IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 November 2023, accepted 30 November 2023, date of publication 4 December 2023,
date of current version 11 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339163

RESEARCH ARTICLE

Research on Robot Monocular Vision-Based 6DOF Object Positioning and Grasping Approach Combined With Image Generation Technology

GUOYANG WAN, JINCHENG CHEN, JIAN ZHANG, BINYOU LIU, HONG ZHANG, AND XIUWEN TAO
School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China

Corresponding author: Guoyang Wan (704610266@qq.com)

This work was supported in part by the Startup Fund for Introducing Talents and Scientific Research of Anhui Polytechnic University under Grant 2022YQQ011, in part by the 2022 Anhui Provincial Key Laboratory Open Fund for Electrical Transmission and Control under Grant DQKJ202206, and in part by the Anhui Polytechnic University Jiujiang Industrial Collaborative Innovation Special Fund Project 2022cyxtb10.

ABSTRACT To address the challenges related to poor positioning accuracy and high usage cost of 6DOF visual measurement systems in industrial settings, this paper presents a monocular vision-based robot vision guidance approach. The goal is to address the issues of expensive 6DOF pose measurement and limited measurement robustness when robots need to manipulate metal objects in industrial environments. The proposed approach enables precise and robust measurement of the 6DOF pose of the target workpiece. The approach integrates two main algorithms: a virtual reality-based image data enhancement algorithm and a 6DOF pose measurement algorithm that combines a multi-keypoint detection model and the Efficient Perspective-n-Points (EPnP) algorithm. The image data enhancement algorithm enhances the data of small-sample industrial objects using image enhancement techniques. This improves the robustness of the detection model by mitigating the challenges of high-cost image acquisition and long acquisition time associated with industrial objects. On the other hand, the 6DOF pose measurement algorithm performs the pose measurement of the target workpiece using a single image, enabling cost-effective 6DOF pose measurement by utilizing only a monocular camera. Experimental results demonstrate that the proposed method achieves measurement errors of 4.21% in the X direction, 2.94% in the Y direction, and 0.39% in the Z direction of the target workpiece. These results highlight the effectiveness of the proposed approach in achieving accurate and reliable pose measurement.

INDEX TERMS Machine vision, industrial robot, data augmentation, object detection, visual guidance.

I. INTRODUCTION

With the advancement of machine vision technology, 2D vision-based robot visual guidance has found widespread application in the industrial sector. Industrial robot systems equipped with visual guidance offer several advantages, including cost reduction in labor, enhancement of work quality, and optimization of production cycles. Currently, existing 2D vision systems are commonly utilized for 3DOF positioning tasks that do not necessitate depth information [1]. In the industrial sector, due to the complex application environments and high stability requirements, single-camera visual 6DOF pose measurement technology is not yet prevalent. However, in recent years, the continuous progress of artificial intelligence technology has injected fresh vitality into vision-based pose measurement techniques [2]. Integration of deep learning-based image processing techniques within the domain of machine vision holds the potential to reduce

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.

the cost associated with implementing robot visual guidance while also paving the way for future upgrades in the manufacturing industry.

The vision sensor based on machine vision technology provides industrial robots with an accurate, efficient, and cost-effective perception system, addressing the weak environmental perception capability of industrial robots [3]. Among the available vision sensors, the 2D vision system based on monocular vision stands out due to its combination of high sampling speed and stability, making it a cost-effective choice for visual systems. However, when it comes to measuring the 6DOF pose of objects, traditional 2D vision systems suffer from low measurement accuracy and poor robustness, which restricts their application in the industrial field [4].

There have been numerous successful cases of integrating robot systems with machine vision. In reference [5], combining stereo vision with virtual reality and heavy-duty industrial robots improved operational efficiency by enabling the robots to handle heavy loads. Reference [6] utilized a combination of 2D and 3D vision to accomplish positioning and grasping tasks for mobile platforms and 6DOF robots. In reference [7], a 2D vision system was employed to achieve high-precision assembly of target workpieces using industrial robots. All of these cases demonstrate the successful utilization of machine vision in the industrial domain. Nevertheless, while there are numerous applications of 2D vision in the field, relatively few incorporate 6DOF pose measurement and grasping of target workpieces by industrial robots. Monocular vision systems face challenges in obtaining accurate 6DOF pose information from 2D images. Compared to other technologies, single-camera vision's high speed and stability provide an irreplaceable edge in the industrial space. Consequently, achieving stable 2D vision 6DOF pose measurement remains a significant research topic in the industrial field.

