score(R_i) = \langle R_i[:,1], -y_{cam} \rangle + \langle R_i[:,0], x_{cam} \rangle \qquad (7)$$

The similarity scoring function, defined in Equation (7), evaluates the alignment between the object orientation and the camera coordinate directions. Specifically, $R_i[:,1]$ and $R_i[:,0]$ denote the unit direction vectors of the object y-axis and x-axis in the camera coordinate frame, while y_{cam} and x_{cam} represent the corresponding unit direction vectors of the camera coordinate system. The inner product computes the cosine of the angle between two vectors; values closer to two indicate better directional alignment. Therefore, the pose with the highest score is selected as the final output.

4 Experiments and Results

4.1 Experiments set up

4.1.1 Efficiency and Symmetry Experiments

An initial evaluation of the proposed pose estimation method is conducted on a single rebar intersection. This setup allows for strict control of experimental variables and provides an accurate assessment of the algorithm's inference efficiency under minimally disturbed conditions. The dataset was generated using ObjectDatasetTools [18]. As shown in Figure 8, thirteen ArUco markers were affixed to a custom 30 × 30 cm wooden board, with the rebar intersection placed at the center. The board is rigidly mounted on a motorized turntable, and an Intel RealSense D435i camera is used to record the scene for approximately 50 seconds during continuous rotation, resulting in 1310 annotated frames.

The generalization capability of the proposed Fuzzy-Oriented FoundationPose in suppressing symmetric ambiguities is further evaluated through pose estimation experiments on four highly symmetric geometric objects: a cube, a pyramid, a cylinder and a cone. The data collection procedure was kept consistent with that used for the rebar intersection experiments. For each object, video data were independently recorded and processed, resulting in approximately 570 annotated frames per dataset. This ensures comparability of experimental conditions across different object categories.

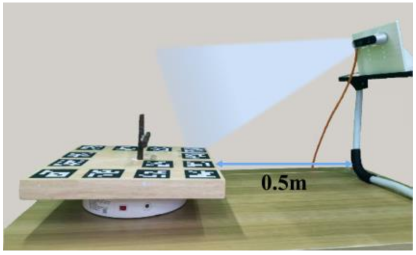

Figure 8. ObjectDatasetTools method for creating datasets and setting up scenarios

4.1.2 Real tying data experiment

Additional experiments were carried out on real rebar cage data to demonstrate the applicability of the proposed method in practical construction scenarios. In the experiments, an Intel RealSense D435i depth camera was rigidly mounted on the end effector of a JAKA C12 six-axis industrial robotic arm to acquire RGB-D data of the rebar cage within the target workspace. This configuration closely emulates practical rebar tying operations, where the collected data inherently include complex backgrounds, occlusions, and realistic lighting conditions, thereby imposing more stringent requirements on pose estimation algorithms. Testing on real rebar cage data enables a systematic assessment of the proposed method in terms of stability and practical applicability under engineering conditions.

4.1.3 Implementation Details

All experiments were conducted on hardware consisting of Intel i5-9400F CPU and an NVIDIA RTX 2060 SUPER GPU.

At the algorithmic level, because the icosphere vertices are discretely sampled, an angular tolerance of 30° is introduced to ensure that at least one vertex falls within the searchable range. The in-plane rotation angle β is set to 30° with a step size of 60°. The coarse pose initialization is further refined through three iterations of the refinement network.

4.2 Experimental Results

To strictly follow the baseline protocol for each setup, the following metrics are considered.

1. The average distance metric (ADD) [10] and ADD-S [5]
2. Frames per second (FPS)
3. Recall of ADD and ADD-S that is less than 0.1 of the object diameter (ADD-0.1d and ADD-S-0.1d) [19-20].

In the efficiency comparison, per-frame pose estimation results and inference time, as well as the total runtime, were recorded for both the proposed method and FoundationPose on the rebar intersection dataset. Our method completed the same task in only 565.396 seconds, whereas FoundationPose required 5964.85 seconds to process 1310 frames.

