A Robust Visual Grasping Method for Robots in Cluttered and Stacked Scenes

Zhiqiang Gao , Mengqi Li , Huihui Bai , Jinze Li , Sifan Li , Jing Han * , Zhengkai Wang

Posted Date: 4 June 2026

doi: 10.20944/preprints202606.0320.v1

Keywords: visual grasping; pose estimation; cluttered stacking scenes; SAM-FoundationPose; iterative closed-loop

 doi:10.20944/preprints202606.0320.v1

Article

A Robust Visual Grasping Method for Robots in Cluttered and Stacked Scenes

Zhiqiang Gao [1], Mengqi Li [1],Huihui Bai [1], Jinze Li [1], Sifan Li [1], Jing Han [2,*] and Zhengkai Wang [2]

[1] Department of Automation, Taiyuan Institute of Technology,Taiyuan 030008, China

[2] College of Mechatronic Engineering, North University of China, Taiyuan 030051, China

* Correspondence: ajingcool@tom.com

Abstract

In complex backgrounds and under severe occlusions, the accuracy of vision-based robotic grasping pose estimation decreases significantly, further making objects difficult to manipulate and grasp. This paper proposes an iterative closed-loop optimization framework that deeply couples SAM with FoundationPose. The framework breaks through the open-loop logic bottleneck of "segmentation first, then estimation" found in traditional vision algorithms and constructs a mask correction mechanism based on rendered projection. By performing 3D rendering of the initially estimated 6D pose, a geometric prior mask of the object is generated and then fed back into SAM's prompt encoder, thereby guiding the model to achieve pixel-level refinement of the target's boundary in the next perception cycle. Meanwhile, to overcome the blind spots of a single metric, the framework designs a multi-dimensional confidence assessment module that integrates both the 2D image domain and the 3D geometric domain to comprehensively evaluate the reliability of the current pose. The SAM prior, the closed-loop iterative mechanism, and the multi-dimensional confidence assessment module work in synergy to form a complete optimization loop. In robustness experiments on pose estimation under cluttered and stacked scenes, the proposed method achieves an overall ADD-S recall rate of 91.7%, with the average translation and rotation errors reduced to as low as 3.5 mm and 2.1°. In 200 real-world robotic grasping verification trials, the overall grasping success rate reaches 96.5%. These experimental results fully validate the effectiveness and robustness of the proposed closed-loop optimization framework in unstructured environments.

Keywords: visual grasping; pose estimation; cluttered stacking scenes; SAM-FoundationPose; iterative closed-loop

1. Introduction

With the deepening evolution of intelligent manufacturing and collaborative robots, the operational capability of robotic arms in unstructured environments such as workpiece sorting and precision assembly has become a research focus in perception technology [1,2]. As the foundation for robots to perform physical interactions, the accuracy of vision-based 6D pose estimation of target objects directly determines the success or failure of grasping tasks [3,4]. Existing approaches to 6D pose estimation can be broadly categorized into three classes: traditional methods based on geometric optimization, regression methods based on deep learning, and methods based on rendering and matching [5,6]. Traditional methods, such as the Iterative Closest Point(ICP) algorithm and its variants, rely on local geometric gradients for point cloud registration and are highly susceptible to local optima in the absence of texture or under severe occlusions [7]. PnP-type methods, although computationally efficient, are highly sensitive to the quality of feature point matching and lack sufficient robustness [8]. Deep learning methods represented by PoseCNN and DenseFusion directly regress pose parameters or establish pixel-level dense correspondences through end-to-end networks, achieving significant breakthroughs in accuracy [9–11]. However, their generalization capability is heavily constrained by the coverage of training data. Wang et al. [12] attempted to expand the

 doi:10.20944/preprints202606.0320.v1

perceptual dimension through viewpoint classification networks and template matching; however, this approach relies on scene-specific pre-trained models and exhibits limited performance under varying illumination or dynamic environments. Liu et al. proposed an RGB-D keypoint cloud estimation scheme that excels in lightweight deployment and real-time performance, yet it remains fundamentally a serial open-loop control and lacks the capability for feedback correction of preceding errors [13].

A common weakness of the aforementioned methods lies in their unidirectional perception paradigm: once deviations occur in the front-end stages of object detection,segmentation, or feature matching due to occlusions, cluttered backgrounds, or illumination variations, such errors are irreversibly amplified and propagated to the final pose output. The entire process lacks self-inspection and correction mechanisms. Consequently, achieving vision perception with error-correction capability in highly variable and cluttered industrial settings remains a critical bottleneck yet to be overcome in the field of automation.

The emergence of visual foundation models has offered new perspectives for overcoming perception challenges. The Segment Anything Model (SAM), with its exceptional zero-shot generalization capability,enables pixel-level extraction of target masks from complex backgrounds, providing a powerful visual front end for filtering out background interference and addressing object recognition difficulties [14]. However, SAM inherently operates within the realm of 2D semantic perception and lacks the capability for pose estimation in 3D geometric space. Meanwhile, the FoundationPose model introduced at CVPR 2024 unifies the model-based and model-free technical routes by integrating neural implicit representations with prior knowledge acquired from large-scale synthetic data training, demonstrating strong cross-object transfer potential. Nevertheless, when handling highly dynamic backgrounds or occluded scenes, its single feed-forward inference results remain susceptible to the quality of initial perception, leading to localization drift [15].

Based on the above analysis, this paper proposes an iterative closed-loop optimization framework that deeply couples SAM with FoundationPose. This architecture abandons the traditional unidirectional perception paradigm and constructs a closed-loop pipeline based on rendering feedback. It leverages semantic extraction capability of SAM to establish a clean input environment for FoundationPose, and then renders the pose estimation results as geometric priors to drive SAM in reverse for refined mask correction. Through this cross-dimensional bidirectional feedback mechanism,the system can progressively converge the error over successive iterations, effectively compensating for the performance shortcomings of a single foundation model in extreme environments. Experiments demonstrate that this method not only effectively improves pose estimation accuracy in complex scenes but also provides more robust algorithmic support for autonomous robotic grasping.

2. Materials and Methods

In complex and heavily occluded scenes, traditional unidirectional feed-forward 6D pose estimation algorithms are susceptible to background interference or the absence of local features, resulting in pose prediction deviations.To address this issue, this paper proposes an iterative closed-loop fusion pose estimation framework based on visual foundation models, namely SAM and FoundationPose. This section first briefly introduces the principles of the adopted foundation models, followed by a detailed exposition of the proposed core system architecture and the design of each module.

2.1. Principles and Applications of SAM

SAM is a computer vision foundation model developed by Meta AI Research. Its core advantage lies in its powerful zero-shot generalization capability,which enables high-quality zero-shot segmentation of arbitrary objects in unseen scenes through prompts such as points, boxes, and text. SAM adopts a three-component collaborative architecture consisting of an image encoder,a prompt encoder,and a mask decoder,which work in concert to complete the entire process from image input

 doi:10.20944/preprints202606.0320.v1

to segmentation mask output,as illustrated in Figure 1. Specifically,the image encoder employs a Vision Transformer (ViT) architecture [16] to extract deep image features; the prompt encoder converts user-provided prompts into feature embeddings. And the mask decoder, based on the bidirectional cross-attention mechanism of Transformers [17], not only fuses the above features to generate high-quality object masks,but also produces an IoU (Intersection over Union) score for each mask to evaluate mask quality.

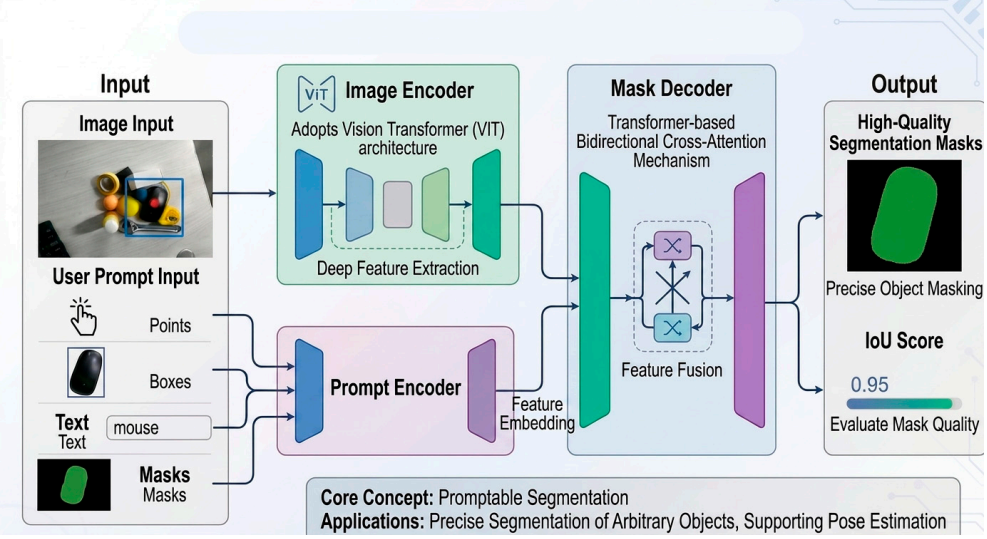


Figure 1. Workflow diagram of SAM.

In this study, SAM primarily serves as a high-precision mask generator at the front end. For an input RGB image with a complex background $I \in R^{H \times W \times 3}$, SAM incorporates specific prompt information PP (such as the guided mask fed back from the previous iteration) and outputs the foreground mask matrix MM of the target object:

$$M = f_{SAM}(I, P) \tag{1}$$

Compared with traditional segmentation models that struggle to accurately extract object contours in complex scenes, SAM demonstrates strong robustness in handling challenging scenarios such as stacked objects and low-contrast regions. It offers high practical value in real-world applications including autonomous driving and complex robotic environment perception,and also provides a solid technical foundation for the pose estimation research presented in this paper.

2.2. Principles and Applications of FoundationPose

FoundationPose is a unified high-precision 6D pose estimation and tracking foundation model that requires no training on specific objects and supports zero-shot generalization. Taking an RGB-D image and the 3D model of the target object as input, the model fuses the rendering and comparison mechanism of neural implicit representations [18,19] to output the 6D pose transformation matrix T of the object in the camera coordinate system.

The pose matrix T belongs to the three-dimensional special Euclidean group $SE(3)$ and consists of a rotation matrix $R \in SO(3)$ and a translation vector $t \in R^3$:

$$T = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in SE(3) \tag{2}$$

Compared with traditional pose estimation methods such as PnP, ICP, and PoseCNN, FoundationPose demonstrates high robustness and zero-shot generalization capability in standard scenes,providing a powerful pose inference foundation for autonomous robotic grasping [20].

 doi:10.20944/preprints202606.0320.v1

However,FoundationPose essentially remains a single feed-forward architecture, where its pose output is heavily dependent on the quality of the input observations.When confronted with severe occlusions or complex background interference,local appearance ambiguities caused by sensor noise or occlusions can directly affect the final inference results,and the error lacks any built-in self-inspection and correction mechanism.This non-negligible open-loop limitation constitutes the core motivation of this study [21]. To address this,we introduce a closed-loop feedback mechanism that renders the inferred pose results back into geometric priors and re-injects them into the front-end perception module,thereby endowing the system with self-inspection and correction capabilities and breaking through the open-loop constraints.