Numerous researchers have conducted studies on 6DOF pose measurement techniques based on 2D vision. Early methods often utilized template matching for measuring the 6DOF pose of the target object. CAD-view [8] and the Line-MOD algorithm [9] are notable examples. The CAD-view algorithm determines the 6DOF pose of the target object by employing its 3D model and has become a well-established method in the industrial field. Subsequently, the Line-MOD method incorporated depth information of the target object, building upon the CAD-view approach and enhancing the accuracy and robustness of the algorithm. However, these algorithms heavily rely on a single contour feature, making it challenging to guarantee localization accuracy and precision when the contour feature of the target object is not visible. In recent years, with the advancement of artificial intelligence, many scholars have explored the use of deep neural networks to measure the 6DOF pose of target objects [10]. SSD-6D [11], YOLO-6D [12], and other methods have emerged as representative approaches in this category. Deep learning-based methods deduce the pose information of the target object from high-dimensional features in the images and exhibit superior generalization performance compared to traditional image processing techniques. Nevertheless, these methods depend on extensive training data and suffer from noticeable drawbacks in terms of measurement accuracy.

To broaden the application of monocular vision in industrial robot guidance, this paper presents a deep learning-based strategy for the 6DOF localization and grasping of robots using monocular vision. The strategy encompasses an image generation technique based on deep learning, which allows for data augmentation of industrial objects with limited samples. Robust pose measurement of surface key points of the target object is achieved through an enhanced anchor-based object detection network [13]. By combining this with the EPnP pose estimation iterative algorithm [14], the robot becomes capable of measuring the 6DOF pose and grasping the target workpiece. The proposed method represents a valuable exploration of utilizing monocular vision for guiding robots in 6DOF positioning and grasping tasks within real-world industrial environments. It has the potential to significantly reduce the cost of visual systems in the industrial field, expand the range of robot applications, and enhance production intelligence in the future.

The proposed algorithm primarily focuses on workpieces with planar features. The main innovations of this paper are as follows:

1) A novel method was proposed that combines virtual reality and image generation technology to augment the data of small-sample industrial objects. By leveraging a virtual engine and generative adversarial networks, data augmentation for industrial objects with limited samples is achieved.
2) A visual-guided grasping strategy for robotic manipulation of target workpieces was introduced, utilizing monocular vision technology. This strategy enables accurate 6DOF pose estimation of objects with planar features, facilitating precise and efficient grasping.
3) The integration of generative adversarial networks and attention mechanisms led to the development of an image enhancement algorithm based on image generation.

This algorithm aims to enhance the quality of images and improve visual perception, leading to more accurate object detection and pose estimation. The remainder of this paper is organized as follows: Section II provides a brief review of the related work in the field. Section III presents our monocular vision detection strategy, which utilizes virtual reality and object detection methods. In Section IV, we describe the experimental results and validate the effectiveness of the proposed method by comparing its performance with other detection methods. Finally, Section V concludes the paper and summarizes the key findings.

II. RELATED WORK

With the ongoing digitalization, informatization, and intelligent transformation of the manufacturing industry, robots are increasingly being utilized in various fields such as welding and polishing. As a result, the performance requirements for robots in these applications are becoming more demanding. Industrial robots equipped with 6DOF pose measurement capabilities have become the preferred automation equipment in the industrial sector due to their expansive workspace, compact design, and high degree of freedom. In the early stages of development, visual localization techniques relied on template matching algorithms that utilized prior knowledge of the target's geometry and pose to establish a model. Methods such as those proposed by Cao et al. [15] and Hinterstoisser et al. [16] required the construction of pre-existing 3D models of the targets and the generation of template image-matching libraries by rendering the 3D models from various viewpoints. During testing, the estimated image would be matched against the template images, and the pose corresponding to the highest match would be considered as the result of pose estimation. While these methods were effective in handling textureless objects, their accuracy heavily depended on the completeness of the matching database. Furthermore, their efficiency and robustness were reliant on the matching strategy employed, and they exhibited limitations in dealing with partially occluded objects and changes in appearance. Consequently, their applicability in complex scenes was hindered. These limitations highlight the need for more advanced techniques that can overcome the challenges posed by complex industrial environments. The proposed algorithm in this paper aims to address these limitations by introducing a deep learning-based strategy for monocular vision 6DOF localization and grasping of robots. By utilizing deep learning techniques and an improved anchor-based object detection network, the algorithm achieves robust pose measurement of surface key points of the target object. In combination with the EPnP pose estimation iterative algorithm, the robot is capable of accurate 6DOF pose measurement and grasping of the target workpiece. The proposed method represents a valuable exploration of utilizing monocular vision to guide robots in 6DOF positioning and grasping tasks within real-world industrial environments, with the potential to reduce the cost of visual systems in the industrial field and enhance production intelligence.