Table 1. Accuracy Comparison with FoundationPose Measured by ADD-0.1d and ADD-S-0.1d on Rebar Intersection Dataset

Method	ADD-0.1d	ADD-S-0.1d
FoundationPose	99.85	100
Ours	99.92	100

ADD-0.1d and ADD-S-0.1d were computed to evaluate the accuracy of the estimated poses, and the final results are summarized in Table 1. Overall, the results demonstrate that the proposed method improves inference speed while maintaining comparable pose estimation accuracy. A further analysis of erroneous estimates reveals that the primary cause stems from the limited quality of the Intel RealSense D435i data. In particular, under reflective surfaces or high-noise conditions, the network may capture incorrect or ambiguous features during prediction, leading to large errors in certain frames. In contrast, the proposed method employs an observational-angle-based optimization strategy during the rotation initialization stage to eliminate redundant poses, effectively mitigating the negative impact of low-quality data on the optimization process. Consequently, for frames with poorer data quality, the pose estimation errors are significantly smaller than those of FoundationPose, demonstrating the robustness of the proposed approach in complex or degraded data environments.

Table 2. Evaluation of Symmetry-Disambiguated Pose Estimation on Four Geometric Objects.

Object	Method	Metric				
		FPS	ADD	ADD-S	ADD-0.1d	ADD-S-0.1d
Cube	FoundationPose	0.136	0.081435	0.002695	6.36	100
	Ours	**0.69**	**0.004903**	**0.002723**	**100**	100
Cone	FoundationPose	0.07	0.026322	0.002329	21.05	100
	Ours	**0.76**	**0.015150**	**0.002332**	**76.8**	100
Cylinder	FoundationPose	0.10	0.050223	0.002173	7.94	100
	Ours	**0.59**	**0.017423**	**0.002068**	**63.2**	100
Pyramid	FoundationPose	0.10	0.027423	0.002068	28.07	100
	Ours	**0.62**	**0.003973**	**0.002061**	**100**	100

Due to manufacturing variations, individual rebar intersection exhibit limited geometric symmetry, so symmetry-related issues have minimal impact in single intersection experiments. However, for highly symmetric geometric objects, spatial symmetry can lead to multiple equivalent poses, significantly affecting pose estimation accuracy. The experimental results are summarized in Table 2: as the degree of symmetry increases, the prediction performance of FoundationPose deteriorates sharply. For example, a cube has 24 equivalent poses, and the ADD-0.1d metric for FoundationPose is only 6.36%, nearly matching the theoretical mean of 1/24. In contrast, when using the proposed Fuzzy-Oriented FoundationPose framework, the ADD-0.1d metric increases to 100% because the rendering process includes only vertical viewpoints and is further refined by the final orientation-similarity correction. It should be noted that for solids of revolution such as cones, absolute correction is not achievable due to axial symmetry; nevertheless, the ADD-0.1d metric still shows a substantial improvement over the baseline method.

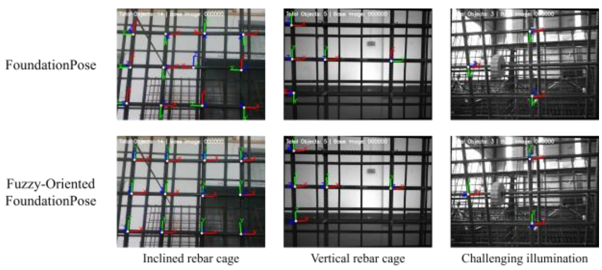

Figure 9. Comparison of Pose Estimation Results with FoundationPose on Rebar Cage Data

In the rebar cage dataset, we further conducted qualitative experiments. As illustrated in Figure 9, several representative scenarios are presented, including inclined, vertical and challenging illumination conditions. When applying FoundationPose directly to rebar cage scenes, the estimated poses of the rebar intersections exhibit noticeable random fluctuations due to their inherent geometric symmetry. Such instability may introduce a risk of mechanical interference during robotic tying operations. In contrast, the proposed Fuzzy-Oriented FoundationPose demonstrates substantially improved stability when handling rebar cage data. The method is able to consistently align the coordinate frames of all rebar intersections to a standard orientation, effectively eliminating symmetry-induced equivalent pose outputs. Moreover, the rotational search space required for rendering is reduced from the original 252 candidate poses to only 16, significantly decreasing the computational overhead of rendering and matching. The proposed approach leads to improved inference speed, providing a practical foundation for future online applications in robotic rebar tying.