2.3. SAM-FoundationPose Closed-Loop Fusion Framework

Traditional deep learning-based 6D pose estimation methods [22–24] predominantly adopt an open-loop control strategy of "segmentation first, then estimation". This approach is highly dependent on the quality of front-end segmentation.Once segmentation fails due to complex backgrounds or severe occlusions,errors are irreversibly propagated downstream and amplified. To overcome this limitation,this paper proposes an iterative closed-loop fusion framework based on SAM and FoundationPose. The framework converts pose estimation results into spatial constraint priors, which in turn guide the vision model to perform mask refinement, thereby enabling mutual correction and co-evolution between segmentation and pose estimation.

2.3.1. Overall Framework Design

The overall workflow of the proposed framework is illustrated in Figure 2. The system takes an RGB-D image with a complex background and the 3D model of the target object as inputs. An initial foreground mask is generated by the SAM segmentation module and then fed into the FoundationPose pose estimation module to produce an initial 6D pose. Subsequently,the pose data are passed into the multi-dimensional confidence assessment module,which evaluates the reliability of the current pose by computing the re-projection error,mask matching degree,and geometric error. If the confidence score meets the threshold, the final pose is directly output and used to guide the robotic arm in visual grasping. If the confidence score falls below the threshold,the system enters the iterative optimization control stage,where the current pose is rendered into a guided mask and fed back to the pose estimation module for optimization and correction,repeating the process until the confidence requirement is satisfied.

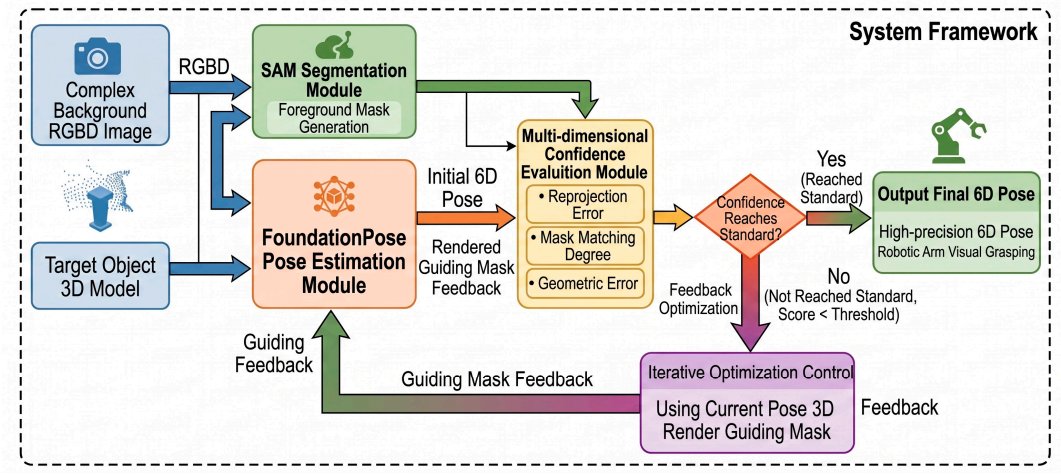


Figure 2. Iterative closed-loop fusion framework based on SAM and FoundationPose.

 doi:10.20944/preprints202606.0320.v1

2.3.2. Dynamic Prompt-Based SAM Segmentation Module

In this closed-loop framework, the SAM segmentation module serves as a front-end high-fidelity region extractor. Unlike traditional static semantic segmentation networks, this module fully exploits the promptable interactive nature of the SAM architecture and incorporates two distinct working mechanisms—initialization with heuristic prompts and guidance via iteratively rendered masks—to address the challenge of severe occlusions in complex scenes.

When the system first encounters a complex RGB-D image $\boldsymbol{I}$, to activate SAM's zero-shot segmentation capability,this module introduces heuristic prompts P_{init} by employing the lightweight 2D object detection algorithm YOLOv5n to generate a coarse 2D bounding box containing the target object as the prompt input [25].Considering that a pure 2D object detector may suffer from missed detections under severe occlusions or extreme background confusion,which would result in the loss of initial prompts, this module incorporates a depth-prior-based fallback mechanism to ensure the continuity of the front-end perception pipeline. When YOLOv5n fails to produce a valid bounding box,the system automatically switches to a depth-guided mode:it performs fast Euclidean clustering on the cluttered point cloud within the workspace using the input depth map, extracts the 3D centroids of potential protruding objects, and reprojects them onto the 2D pixel plane to serve as point prompts for SAM.This dual heuristic strategy that combines RGB-D heterogeneous information effectively compensates for the robustness shortcomings of a single 2D detection front end under extreme conditions,thereby ensuring that the closed-loop iterative framework can be reliably triggered.

Let ε_{img} denote the image encoder of SAM, ε_{prompt} the prompt encoder, and D_{mask} the mask decoder. The generation process of the initial mask M_0 can be mathematically formalized as:

$$F_I = \varepsilon_{\text{img}}(I) \tag{3}$$

$$F_{P_0} = \varepsilon_{prompt}(P_{init}) \tag{4}$$

$$M_0 = D_{mask}(F_I, F_{P_0}) \tag{5}$$

The mask generated at this stage can effectively remove the majority of background point clouds and irrelevant textures,thereby tightly constraining the search space of the subsequent FoundationPose estimation module around the target object. In practical grasping scenarios, however, the initial mask generated solely by a single-shot heuristic prompt often falls short of ideal segmentation quality. When the target object is severely occluded or its color closely resembles the background, SAM's segmentation results may suffer from over-segmentation or under-segmentation, and the single feed-forward mechanism will irreversibly propagate such errors to the pose estimation module, leading to the accumulation of pose errors [26–28].

To overcome this limitation, this paper deeply embeds this module into the closed-loop feedback pipeline. When the system enters the k-th iteration (k≥1), this module receives the guided mask from the iterative optimization control module $M_{guided}^{(k-1)}$.This guided mask is rendered from the imperfect pose T_{k-1} estimated in the (k−1)-th iteration and the 3D model,and it provides the ideal geometric topology of the target under the current viewpoint. Injecting $M_{guided}^{(k-1)}$ as a spatial prior prompt into the system can force SAM to re-correct the segmentation boundaries in edge-ambiguous regions. The mask generation equation for the k-th iteration is updated as:

$$F_{P_k} = \varepsilon_{prompt}(P_{init}, M_{guided}^{(k-1)}) \tag{6}$$

 doi:10.20944/preprints202606.0320.v1

$$M_k = D_{mask}(F_1, F_{P_k}) \tag{7}$$

By implementing the closed-loop interaction mechanism described above,the SAM segmentation module evolves from an isolated preprocessing step into a dynamically self-optimizing perception module.After the refined mask M_k it produces is fed into the FoundationPose pose estimation module,it can substantially reduce the interference of outliers and enhance the robustness of feature matching,thereby laying a foundation for achieving high-precision robotic grasping.

2.3.3. FoundationPose Pose Estimation Module

After generating the high-precision foreground mask of the target object,the system employs FoundationPose as the core 6D pose inference engine. Traditional single networks often require lengthy fine-tuning for specific objects, whereas FoundationPose, as a visual foundation model,leverages its powerful large-scale generalization prior to directly perform zero-shot pose estimation for novel objects.Within this closed-loop framework,the module first extracts local features through mask-guided region-of-interest cropping, and then generates and selects the optimal 6D pose hypotheses via the rendering and comparison mechanism [29,30].

In the k-th iteration, this module receives the global observation data from the environment (i.e., the RGB image ***I*** and the depth map ***D***), the 3D model M_{3D} of the target object, and the current frame foreground mask M_k output by the SAM module. To eliminate the interference of complex backgrounds on pose estimation, the module first performs a bitwise AND operation between M_k and the global observation data to extract local observation features I_{ROI} and D_{ROI} that are strictly constrained within the target region:

$$I_{ROI} = I \odot \mathrm{M_k} \tag{8}$$

$$D_{ROI} = D \odot \mathrm{M_k} \tag{9}$$

Through this mask-guided feature space constraint,the feature extraction of the FoundationPose module is highly concentrated on the effective pixels of the target object. This significantly reduces the probability of false feature matches caused by external occluding objects or similar background textures, thereby providing a relatively clean input source for subsequent pose estimation.

FoundationPose internally adopts a network architecture that combines neural implicit representations with rendering-based comparison. Let F_{pose} denote the core inference network of FoundationPose. By performing deep feature extraction on the input local observation features (I_{ROI}, D_{ROI}) and the 3D model M_{3D}, the network generates a set of initial 6D pose hypotheses within the vast pose search space.Through the rendering and comparison mechanism, the feature similarity between the rendered views of the 3D model under each hypothesized pose and the actual 2D observations is evaluated at the feature level, and the highest-scoring pose is output as the estimated result T_{k} for the current round. This process can be mathematically abstracted as:

$$T_{\mathrm{k}} = F_{pose}(I_{ROI}, D_{ROI}, M_{3D}) \tag{10}$$

Where T_k represents the 6D rigid-body transformation matrix of the target object in the camera coordinate system, belonging to the three-dimensional special Euclidean group SE(3). It consists of a 3×3 orthogonal matrix $R_k \in SO(3)$ representing rotation and a 3×1 vector $t_k \in R^3$ representing translation:

 doi:10.20944/preprints202606.0320.v1

$$T_k = \begin{bmatrix} R_k & t_k \\ 0 & 1 \end{bmatrix} \tag{11}$$

In traditional open-loop applications, the output of FoundationPose is taken directly as the final result. However, this output is constrained by the inherent limitation of single feed-forward inference: when the initial mask M_0 contains noise, the F_{pose} output by T_0 is highly prone to falling into a local optimum. Within the closed-loop framework constructed in this study, as the number of iterations k increases, the fed-back mask M_k increasingly approaches the true physical contours in terms of boundary delineation.The FoundationPose module performs re-estimation using the progressively purified I_{ROI} and D_{ROI}, enabling the output pose T_k to gradually converge toward the global optimum in both translation and rotation dimensions. The computed current pose T_k is then fed into the multi-dimensional confidence assessment module to determine whether to initiate the next round of mask refinement and pose iteration.

2.3.4. Multi-Dimensional Confidence Assessment Module

In the transition from a feed-forward network to a closed-loop iterative system, accurately evaluating the reliability of the current pose and deciding whether to terminate the iteration lies at the core of the entire framework.Traditional single evaluation metrics often suffer from inherent limitations.For example, relying solely on 2D re-projection error can easily fall into local optima in textureless regions,while relying solely on 3D depth error struggles to handle the pose ambiguity of symmetric objects.To overcome these limitations, this framework designs a multi-dimensional confidence assessment module that integrates both the 2D image domain and the 3D geometric domain. The module comprises three mutually orthogonal and complementary sub-metrics: re-projection error, mask matching degree, and geometric error [31].