Researchers have made significant efforts to overcome the limitations of template matching techniques. One notable contribution is the method proposed by Crivellaro [17], which focuses on estimating object pose in the presence of cluttered backgrounds. An interesting aspect of this method is that it eliminates the requirement for a color camera, enabling real-time object pose estimation using only a grayscale camera. The approach achieves this by predicting the three-dimensional pose of each object part through the projection of multiple key points in two dimensions.

To address the issue of slow matching speed encountered with template matching, especially when dealing with a large number of templates, Konishi [18] introduced a novel monocular image 6D pose estimation method based on PCOF (Pose Cluster and Outlier Filtering) and HPT (Hierarchical Pose Tree). By efficiently clustering object poses and utilizing a hierarchical pose tree, the method achieved faster pose estimation even when dealing with a large number of templates. The proposed approach represents a valuable contribution to the field of monocular image 6D pose estimation, offering a practical solution to overcome the challenges associated with slow matching speed.

He Zaixing [19] and his colleagues utilized special feature points, specifically the endpoints of straight contours, to accurately estimate the 6D pose based on geometric features. The method begins by matching the target object image with CAD templates using simple geometric features. It then employs specifically positioned points for precise matching. As a result, this method achieves high-precision pose estimation while employing a smaller number of templates. Additionally, the algorithm demonstrates excellent scalability as it can be combined with various geometric features and key points.

In recent years, the progress of artificial intelligence technology has led to remarkable research achievements in various fields, with deep learning playing a prominent role. Within the domain of 6DOF pose estimation, many studies have focused on harnessing the power of deep learning. Among the most direct approaches in these studies is the construction of end-to-end CNN models that aim to regress the 6D pose targets. Similarly, camera pose estimation, which shares similarities with pose estimation, involves leveraging CNN models to regress the camera pose. A pioneering work in this area is PoseNet [20], which directly estimates the camera pose using RGB images.

The objective of 6DOF pose estimation is to detect objects and estimate the rigid transformation parameters, encompassing translation and rotation, from the object coordinate system to the camera coordinate system. Algorithms such as PoseCNN [21], ConvPoseCNN [22], and SilhoNet [23] employ a separate prediction approach for estimating translation and rotation parameters, reducing the complexity of estimation. To address the challenge of insufficient depth information in RGB images, PoseCNN not only predicts the mapping position of the object center in the image but also incorporates an additional branch to estimate the depth of the object center. This enhancement improves the accuracy of the translation parameters. ConvPoseCNN improves upon the rotation parameter prediction branch of PoseCNN by utilizing dense (per-pixel) prediction. The adoption of dense prediction has become a prevalent trend for handling partial occlusion in target objects. SilhoNet extracts features from rendered images and regions of interest (RoI), combining them to construct two additional branches. These branches facilitate object segmentation and mask prediction for the

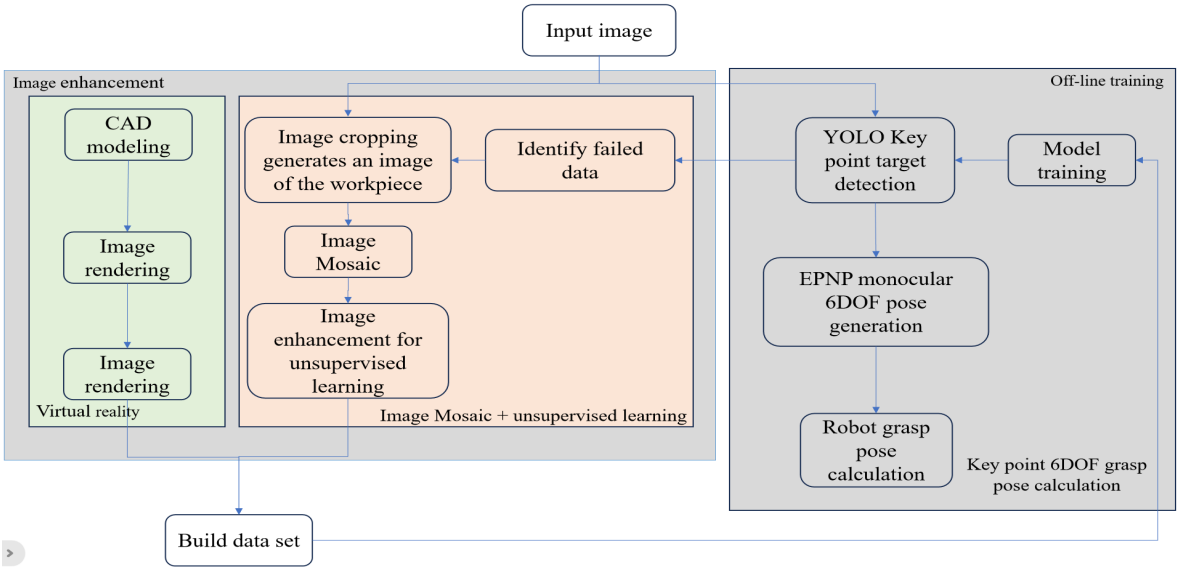

FIGURE 1. Workflow diagram of the monocular 6DOF robot localization and grasping strategy.