5 Conclusion and Future Research

This paper presents a 6DoF pose estimation framework for multi-axis robotic rebar tying. Existing studies primarily focus on methods based solely on RGB or depth information, which limit pose estimation accuracy in complex construction environments. The framework integrates object detection for determining the tying sequence and generating pixel-level segmentation masks. More importantly, a novel observation-angle-guided algorithm, termed Fuzzy-Oriented FoundationPose, is introduced. Experimental results demonstrate that, for non-rotationally symmetric objects, the proposed method improves the ADD-0.1d metric to 100%, effectively resolving pose ambiguities caused by geometric symmetry. In real rebar cage scenarios, the proposed framework demonstrates improved pose stability and inference efficiency, thereby providing strong support for robotic rebar tying applications.

Despite the promising results, the proposed framework still relies on the assumption that the fitted rebar layer plane is approximately perpendicular to the object z-axis. Future work will focus on developing more flexible geometric formulations to improve robustness in complex rebar configurations.

Acknowledgment

This research was supported by the National Key Research and Development Program of China (Grant No.2023YFC3806800) and the National Natural Science Foundation of China (Grant No. 52278177).

References

[1] G. Biswas, A. Bhattacharya, and R. Bhattacharya. Occupational health status of construction workers: A review. *Int J Med Sci Public Health*, 2017.

[2] Tybot. TyBot: REBAR TYING ROBOTS. On-line: https://www.constructionrobots.com/tybot, Accessed: 25/12/2024.

[3] T. Takahashi, T. Sugimoto, H. Tanimoto, and S. Yoshizawa, "A Proposal for Downconverting A-to-D Converter Based on Even-Harmonic Mixer and ΔΣ-TDC," In *Proceedings of the IEEJ Trans.EIS*, pages 50–56, Singapore, Singapore, 2018.

[4] J. Jin. "Robotic binding of rebar based on active perception and planning," *Automation in Construction*, vol. 132, 2021.

[5] X. Yu, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *arXiv preprint*, arXiv:1711.00199, 2018.

[6] B. Wen, W. Yang, J. Kautz, and S. Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868-17879, Seattle WA, USA, 2024.

[7] D. G. Lowe. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, Kerkyra, Greece, 1999.

[8] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision – ECCV*, pages 404–417, Berlin, Heidelberg, 2006.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision*, pages 2564–2571, Barcelona, Spain, 2011.

[10] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, K. Konolige, and N. Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. *Asian conference on computer vision*. pages 548–562 Berlin, Heidelberg, 2012.

[11] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4556–4565, Long Beach, USA, 2019.

[12] C Wang, D Xu, Y Zhu, *et al*. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3338–3347, Long Beach, USA, 2019.

[13] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages11629–11638. Seattle, USA, 2020.

[14] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. *arXiv*: arXiv:2103.02242, 2021.

[15] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. pages 381–395, New York, USA, 1981.

[16] G Jocher. Ultralytics. Ultralytics. On-line:. https://github.com/ultralytics/ultralytics, Accessed: 01/10/2023.

[17] A. Yershova and S. M. LaValle. Deterministic sampling methods for spheres and SO(3). In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3974-3980, New Orleans, USA, 2004.

[18] F2Wang. ObjectDatasetTools. ObjectDatasetTools. On-line: https://github.com/F2Wang/ObjectDatasetTools, Accessed: 31/12/2020.

[19] X. He, J. Sun, Y. Wang, D. Huang, and X. Zhou. OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models. *Advances in Neural Information Processing Systems* 35 (2022): 35103-35115, 2022.

[20] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen. FS6D: Few-Shot 6D Pose Estimation of Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6804–6814, New Orleans, USA, 2022.