The re-projection error is used to measure the alignment accuracy of the pose on the 2D pixel plane,primarily constraining the translation of the target object along the X- and Y-axes as well as the pitch or yaw angles.Let the set of sampled points on the surface of the target object's 3D model be denoted as $P_{3D} = \{p_i\}_{i=1}^{N} \in R^3$ .In the k-th iteration,given the known camera intrinsic matrix K,the system projects the 3D point cloud onto the 2D pixel plane according to the current predicted pose $T_k = [R_k \mid t_k]$ .Let $\pi(\cdot)$ be the perspective projection function from homogeneous coordinates to pixel coordinates. Then the average re-projection error u_i^{proj} between the projected point u_i^{obs} and the matched actual observed feature point $E_{reproj}^{(k)}$ in the image is defined as:

$$E_{reproj}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left\| \pi(K(R_k p_i + t_k)) - u_i^{obs} \right\|_2 \tag{12}$$

Where $||\cdot||_2$ denotes the L2 norm. A smaller $E_{reproj}^{(k)}$ indicates a higher degree of alignment of the 2D visual features.

To constrain the pose from a global topological perspective, this module introduces the mask matching degree. It primarily evaluates the plausibility of the predicted pose in terms of the object's overall contour and boundary scale. Using the current predicted pose T_k and the 3D model, the system renders a predicted foreground mask $M_{render}^{(k)}$ of the target object from the camera viewpoint. The Intersection over Union (IoU) between this rendered mask and the observed mask M_k output by the current round of the SAM module is then computed as:

 doi:10.20944/preprints202606.0320.v1

$$S_{iou}^{(k)} = \frac{\left| M_{render}^{(k)} \cap M_k \right|}{\left| M_{render}^{(k)} \cup M_k \right|} \tag{13}$$

The value of $S_{iou}^{(k)}$ ranges from 0 to 1. A higher score indicates a greater consistency between the object contour under the predicted pose and the actual observed contour, effectively preventing the algorithm from producing scale divergence under severe occlusions.

Although the above two 2D-domain metrics can effectively constrain the planar pose, they lack sensitivity to depth translation along the camera's optical axis (Z-axis). Therefore, the module introduces a depth-map-based geometric error to perform verification directly in the 3D physical space.Let $D_{render}^{(k)}$ be the depth map rendered using T_k, and D_{obs} be the observed depth map actually captured by the depth camera. To exclude background interference, the error computation is performed only within the valid overlap region $\Omega = M_{render}^{(k)} \cap M_k$. The geometric error $E_{geo}^{(k)}$ is quantified using the root mean square error:

$$E_{geo}^{(k)} = \sqrt{\frac{1}{|\Omega|} \sum_{x \in \Omega} (D_{render}^{(k)}(x) - D_{obs}(x))^2} \tag{14}$$

where $\boldsymbol{x}$ represents the pixel coordinates,and $|\boldsymbol{\Omega}|$ denotes the total number of pixels within the valid region.The geometric error ensures a tight alignment between the predicted pose and the true 3D surface of the object.

To establish a unified convergence criterion, the system fuses the above three metrics, which have different dimensions, into a single scalar score. Since the re-projection error and the geometric error are the smaller the better, while the mask matching degree is the larger the better, the system first applies a bounded normalization to the two error metrics using a nonlinear negative exponential function $N(x) = \exp(-\lambda x)$ (where λ is a tuning coefficient). The final comprehensive confidence score S_k for the k-th iteration is computed as follows:

$$S_k = \omega_1 \exp(-\lambda_1 E_{reproj}^{(k)}) + \omega_2 S_{iou}^{(k)} + \omega_3 \exp(-\lambda_2 E_{geo}^{(k)}) \tag{15}$$

Where ω_1, ω_2, and ω_3 are the weight coefficients of the respective metrics, satisfying $\omega_1 + \omega_2 + \omega_3 = 1$. The comprehensive confidence score $S_k \in [0,1]$ can comprehensively and objectively reflect the accuracy of the current 6D pose. This score is then passed to the iterative optimization control module, serving as the sole trigger condition for the system's closed-loop operation.

2.3.5. Iterative Optimization Control

The iterative optimization control module serves as the core hub for controlling the closed-loop feedback pipeline of the system.In the feed-forward networks of traditional visual pose estimation, errors from front-end perception and segmentation are irreversibly propagated downstream and amplified. To break through this unidirectional data flow barrier, this module acts as a logic gate that governs the convergence state of the system based on the comprehensive confidence score output by the multi-dimensional confidence assessment module.When the comprehensive confidence score falls below the preset qualification threshold, it indicates that the currently inferred pose suffers from local optimum collapse caused by extreme occlusions or severe background confusion, and the system immediately activates the closed-loop feedback optimization loop. Once the iterative optimization stage is triggered,the module first extracts the imperfect 6D pose matrix output by FoundationPose in the current round and deeply couples it with the 3D model of the target object.Using a graphics rendering engine, the system reprojects the current 3D pose hypothesis back

 doi:10.20944/preprints202606.0320.v1

onto the 2D camera observation plane,thereby generating a guided mask that possesses the ideal geometric topology of the target.The essence of this rendering feedback operation is the dimensionality reduction of error correction information from the 3D spatial domain into a 2D geometric prior. Subsequently,this guided mask is injected in reverse into the prompt encoder of the front-end SAM segmentation module as a dynamic spatial prompt.

Under this cross-dimensional rendering guidance, SAM is forced to re-constrain its feature matching scope in edge-ambiguous and highly interfered regions,thereby effectively eliminating the interference of erroneous background textures and outputting a pixel-level refined high-fidelity foreground mask.Subsequently, the FoundationPose module receives the purified effective region features and performs a new round of pose estimation, driving the output pose toward the global optimum in both translation and rotation dimensions.The system undergoes successive linear iterations within the bidirectional closed-loop pipeline based on rendering feedback until the multi-dimensional comprehensive confidence score of the output pose meets the threshold criterion. At this point,the iterative control mechanism terminates and outputs the final high-precision 6D pose.

The detailed execution procedure of the entire SAM-FoundationPose iterative optimization control is presented in Table 1.

Table 1. SAM-FoundationPose Iterative Closed-Loop Optimization.

Input: *RGB-D image I, 3D model* M_{cad}, *camera intrinsics K*
Output: *Optimal 6D pose T**
Parameters: $\tau \leftarrow 85$ *(confidence threshold),* $k_{max} \leftarrow 10$ *(maximum iterations)*

1: $k \leftarrow 0$
2: *// Initialization phase*
3: $bbox \leftarrow$ YOLOv5n(I) ▷ *Heuristic bounding box prompt*
4: $M_0 \leftarrow$ SAM$(I, bbox)$ ▷ *Initial foreground mask*
5: $T_0 \leftarrow$ FoundationPose(I, M_0, M_{3d}) ▷ *Initial 6D pose estimation*
6: $S_0 \leftarrow$ ComputeScore(T_0, M_0, I, M_{3d}, K) ▷ *Multi-dimensional confidence score*
7: *// Iterative refinement phase*
8: ***while*** $S_k < \tau$ *and* $k < k_{max}$ ***do***
9: $k \leftarrow k + 1$
10: $M_{guide} \leftarrow$ RenderMask$(T_k{-}1, M_{3d}, K)$ ▷ *Render geometric prior mask*
11: $M_k \leftarrow$ SAM(I, M_{guide}) ▷ *Mask refinement with spatial prior*
12: $T_k \leftarrow$ FoundationPose(I, M_k, M_{3d}) ▷ *Pose re-estimation with refined mask*
13: $S_k \leftarrow$ ComputeScore(T_k, M_k, I, M_{3d}, K) ▷ *Update confidence score*
14: ***end while***
15: ***return*** $T^* \leftarrow T_k$

3. Experiments and Analysis

To thoroughly validate the accuracy and robustness of the 6D pose estimation method under the SAM-FoundationPose iterative closed-loop fusion framework,this section takes ten common objects—including a computer mouse, tape measure, adhesive tape, and earphones—as experimental targets. Using the constructed robotic arm visual grasping platform, a series of experiments are conducted successively:object pose estimation in cluttered but non-stacked scenes, object pose estimation in cluttered and stacked scenes, comparative experiments with different estimation methods,core module ablation experiments, and robotic arm visual grasping verification experiments in real-world scenarios.Through the above series of experiments, the self-correction capability of the proposed closed-loop framework when facing extreme visual interference and severe occlusions is comprehensively validated, effectively demonstrating the decisive role of introducing the iterative

 doi:10.20944/preprints202606.0320.v1

rendered feedback mechanism in improving 6D pose estimation accuracy and ensuring reliable robotic grasping in unstructured environments.

3.1. Platform Setup and Experimental Preparation

The experimental hardware system consists of three main components:a vision sensor, an edge computing core,and a collaborative robotic arm. As shown in Figure 3, an Intel RealSense D435i depth camera is selected as the vision sensor,providing color features for the SAM segmentation module and supplying 3D spatial geometric information for the FoundationPose and multi-dimensional confidence assessment modules.An NVIDIA Jetson NX is adopted as the computing core to handle visual inference tasks,and a Realman GEN72 series ultra-lightweight collaborative robotic arm serves as the grasping verification platform to execute physical manipulation tasks after pose output.

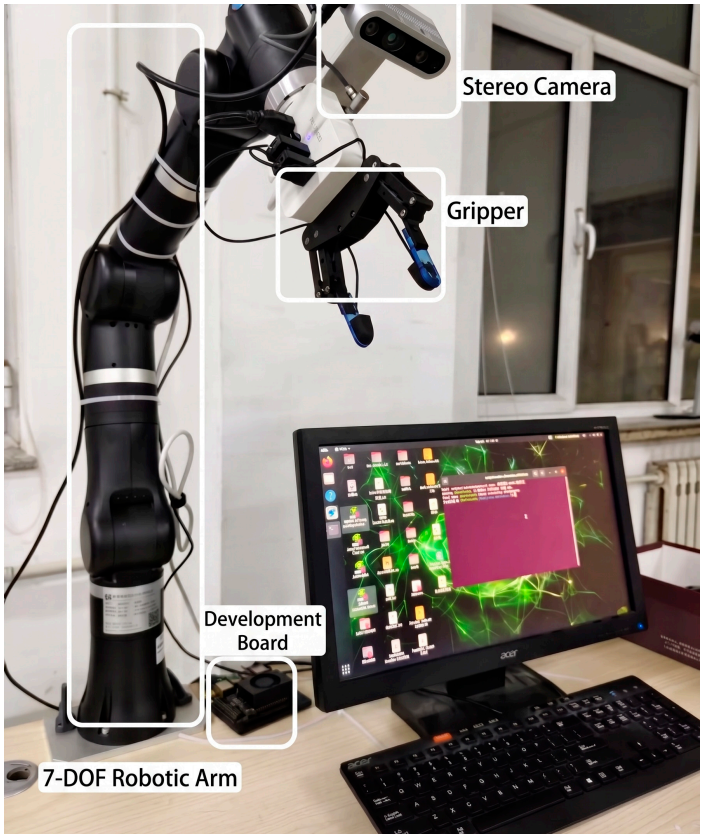


Figure 3. Experimental hardware system composition.

In terms of experimental targets,ten representative everyday objects were selected, including a computer mouse,tape measure,adhesive tape,earphones,a ping-pong ball,etc. The visual images and 3D geometric models of these objects are shown in Figure 4. The primary rationale for selecting these objects lies in their strong representativeness in terms of geometric shape complexity(e.g., cylinders, flat objects, irregular geometries), surface texture characteristics(e.g., smooth and reflective, richly textured), and size range(from a small ping-pong ball to a large detergent bottle),enabling effective evaluation of the accuracy and robustness of the proposed framework for object pose estimation in complex scenes.

 doi:10.20944/preprints202606.0320.v1

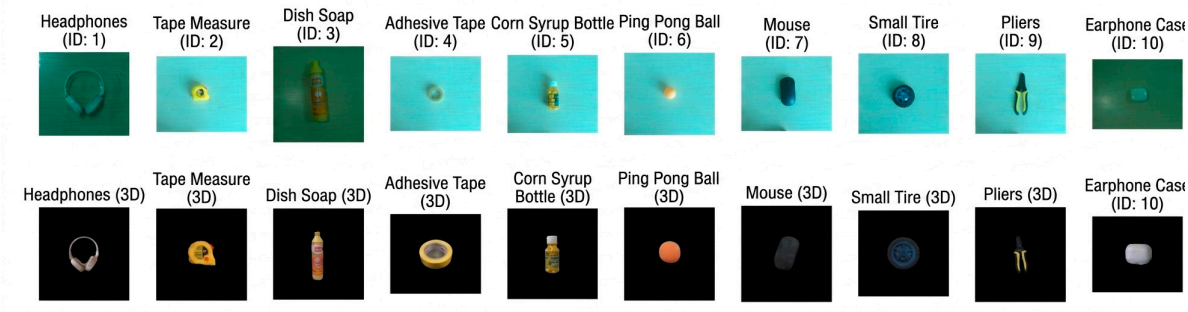


Figure 4. Comparison of real images (top) and 3D models (bottom) of the ten objects.

In terms of the underlying runtime environment,the edge computing core is equipped with the JetPack 5.1.2 operating system and configured with CUDA 11.4 and cuDNN computational acceleration libraries,fully harnessing the GPU computing power of the Jetson platform.The system's algorithm development is primarily based on Python 3.8, and both the model's forward inference and the closed-loop iterative optimization computations are natively deployed within the PyTorch 1.13.0 deep learning framework. To balance perception accuracy and operational efficiency under constrained edge computing resources, the system adopts the lightweight segmentation foundation model MobileSAM at the front end and utilizes PyTorch's native automatic mixed precision capability to accelerate the tensor operations of FoundationPose.

3.2. Object Pose Estimation in Cluttered but Non-Stacked Scenes

In practical robotic grasping applications,target objects are often subject to interference from surrounding clutter and confusion caused by similar background textures. Such environments can severely disrupt the feature extraction network's ability to correctly identify object boundaries and surface characteristics. To validate the adaptability of the proposed algorithm in non-ideal environments,this section conducts a robustness evaluation on object pose recognition under complex background interference. The experiment aims to verify the perception accuracy of the SAM-FoundationPose framework under conditions with no external physical occlusion but with severe background color and high-frequency texture confusion. The experimental scene is set up as a desktop with complex patterns and reflective interference. The ten selected categories of target objects are randomly placed within the camera's field of view. The core evaluation metrics employed are the ADD(-S)average recall rate(with a threshold of 2 cm), the average translation error, and the average rotation error.

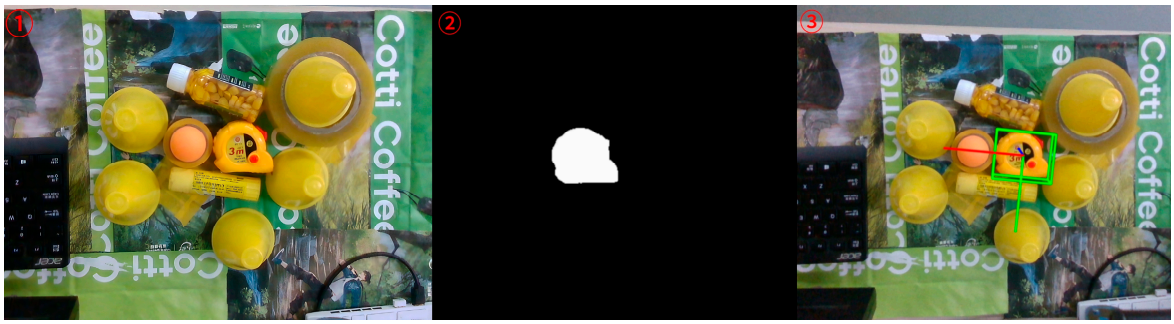
3.2.1. Robustness Analysis

Under visual inputs with strong background texture interference,traditional feed-forward segmentation algorithms based on edges or color gradients are highly prone to misidentifying background patterns as object contours, thereby causing the collapse of subsequent pose estimation. In the proposed framework, benefiting from the exceptional zero-shot semantic generalization capability of the SAM front end,the system can still effectively resist 2D background noise and achieve accurate separation of foreground regions.

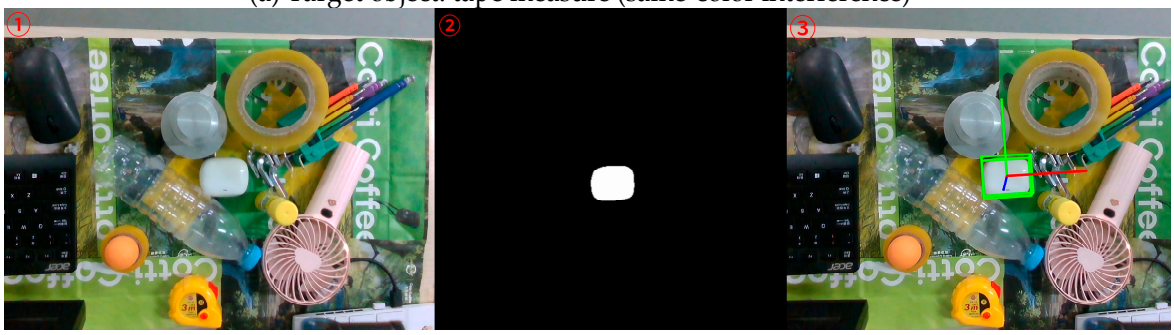
To intuitively demonstrate the system's perceptual robustness under 2D visual interference, Figure 5 presents the intermediate processing and pose rendering results for individual target objects in three typical challenging scenarios. The three scenarios are as follows. Figure 5(a) shows the same-color and background texture confusion scenario, with the tape measure as the target, the test environment contains background color blocks highly similar to the target color, with multiple same-color distractors scattered around. Figure 5(b) shows the high-reflectivity and transparent material interference scenario:focusing on the earphone case with a smooth surface, under the combined refraction of multiple transparent and reflective objects in the surroundings, the specular reflections

 doi:10.20944/preprints202606.0320.v1

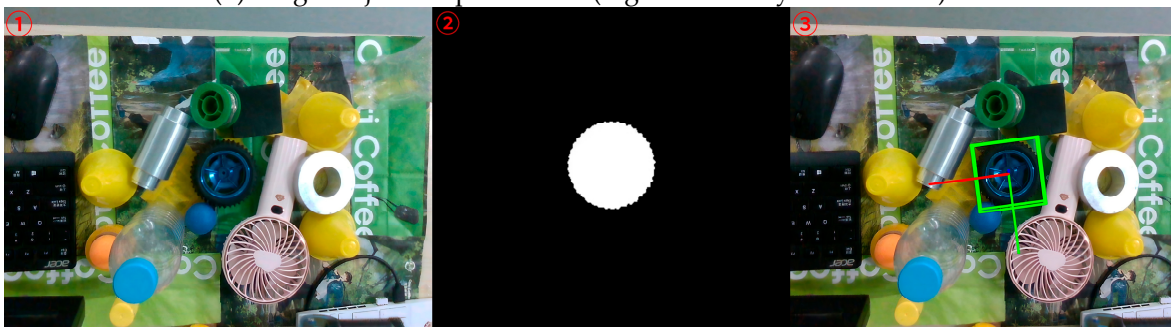
and background color transmission on the target surface pose severe challenges to mask boundary extraction. Figure 5(c) shows the strong feature texture and clutter interference scenario:centering on a small tire with strong jagged textures, this scenario verifies the system's ability to eliminate high-frequency background noise and accurately capture the effective geometric contours of the target when surrounded by scattered objects of diverse shapes.



(a) Target object: tape measure (same-color interference)



(b) Target object: earphone case (high-reflectivity interference)



(c) Target object: small tire (heterogeneous clutter interference)

Figure 5. Qualitative results of 6D pose estimation in three typical challenging scenarios.

In Figure 5, ①denotes the RGB input image from the camera, ②denotes the binary mask after segmentation by the SAM module, and③denotes the final rendered 6D pose output. It can be observed from Figure 5 that the system still delivers relatively high-quality perception results in the three typical challenging scenarios described above. In scenario(a), despite the color overlap between the tape measure body and the background pattern as well as interference from multiple same-color objects,the final mask generated by the system still accurately segments the target contour without any background adhesion, and the pose rendering box tightly fits the target center. In scenario(b), faced with high-reflectivity interference from the earphone case surface and surrounding objects,traditional algorithms are prone to breakage in highly reflective and bright regions, whereas the mask output by the proposed framework maintains high integrity and smoothness,demonstrating that this method effectively overcomes the 2D feature ambiguity caused by background color transmission. In scenario(c), targeting the small tire with strong jagged edges,under dense occlusion by surrounding clutter,the system still successfully locks onto the target;its mask contour faithfully restores the geometric jagged features of the tire edges, and the pose rendering box output by the system tightly fits the physical object,which verifies that when

 doi:10.20944/preprints202606.0320.v1

confronting high-frequency texture interference,the framework can effectively guide the model to traverse local minima through rendering feedback, ensuring convergence in the three-dimensional spatial domain.

3.2.2. Positioning Accuracy Analysis

To objectively quantify the system's anti-interference capability under complex backgrounds, the experiment collected statistics on the multi-dimensional pose errors of the ten categories of test objects in a desktop environment with strong interference. The quantitative evaluation results are presented in Table 2.

Table 2. Quantitative results of pose estimation in cluttered scenes.

Target Object (Category)	Object Type Characteristics	ADD(-S) Recall Rate (%)	Average Translation Error (mm)	Average Rotation Error (°)
Headphones	Heterogeneous shape / background texture	98.2	3.2	1.9
Tape measure	Same-color background confusion	97.5	4.1	2.3
Detergent bottle	Cylindrical symmetry / high-frequency reflections	96.8	3.8	2.1
Adhesive tape	Low-contrast edges	94.1	5.2	3.4
Corn bottle	Complex internal texture	98.0	2.9	1.7
Ping-pong ball	Textureless solid-color sphere	99.2	1.8	1.2
Computer mouse	Weak texture / dark color	98.5	2.6	1.8
Small tire	Strong geometric jagged texture	97.4	3.5	2.0
Pliers	Slender irregular structure	96.1	4.4	2.6
Earphone case	Multi-source lighting / specular reflections	95.3	4.8	3.1
Mean	-	97.1	3.63	2.21

According to the quantitative statistical results in Table 2, the SAM-FoundationPose closed-loop framework demonstrates high perceptual accuracy and 3D spatial alignment capability under complex backgrounds.In comprehensive tests on ten target objects with different physical characteristics, the overall average ADD(-S) recall rate of the system reaches 97.1%,while the average translation error is controlled at a low level of 3.63 mm and the average rotation error is as low as 2.21°. When encountering samples with optical and color ambiguities, the proposed algorithm still maintains strong robustness. For instance,for the adhesive tape that is transparent and has extremely low edge contrast, the earphone case subject to specular reflections,and the tape measure deeply

 doi:10.20944/preprints202606.0320.v1

entangled in same-color background confusion, their translation errors are all effectively constrained to within 5.2 mm, and the recall rates remain consistently above 94%.

This result fully demonstrates that the self-optimizing closed-loop logic based on rendering feedback constructed in this paper can effectively resist 2D visual deceptions (such as high-frequency textures, similar colors,and multi-source reflections) under unstructured working conditions. Through continuous correction via rendering priors,the system achieves effective convergence in 3D physical space pose estimation,thereby establishing a reliable perception foundation for the subsequent more challenging grasping tasks involving severe occlusions and cluttered stacking scenarios.

3.3. Object Pose Estimation in Cluttered and Stacked Scenes

Having established the baseline performance, this section focuses on evaluating the robustness of pose estimation under extreme conditions such as severe occlusions,cluttered stacking, and complex background interference.These experiments correspond to the highly challenging robotic grasping environments encountered in real-world scenarios and also validate the core value of the proposed "iterative closed-loop fusion mechanism."

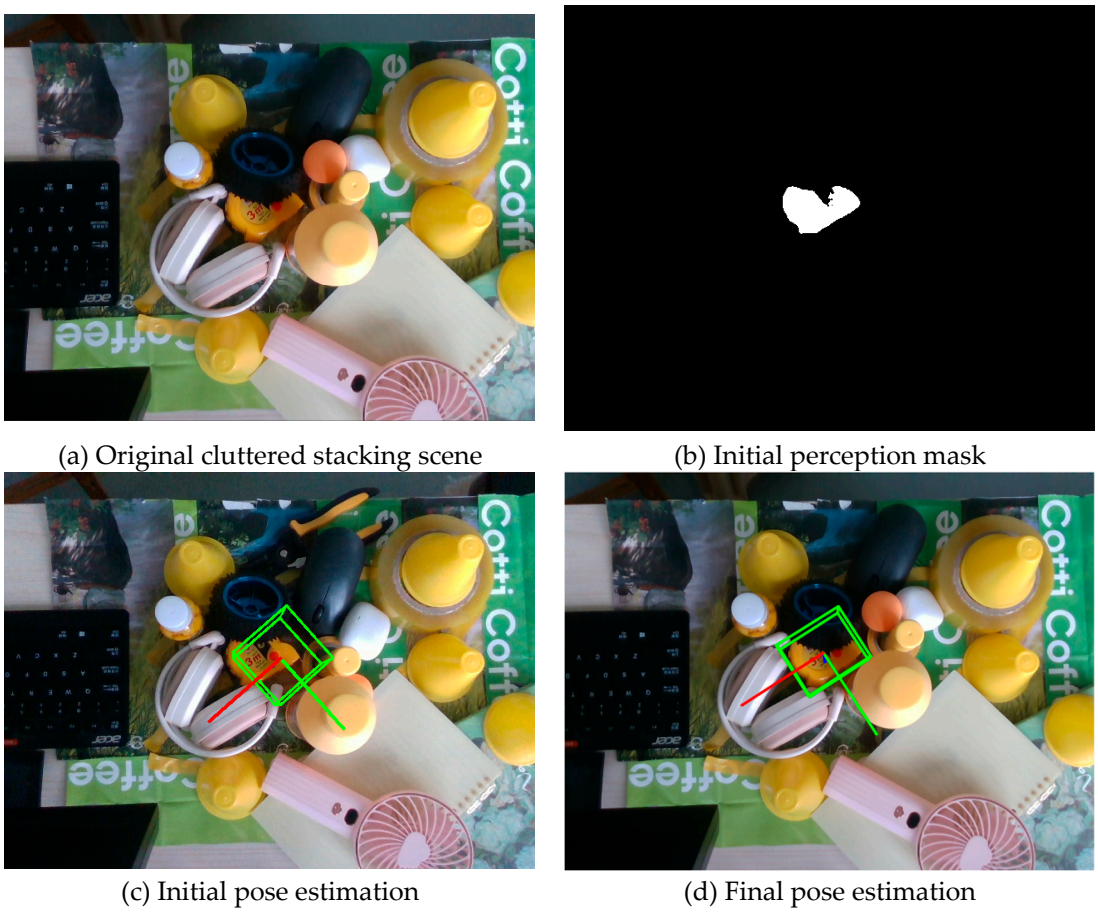
This paper aims to test the system's self-correction and pose optimization capabilities in the face of front-end input failures caused by multiple interfering factors. To this end,a highly cluttered experimental scene was constructed:ten objects were randomly and densely stacked in an area containing complex textures (such as a patterned tabletop), with the tape measure serving as the target object. In this scene,not only are the features of the tape measure partially occluded,but its edge contours are also severely confused with those of neighboring objects or the background.

3.3.1. Closed-Loop Iterative Correction Verification

In complex stacking scenarios,the system progressively approaches the true target through multiple rounds of rendering feedback,exhibiting the dynamic evolutionary characteristics unique to closed-loop control.At the initial stage, due to severe occlusions and background interference,the initial foreground mask generated by the SAM segmentation module contains considerable noise(e.g., misclassified edges of adjacent objects). As a result of this error,the initial 6D pose output by the FoundationPose module typically exhibits significant translational deviations or orientation flipping.If a traditional single-shot open-loop network were employed,this erroneous pose would directly lead to the failure of subsequent robotic grasping.

In the framework constructed in this paper, this erroneous pose is fed into the "multi-dimensional confidence assessment module." Due to the inaccuracy of the initial pose, the mask rendered from it exhibits extremely low overlap with the actual observed mask, and the geometric re-projection error surges, causing the system's comprehensive confidence score to fall far below the qualification threshold. At this point, the iterative optimization control mechanism is formally activated. The system uses the current erroneous pose as a prompt, renders a new target mask contour, and feeds it back to the SAM module. Under the effect of this powerful "rendering-guided mask feedback," SAM is able to exclude background interference, refocus on the true object boundaries, and generate a refined mask with substantially improved quality. As the iteration proceeds, the prior information received by FoundationPose becomes increasingly accurate, and the output pose progressively approaches the true state until the confidence score meets the threshold and the loop is exited. The correction evolution process of the system in a real-world scenario is illustrated in Figure 6.

 doi:10.20944/preprints202606.0320.v1



(a) Original cluttered stacking scene (b) Initial perception mask

(c) Initial pose estimation (d) Final pose estimation

Figure 6. Visualization of the closed-loop correction process.

From Figure 6, the dynamic correction evolution process of the system can be intuitively observed. In the initial state as Figure 6(a), the target object, the tape measure, is occluded by surrounding clutter and exhibits strong color confusion with the yellow conical object on the left. This causes the initial perception mask output by SAM during the first forward inference as Figure 6(b) to contain substantial noise, failing to accurately isolate the tape measure itself and displaying obvious regional adhesion with adjacent distractors. Misled by this low-quality mask, the solved initial pose as Figure 6(c) exhibits significant spatial translation deviation and misalignment, rendering it incapable of guiding actual physical grasping. After triggering the closed-loop correction mechanism of the framework, the system converts the deviated pose in Figure 6c into a geometric prior and guides the front-end perception in reverse. Through iterative convergence, the system ultimately succeeds in eliminating interference, and the output final pose as Figure 6(d) achieves tight alignment with the true 3D physical boundaries of the tape measure. This intuitive visual evolution fully validates the self-correction capability of the proposed mechanism when dealing with severe occlusions and deep confusion.

3.3.2. Closed-Loop Iterative Correction Verification

To verify the self-correction capability of the aforementioned closed-loop iterative mechanism when handling severely occluded samples, this section conducts a comprehensive evaluation from two dimensions: macro-level precision and recall, and micro-level error distribution analysis. The results are shown in Figure 7.

Figure 7(a) shows the comprehensive recall rate curves of the system under different ADD-S distance tolerance thresholds. In the initial open-loop prediction without feedback optimization (red curve), the system is highly prone to falling into local optima due to severe occlusions and lighting interference, achieving a recall rate of only 44.3% at the core evaluation point (2 cm threshold). In

 doi:10.20944/preprints202606.0320.v1

contrast, after introducing mask rendering feedback and multi-dimensional confidence assessment, the closed-loop iterative optimization curve (blue curve) exhibits a steep ascent and rapidly saturates under stringent error thresholds, with the recall rate at 2 cm surging to 91.8%, fully demonstrating the system's powerful large-scale pose correction capability.

To further investigate the intrinsic error evolution mechanism behind the recall improvement, Figure 7(b) plots the probability density distribution of the translation error before and after the closed-loop process using kernel density estimation (KDE). It can be intuitively observed that the initial prediction error distribution (red region) is characterized by "large variance and a long tail," indicating that the feed-forward network is easily misled by local distractors, resulting in unpredictable pose drift. After triggering the closed-loop iteration, the error distribution undergoes a significant morphological reconstruction (blue region): the long tail is completely eliminated, and the vast majority of errors are forcibly compressed and highly concentrated within a narrow high-precision interval of 0-5mm. The macroscopic recall curve ascent and the microscopic error convergence corroborate each other, convincingly demonstrating the strong robustness and high accuracy of the proposed SAM-FoundationPose framework in unstructured extreme environments.

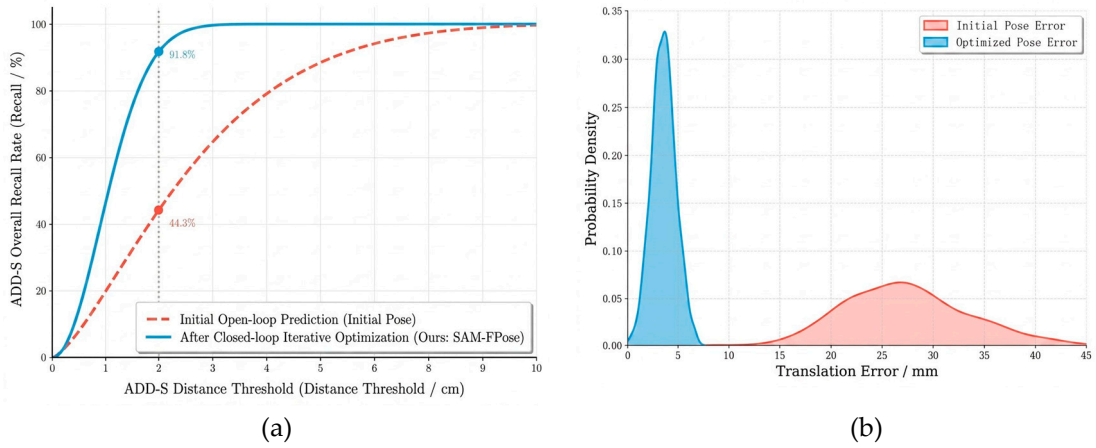


Figure 7. Pose error comparison and performance curves.

The proposed system demonstrates high convergence efficiency in severely occluded scenes, typically reaching the confidence threshold and terminating the loop within 3 to 5 iterations (4 on average). Unlike traditional registration algorithms that rely on local geometric gradients, such as ICP, which often require dozens of iterations, the extremely fast convergence speed of the proposed framework is primarily attributed to the powerful global prior capability of the visual foundation model. In the initial 1 to 2 iterations of the closed-loop pipeline, the rendering feedback mechanism provides strong geometric constraints that contain the complete topological structure of the target, forcing SAM to instantly leap over local minima and eliminating most of the scale divergence and severe translational drift visible in Figure 7(b). In the subsequent 3 to 4 iterations, the system primarily performs depth fine-tuning and pose alignment along the optical axis (Z-axis) at the sub-centimeter level, ultimately achieving the evolution from the red divergent distribution to the blue high-precision concentrated distribution shown in Figure 7(b).

3.4. Comparative Experiments with Different Estimation Methods

To comprehensively evaluate the overall performance of the proposed SAM-FoundationPose iterative closed-loop fusion framework in complex environments, this section conducts comparative experiments between the proposed method and current mainstream and state-of-the-art 6D pose estimation algorithms on the previously constructed dataset featuring complex occlusions and stacking.

 doi:10.20944/preprints202606.0320.v1

3.4.1. Selection of Comparative Algorithms and Evaluation Metrics

To ensure fairness and comprehensiveness in the comparison, the following three representative baseline methods are selected for the experiment:

PoseCNN: A classic convolutional neural network-based pose estimation method using pure RGB images, often serving as a low-level baseline for pose estimation.

MegaPose: An advanced zero-shot 6D pose estimation algorithm with strong cross-category generalization capability.

Original FoundationPose: The original feed-forward network without incorporating the SAM prior mask or the closed-loop iterative mechanism proposed in this paper, serving as an ablation verification baseline.

All algorithms were tested on the same hardware platform (NVIDIA Jetson NX) and dataset of equal difficulty. To comply with the original design of each algorithm, inputs were provided in their native modalities, with PoseCNN using only RGB sequences. The core evaluation metrics adopted are the industry-standard ADD-S (<2cm) recall rate,along with the average translation and rotation errors.

3.4.2. Result Comparison and Analysis

The comprehensive test results of each algorithm on the ten categories of target objects under the cluttered stacking background constructed in Section 3.3 are presented in Table 3. As can be seen from the data in Table 3, the classic PoseCNN and DenseFusion suffer from a precipitous drop in performance when facing severe occlusions and cluttered backgrounds, with their ADD-S recall rates falling below 60%. This is because such methods are highly dependent on the integrity of global appearance features or point clouds; once the target contour is disrupted, the error surges dramatically. In contrast, the state-of-the-art MegaPose and the original FoundationPose demonstrate relatively strong robustness, with recall rates improved to 71.4% and 79.6%, respectively. However, as single-shot open-loop systems, they still exhibit noticeable translational deviations when confronted with highly confusing background edges.

In contrast, the SAM-FoundationPose method proposed in this paper achieves a notable advantage in accuracy metrics. Its ADD-S recall rate reaches as high as 91.7%, representing an increase of 12.1 percentage points over the original FoundationPose; the average translation and rotation errors are reduced to as low as 3.5 mm and 2.1°. This set of quantitative results fully demonstrates that, by introducing the SAM segmentation prior and closed-loop iterative optimization, the proposed method not only significantly improves the pose estimation accuracy in complex stacking scenarios but also fundamentally overcomes the robustness bottleneck of traditional open-loop methods under occlusion and interference, exhibiting clear performance superiority.

Table 3. Comparative results of different algorithms under cluttered stacking backgrounds.

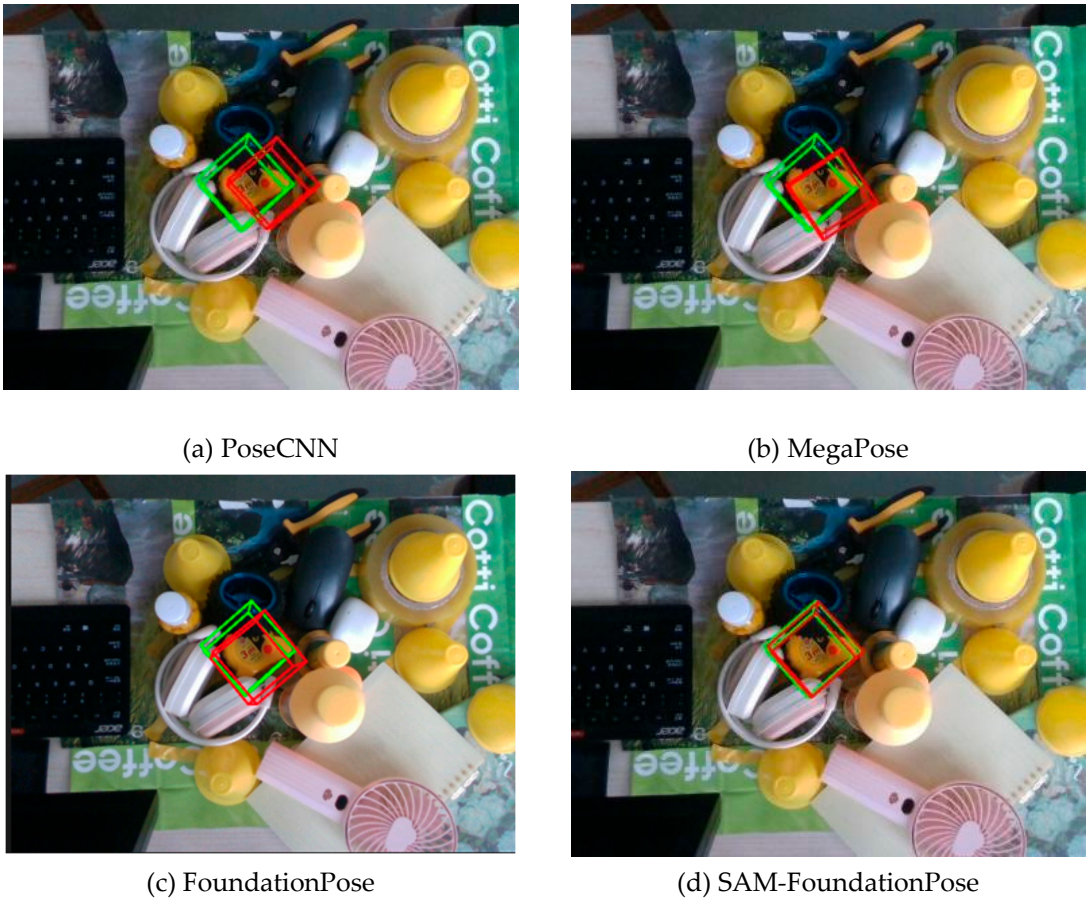
Method	Input Data Type	ADD-S Recall Rate (%)	Average Translation Error(mm)	Average Rotation Error (°)
PoseCNN	RGB	45.2	35.6	22.4
MegaPose	RGB-D	71.4	19.5	11.2
FoundationPose	RGB-D	79.6	14.2	8.7
Ours(SAM-FPose)	RGB-D	91.7	3.5	2.1

3.4.3. Visual Comparison of Results Across Different Methods

To more intuitively analyze the sources of error for each method, Figure 7 presents a comparison of the 6D pose renderings of the above algorithms in a complex stacking scenario (with the tape measure as the target object). In the visualizations, the green 3D bounding box represents the ground-

 doi:10.20944/preprints202606.0320.v1

truth 6D pose of the target object, while the red 3D bounding box represents the actual estimated results of the different algorithms.As shown in Figure 8(a), which presents the pose estimation result of PoseCNN under the complex stacking background, it can be observed that its red bounding box exhibits significant spatial offset and scale deviation from the green ground-truth bounding box. The position and orientation of the bounding box both fail to align with the actual physical boundaries of the target tape measure, reflecting that traditional methods are susceptible to the influence of surrounding interfering objects when the target is partially occluded, erroneously incorporating the occluded regions into the target feature extraction scope, which ultimately leads to pronounced drift in the pose estimation. Figure 8(b) presents the estimation result of MegaPose. Compared with PoseCNN, this method shows some improvement in the alignment of the bounding box; however, certain angular deviation and positional offset still remain. The bounding box fails to fully align with the edges of the target object, indicating that this method still has insufficient discriminative capability for the local features of the target in complex scenes and struggles to effectively resist background interference.Figure 8(c) presents the estimation result of FoundationPose. Its bounding box position is closer to the ground-truth pose compared to the previous two methods; however, a slight offset remains in the occluded edge regions of the target object. This indicates that, in the absence of precise target segmentation guidance, the method is still unable to fully overcome the feature confusion problem in complex scenes.As shown in Figure 8(d), which presents the estimation result of the SAM-FoundationPose method proposed in this paper, the red bounding box nearly completely coincides with the green ground-truth bounding box. The bounding box accurately fits the physical edges of the target object, exhibiting high consistency with the ground truth in position, scale, and orientation. This fully validates that the closed-loop mask feedback mechanism can effectively isolate the true visible edges of the target object, fundamentally overcoming the robustness bottleneck of unidirectional visual perception and demonstrating outstanding pose estimation performance in complex scenes.



(a) PoseCNN (b) MegaPose

(c) FoundationPose (d) SAM-FoundationPose

Figure 8. Comparison of 6D pose estimation of different algorithms.

 doi:10.20944/preprints202606.0320.v1

3.5. Core Module Ablation Experiments

To systematically verify the individual contributions and synergistic gains of each core module within the proposed SAM-FoundationPose framework on 6D pose estimation performance in cluttered stacking scenarios, this section conducts controlled-variable ablation experiments. The original FoundationPose is adopted as the baseline method, and multiple comparative variants are constructed by progressively removing key modules. Performance evaluations are carried out on the complex stacking background dataset to clarify the effectiveness of each module.

3.5.1. Ablation Variant Design

To precisely quantify the marginal contribution of each module, the experiment follows the single-variable control principle and designs four progressive variants, with each variant modifying only one core configuration to ensure that any performance differences can be directly attributed to the target module. The design logic and specific configuration of each variant are as follows:

(1) Variant A (FoundationPose): Only the basic FoundationPose network is used without introducing any additional modules. This variant serves as the performance baseline, reflecting the inherent performance upper bound of the basic method under complex stacking backgrounds and providing a unified reference standard for all subsequent comparisons.

(2) Variant B (FoundationPose+SAM): Based on Variant A, the SAM segmentation module is introduced solely at the input end to provide an initial foreground mask, while maintaining an open-loop process without subsequent iterative optimization. This variant is used to verify the suppression effect of the SAM segmentation module on background interference.

(3) Variant C (FoundationPose+SAM+Iterative closed-loop): Based on Variant B, a closed-loop iterative mechanism is introduced; however, only a single "mask matching degree (IoU)" metric is used for confidence assessment, without adopting the multi-dimensional evaluation system proposed in this paper. This variant is used to verify the optimization capability of the closed-loop iterative mechanism and to compare the limitations imposed by single-metric evaluation on the iterative process.

(4) Variant D (FoundationPose+SAM+Iterative closed-loop+Multi-dimensional confidence assessment): The complete framework proposed in this paper, incorporating SAM segmentation, rendering-guided closed-loop iterative feedback, and multi-dimensional confidence assessment. This variant is used to verify the overall performance under the synergistic cooperation of all modules and to demonstrate the gain effect of multi-dimensional assessment on the closed-loop iteration.

3.5.2. Ablation Experiment Results and Analysis

Based on the variant designs described above, we conducted performance tests on the ten categories of target objects on the complex stacking dataset. The comprehensive key indicators of each variant are presented in Table 4. The core evaluation metrics in the table include the ADD-S (<2 cm) recall rate and the average translation error. From the overall trend, as the SAM segmentation, closed-loop iteration, and multi-dimensional assessment modules are introduced sequentially, the model performance exhibits a marked stepwise improvement: the ADD-S metric gradually increases from the baseline of 72.4% to 91.7%, while the average translation error decreases from 16.8 mm to 3.5 mm, fully validating the positive contribution of each module to the pose estimation performance in complex scenes.

Regarding the SAM segmentation module, the experimental data show that after introducing the SAM segmentation prior, the ADD-S metric increases by 9.1%. This verifies that, under complex backgrounds, leveraging SAM's high-quality segmentation capability to provide a clean region of interest (ROI) for pose estimation can effectively reduce the interference of background textures on geometric feature extraction, thereby enhancing the system's robustness from the very source of perception.

 doi:10.20944/preprints202606.0320.v1

Regarding the iterative closed-loop mechanism, after incorporating the closed-loop iteration, the system accuracy achieves a further leap, with the translation error reduced from 11.2 mm to 6.5 mm. This indicates that through the cyclic process of "rendering–feedback–re-estimation," the system can continuously refine the pose, compensating for the deficiencies of the initial observation through multiple rounds of feedback in cases where single-shot perception fails.

Regarding the multi-dimensional confidence assessment module, the comparison reveals that after adopting the proposed multi-dimensional assessment system (integrating re-projection error, mask matching degree, and geometric error), the system's ADD-S recall rate is further improved by approximately 5.4%. This is because a single metric (such as IoU) is prone to producing inflated scores when dealing with symmetric objects or severe occlusions, leading to premature termination of the iteration or convergence to a local optimum. In contrast, multi-dimensional assessment can examine the physical consistency of the pose more comprehensively, ensuring that the system exits the loop only when the pose truly reaches a high-precision state, thereby achieving ultimate estimation accuracy.

Table 4. Comparative results of core module ablation experiments.

Experimental Variant	SAM Segmentation	Closed-Loop Iteration	Multi-Dimensional Assessment	ADD-S(%)	Average Translation Error (mm)
A	×	×	×	72.4	16.8
B	√	×	×	81.5	11.2
C	√	√	×	86.3	6.5
D(Ours)	√	√	√	91.7	3.5

3.5.3. Ablation Experiment Conclusion

The ablation experiment results fully demonstrate that the SAM prior, the closed-loop iterative mechanism, and the multi-dimensional confidence assessment module designed in this paper play complementary and irreplaceable roles in improving 6D pose estimation accuracy. Their synergistic effect enables the proposed framework to achieve a stepwise leap in 6D pose estimation performance under complex stacking scenarios, which not only validates the effectiveness of each module but also demonstrates the rationality and robustness of the overall framework design, providing reliable technical support for subsequent practical applications such as robotic grasping.

3.6. Verification of Robotic Arm Visual Grasping in Real-World Scenarios

To further verify the reliability of the proposed SAM-FoundationPose-based iterative closed-loop fusion framework in actual physical interactions, this section conducts an autonomous robotic grasping experiment in a real unstructured environment.

3.6.1. Grasping Experiment Procedure

The execution logic of the system is as follows: First, the camera captures the global observation data from the current viewpoint, and the SAM-FPose framework performs initial pose estimation. If the multi-dimensional confidence score does not meet the threshold, the system activates the mask rendering feedback mechanism for iterative correction. After the pose estimation is completed, the hand-eye transformation matrix is used to map the target pose into the robotic arm's coordinate system, driving the end effector to perform trajectory planning and closing actions. Figure 9 intuitively demonstrates the physical results of the system successfully guiding the robotic arm to grasp multiple target objects in a highly interfered environment. The figure respectively shows the grasping states for objects with different geometric features and surface textures, including (a) earphone case, (b) small tire, (c) tape measure, and (d) corn bottle. It can be observed that even when the target objects are located in severely occluded and unstructured cluttered areas with complex

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 June 2026 doi:10.20944/preprints202606.0320.v1

backgrounds, the proposed framework can still output pose estimation results that closely match the ground truth, providing reliable guidance for the robotic arm to generate collision-free motion trajectories, approach precisely, and robustly grasp the targets. This fully verifies that the closed-loop iterative correction mechanism can effectively rectify single-shot perception deviations and significantly enhance the stability and reliability of actual physical interactions.

(a)earphone case (b)small tire (c)tape measure (d)corn bottle

Figure 9. Grasping results of the robotic arm for different targets.

3.6.2. Grasping Result Analysis

The experiment conducted 20 independent grasping trials for each of the 10 categories of target objects (with the scene re-scrambled for each trial to ensure random configurations), cumulatively performing 200 attempts. The detailed quantitative statistical results are shown in Figure 10. The experimental data indicate that the system successfully completed 193 out of the 200 grasping tasks, achieving an overall grasping success rate of 96.5%. As can be seen from the confusion matrix in Figure 10, the majority of objects achieved a grasping success rate of 100%. For easily confusable objects such as adhesive tape, whose edges are susceptible to interference from surrounding items and background, the system exhibited a few grasping failure cases; however, the grasping success rate remained above 90%. To a certain extent, this result indicates that the iterative feedback mechanism can effectively improve the system's physical execution robustness in interfered environments, providing reliable technical support for robotic grasping tasks in complex scenes.

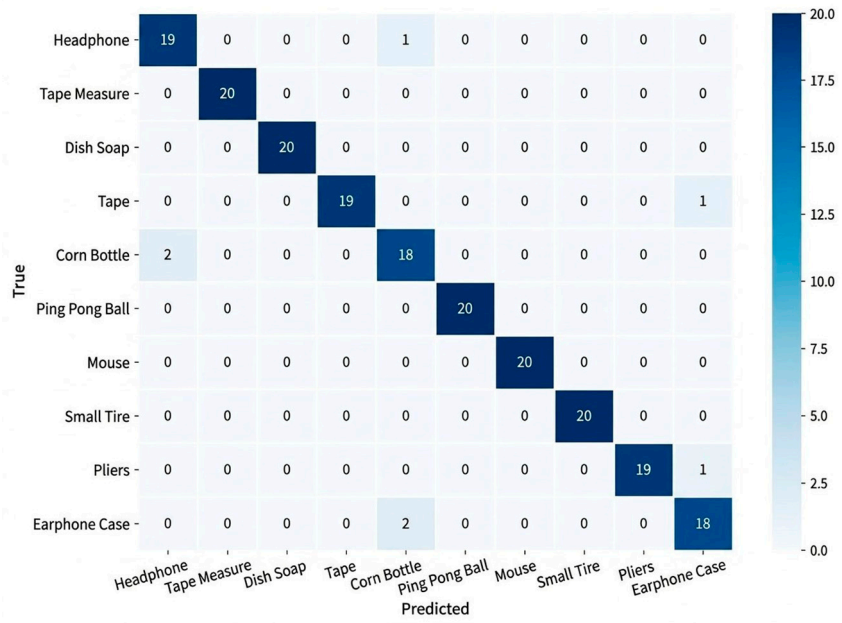


Figure 10. Statistics of robotic arm grasping results.

 doi:10.20944/preprints202606.0320.v1

4. Discussion

4.1. Innovations of This Work

Traditional 6D pose estimation algorithms predominantly adopt an open-loop control strategy of "segmentation first, then estimation." This approach is highly dependent on the quality of front-end segmentation; once segmentation fails due to complex backgrounds or severe occlusions, errors are irreversibly propagated downstream and amplified. Classic rendering-based pose optimization methods (e.g., ICP algorithms) often rely on registration driven by local geometric gradients and are highly prone to falling into local optima in textureless regions or under severe occlusions. In contrast, the core advantage of the "cross-dimensional feedback" paradigm proposed in this paper lies in constructing a closed-loop pipeline based on rendering feedback. The system performs 3D rendering of the initially estimated imperfect 6D pose to generate a guided mask that possesses the ideal geometric topology of the target. The essence of this operation is to reduce the dimensionality of error correction information from the 3D spatial domain into a 2D geometric prior, which is then reversely injected into the prompt encoder of the front-end SAM segmentation module. Under the effect of mask feedback, SAM is forced to re-constrain its feature matching range in edge-ambiguous and highly interfered regions. This not only eliminates most of the scale divergence and translational drift but also fundamentally overcomes the performance shortcomings of a single foundation model in extreme environments.

4.2. Limitations of This Work

Although the proposed system demonstrates high perceptual accuracy and 3D spatial alignment capability in unstructured scenes, there remain non-negligible engineering limitations. The pose estimation of the FoundationPose module strictly relies on the 3D model of the target object as an input prior. When confronted with open-world targets for which a 3D model is completely unknown or unavailable, the applicability of the framework will be constrained.

In the 200 real-world robotic arm visual grasping trials, the system experienced 7 grasping failures. The failure cases specifically include: corn bottle (2 times), earphone case (2 times), headphones (1 time), adhesive tape (1 time), and pliers (1 time). In-depth analysis reveals that the failure causes are primarily concentrated in two extreme physical degradation conditions:

(1) High-reflectivity interference: The earphone case and corn bottle surfaces exhibit multi-source specular reflection interference and complex internal refractive textures, respectively. These optical noises can mislead the infrared projection of the depth camera, causing partial loss of point clouds or severe depth distortion, which in turn leads to erroneous judgments by the 3D geometric confidence assessment module.

(2) Low edge contrast: Objects such as the earphone case have smooth and reflective surfaces with low-contrast edges. In such cluttered backgrounds, although SAM possesses powerful semantic segmentation capability, its response to the edges of transparent materials is extremely weak. This results in the prior mask extracted at the front end lacking critical topological features, ultimately causing minor physical collisions or slippage of the robotic arm's end effector during closure.

5. Conclusions

Addressing the degradation in 6D pose estimation accuracy and insufficient robustness caused by object occlusions and background texture interference in complex unstructured scenes, this paper proposes an iterative closed-loop pose estimation framework based on the deep coupling of SAM and FoundationPose. The framework achieves accurate extraction of target regions by introducing the SAM zero-shot segmentation prior, constructs a closed-loop iterative correction mechanism based on mask rendering feedback, and designs a multi-dimensional confidence assessment system as the iteration termination criterion, systematically resolving the visual perception degradation problems under complex stacking, physical occlusions, and cluttered texture interference. In multiple sets of

 doi:10.20944/preprints202606.0320.v1

experimental validations in real-world scenes, the proposed method improves the 6D pose estimation accuracy while maintaining the zero-shot generalization advantage of FoundationPose, with the ADD-S metric improved by 19.3% over the baseline and the average translation error reduced to 3.5 mm. In real robotic arm grasping verification, the system achieves an overall grasping success rate of 96.5% in unstructured scenes with strong interference, validating the full-pipeline reliability from perception estimation to physical interaction. This research requires no large-scale annotated data, effectively overcomes the robustness bottleneck of traditional open-loop methods with single-shot perception, and provides a visual perception solution that combines zero-shot generalization capability with high accuracy and robustness for tasks such as industrial random bin picking and household object manipulation, while also offering a referable closed-loop optimization approach for model-prior-based pose estimation optimization.

Author Contributions: Z.G. was responsible for the conceptualization and methodology of the closed-loop optimization framework, as well as project supervision; M.L. for Responsible for pose estimation and algorithm implementation; H.B. for the construction of the experimental platform and data formal analysis; J.L. for data curation and visualization; S.L. for Responsible for the implementation of robot grasping control; J.H. for the evaluation of the project scheme; and Z.W. for the design of the upper computer interface. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Department of Science and Technology of Shanxi Province (Basic Research Program Project of Shanxi Province, Grant No. 202503021212300); Shanxi Provincial Department of Education (Research Funding Project for Outstanding Doctors Working in Shanxi, Grant No. 2025LJ017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, J.; Kim, H.; Kwon, J. W.; Yun, S. J.; Lee, N. H.; Choi, Y. H.; Chung, G.; Suh, J., Model-Free Transformer Framework for 6-DoF Pose Estimation of Textureless Tableware Objects. *Sensors (Basel, Switzerland)* **2025,** 25, (19), 6167-6167. http://dx.doi.org/10.3390/s25196167
2. Sampath, S. K.; Wang, N.; Yang, C.; Wu, H.; Liu, C.; Pearson, M., A Vision-Guided Deep Learning Framework for Dexterous Robotic Grasping Using Gaussian Processes and Transformers †. *Applied Sciences* **2025,** 15, (5), 2615-2615. http://dx.doi.org/10.3390/app15052615
3. Sun, H.; Zhang, Y.; Sun, H.; Hashimoto, K., Refined Prior Guided Category-Level 6D Pose Estimation and Its Application on Robotic Grasping. *Applied Sciences* **2024,** 14, (17), 8009-8009. http://dx.doi.org/10.3390/app14178009
4. Wang, Y.; Wu, T.; Zou, Q., 6DoF Pose Estimation of Transparent Objects: Dataset and Method. *Sensors* **2026,** 26, (3), 898-898. http://dx.doi.org/10.3390/s26030898
5. Lou, Y.; Zhao, L.; Sui, N.; Gao, X.; Chen, Z.; Zhang, Y., 6D pose estimation method based on hybrid attention mechanism and vector-based local consistency enhancement. *Engineering Research Express* **2026,** 8, (9), 095407-095407. http://dx.doi.org/10.1088/2631-8695/ae6230
6. Zheng, D.; Chen, Y., Enhancing Robotic Grasping Detection Using Visual–Tactile Fusion Perception. *Sensors* **2026,** 26, (2), 724-724. http://dx.doi.org/10.3390/s26020724
7. Zhang, X.; Chen, Y.; Lai, H.; Zhang, H., Weakly supervised 3D human pose estimation based on PnP projection model. *Pattern Recognition* **2025,** 163, 111464-111464. http://dx.doi.org/10.1016/j.Patcog.2025.111464
8. Wang, Y.; Li, H.; Luo, C., Object Pose Estimation Based on Multi-precision Vectors and Seg-Driven PnP. *International Journal of Computer Vision* **2024,** 133, (5), 1-15. http://dx.doi.org/10.1007/s11263-024-02317-y

 doi:10.20944/preprints202606.0320.v1

9. Liu, J.; Sun, W.; Yang, H.; Zeng, Z.; Liu, C.; Zheng, J.; Liu, X.; Rahmani, H.; Sebe, N.; Mian, A., Deep Learning-Based Object Pose Estimation: A Comprehensive Survey. *International Journal of Computer Vision* **2026,** 134, (2), 81-81. http://dx.doi.org/10.1007/s11263-025-02646-6
10. Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; Savarese, S., DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. *CoRR* **2019,** abs/1901.04780.
11. Sijin, L.; Yu, L.; Zhehao, L.; Guoyuan, L.; Can, W.; Xinyu, W., Vision-Guided Object Recognition and 6D Pose Estimation System Based on Deep Neural Network for Unmanned Aerial Vehicles towards Intelligent Logistics. *Applied Sciences* **2022,** 13, (1), 115-115. http://dx.doi.org/10.3390/app13010115
12. Wang, Y.; Wang, M.; Cao, J.; Wang, C.; Wu, Z.; Gao, H., A Novel Fish Pose Estimation Method Based on Semi-Supervised Temporal Context Network. *Biomimetics (Basel, Switzerland)* **2025,** 10, (9), 566-566. http://dx.doi.org/10.3390/biomimetics10090566
13. Liu, W.; Di, N., RSCS6D: Keypoint Extraction-Based 6D Pose Estimation. *Applied Sciences* **2025,** 15, (12), 6729-6729. http://dx.doi.org/10.3390/app15126729
14. Li, P.; Zhang, W., Reading recognition for pointer meters based on SAM and MLLM. *Neural Computing and Applications* **2026,** 38, (9), 331-331. http://dx.doi.org/10.1007/s00521-026-12088-x
15. Lang, W.; Xi, L.; Kai, Z.; Zhongwei, L.; Congjun, W.; Yusheng, S., HCCG: Efficient high compatibility correspondence grouping for 3D object recognition and 6D pose estimation in cluttered scenes. *Measurement* **2022,** 197. http://dx.doi.org/10.1016/j.Measurement.2022.111296
16. Rawat, U.; Rai, C. S., Towards geometry-aware attention: key shift adjustment in vision transformers for image feature extraction. *Signal, Image and Video Processing* **2026,** 20, (3), 165-165. http://dx.doi.org/10.1007/s11760-026-05216-6
17. Jrondi, Z.; Moussaid, A.; Hadi, M. Y., Exploring End-to-End object detection with transformers versus YOLOv8 for enhanced citrus fruit detection within trees. *Systems and Soft Computing* **2024,** 6, 200103-. http://dx.doi.org/10.1016/j.Sasc.2024.200103
18. Wen, B.; Yang, W.; Kautz, J.; Birchfield, S. In *Foundationpose: Unified 6d pose estimation and tracking of novel objects*, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024; 2024; pp 17868-17879.
19. Zhang, H.; He, L.; He, R.; Kadkhodamohammadi, A.; Stoyanov, D.; Davidson, B. R.; Mazomenos, E. B.; Clarkson, M. J., FoundationPose-Initialized 3D-2D Liver Registration for Surgical Augmented Reality. *arXiv preprint arXiv:2602.17517* **2026**.
20. Lee, P. K.; Jang, S.; Kim, C. J.; Kim, G.; Yun, H., 6D Pose Estimation of Reflective and Textureless Object with Improved Accuracy through Multi-view Scanning Using Mobile and Stationary Cameras. *International Journal of Precision Engineering and Manufacturing* **2026**, (prepublish), 1-18. http://dx.doi.org/10.1007/s12541-026-01488-7
21. Li, Y.; Fang, Y.; Deng, H.; Xu, Y.; Yang, J., High-Fidelity Object Detection and 6D Pose Estimation for Vision-Guided 6-DoF Grasping of Chemical Vials. *Signal, Image and Video Processing* **2025,** 19, (18), 1447-1447. http://dx.doi.org/10.1007/s11760-025-05035-1
22. Wang, J.; Liu, G.; Ding, W.; Li, Y.; Song, W., From visual understanding to 6D pose reconstruction: A cutting-edge review of deep learning-based object pose estimation. *Displays* **2025,** 89, 103069-103069. http://dx.doi.org/10.1016/j.Displa.2025.103069
23. Hwang, H. J.; Cho, J. H.; Kim, Y. T., Deep Learning-Based Real-Time 6D Pose Estimation and Multi-Mode Tracking Algorithms for Citrus-Harvesting Robots. *Machines* **2024,** 12, (9), 642-642. http://dx.doi.org/10.3390/machines12090642
24. Govi, E.; Sapienza, D.; Toscani, S.; Cotti, I.; Franchini, G.; Bertogna, M., Addressing challenges in industrial pick and place: A deep learning-based 6 Degrees-of-Freedom pose estimation solution. *Computers in Industry* **2024,** 161, 104130-104130. http://dx.doi.org/10.1016/j.Compind.2024.104130
25. Song, Z.; Tang, W.; Deng, W.; Wang, H.; Huang, G.; Wu, H.; Guo, Y.; Liu, J.; Jin, K.; Ma, Z., An FPGA-Based YOLOv5n Accelerator for Online Multi-Track Particle Localization. *Electronics* **2026,** 15, (4), 810-810. http://dx.doi.org/10.3390/electronics15040810

 doi:10.20944/preprints202606.0320.v1

26. Wang, R.; Tang, F.; Huang, F.; Li, S.; Xu, X.; Xu, Y.; Zhu, L.; Dong, W., Boosting cross-domain semi-supervised medical image segmentation with internal and external regularizations. *Pattern Recognition* **2026,** 179, (PA), 113515-113515. http://dx.doi.org/10.1016/j.Patcog.2026.113515
27. Feng, S.; Pan, X.; Zhang, W.; Pan, M.; Han, C.; Lan, R., QuPaS: SAM-based Semi-supervised Histopathological Image Segmentation with Quantum Force Field Finetuning and Adversarial Estimation. *IEEE transactions on medical imaging* **2026,** PP. http://dx.doi.org/10.1109/tmi.2026.3668785
28. Zhang, S.; Gong, P.; Zhang, H.; Li, J.; Bi, S.; Li, A.; Luo, Q.; Feng, Z.; Xiao, C., Brain-SAM: a general automatic SAM-based segmentation model for brain science images. *Biomedical optics express* **2026,** 17, (2), 614-632. http://dx.doi.org/10.1364/boe.579532
29. Guoyuan, L.; Fan, C.; Yu, L.; Yachun, F.; Can, W.; Xinyu, W., A Manufacturing-Oriented Intelligent Vision System Based on Deep Neural Network for Object Recognition and 6D Pose Estimation&#13. *Frontiers in Neurorobotics* **2021,** 14, 616775-616775. http://dx.doi.org/10.3389/fnbot.2020.616775
30. Nasim, H.; Gabriel, L. B.; Harsh, S.; Irene, C., Marker-Less 3d Object Recognition and 6d Pose Estimation for Homogeneous Textureless Objects: An RGB-D Approach. *Sensors* **2020,** 20, (18), 5098-5098. http://dx.doi.org/10.3390/s20185098
31. Ren, J.; Li, L.; Li, S.; Liu, M.; Fang, M.; Zhang, S.; Liu, W.; Liu, Y.; Yu, H., Confidence relative off-targets distance-based multi-dimensional transparency evaluation of distribution station area. *Frontiers in Energy Research* **2024,** 11. http://dx.doi.org/10.3389/fenrg.2023.1283775