complete object structure, thereby enabling regression of the rotation parameters

III. PROPOSED METHOD

Virtual reality technology has the capability to generate a 3D model of the target workpiece using a virtual engine. By employing random algorithms, this technology enables the acquisition of randomly distributed images of the workpiece within the camera's field of view. This approach effectively addresses the challenge of acquiring images of small sample objects in visual systems used in industrial applications. Additionally, the generation of adversarial networks aids in eliminating the stitching effect caused by image splicing. This further enhances the detection effectiveness and accuracy of the target detection network, creating favorable conditions for achieving high-precision 6DOF pose detection of the target workpiece.

A monocular 6DOF robot visual guidance grasping strategy was developed, encompassing data augmentation and target detection. The workflow diagram illustrating this strategy is depicted in Figure 1.

The 6DOF positioning and grasping strategy can be divided into two main workflow components: dataset establishment and monocular 6DOF grasping pose calculation. Dataset establishment further comprises two parts: image augmentation using virtual reality technology and image stitching combined with unsupervised learning image augmentation. Virtual engines enable the direct creation and rendering of workpiece models, generating virtual images of the workpiece in diverse backgrounds and lighting conditions. This approach significantly reduces the cost of obtaining image samples for small sample objects in industrial settings. Additionally, this article explores the combination of image stitching algorithms and random generation algorithms to obtain stitched images of multiple workpieces placed in different positions and backgrounds. By utilizing an unsupervised algorithm-based image generation model to eliminate stitching artifacts, a pseudo ground truth map of the workpieces is obtained. This article leverages the aforementioned methods to acquire workpiece images, along with images collected by the camera, to form a workpiece sample dataset. This dataset serves for model training and validation in the subsequent robot pose grasping calculation process. In the grasping pose calculation section, a single-stage keypoint object detection algorithm is employed to detect multiple keypoints on the workpiece's surface in 2D images. Subsequently, the EPNP algorithm is utilized to calculate the 6DOF pose of the workpiece relative to the camera. Finally, in conjunction with the robot hand-eye calibration parameters, the grasping pose of the workpiece relative to the robot base coordinate system can be determined.

A. DATA GENERATION COMBINING VIRTUAL REALITY AND GENERATIVE ADVERSARIAL NETWORKS(GANS) DATA

The initial step in enabling a robot to locate and grasp target objects involves establishing a dataset. However, due to the high cost associated with acquiring image data for industrial objects, target detection for such objects often falls under the category of small-sample object detection. Data augmentation techniques play a crucial role in enriching small-sample data. In this paper, we propose a combination of virtual reality technology and generative adversarial networks to achieve data augmentation for industrial small-sample objects. The main methods employed are as follows:

1) DATA AUGMENTATION FOR INDUSTRIAL SMALL-SAMPLE OBJECTS BASED ON VIRTUAL REALITY TECHNOLOGY

Virtual reality technology is assuming an increasingly significant role in the industrial field. The method proposed in this paper involves the direct creation of small-sample images

of target objects using a virtual engine. By leveraging the rendering capabilities of the virtual engine, images of objects captured under various backgrounds and lighting conditions can be obtained.

Upon observing the image, it becomes evident that employing a virtual engine enables the generation of virtual images exhibiting diverse backgrounds, environments, and quantities.

FIGURE 2. Artifact image created by a virtual engine.

2) DATA AUGMENTATION OF TRADITIONAL IMAGE STITCHING AND GENERATIVE ADVERSARIAL NETWORKS

This paper employs image cropping, image stitching techniques, and random allocation algorithms to generate uniformly distributed images of objects in various positions. Enriching the image data mitigates the likelihood of deep neural networks becoming trapped in local minima during training.

Figure 3 depicts the process of data augmentation using traditional image processing methods and generative adversarial networks. Figure 3(a) showcases the object image of the workpiece captured directly by a monocular camera. Through image cropping, individual workpiece images are obtained, as illustrated in Figure 3(b). By employing a combination of random allocation algorithms and image stitching techniques, the resulting image shown in Figure 3(c) is obtained. Figure 3(d) represents the workpiece image generated by applying the generative adversarial network to the image in Figure 3(c). The integration of traditional image processing methods and generative adversarial networks enables the generation of randomly distributed images depicting different quantities of target workpieces within the visual range of the visual system. When compared to images generated solely using virtual reality technology, the optimized images produced by the generative adversarial networks exhibit a higher degree of similarity to real images. This makes them a valuable complement to virtual reality technology, offering enhanced realism and serving as an important augmentation technique.

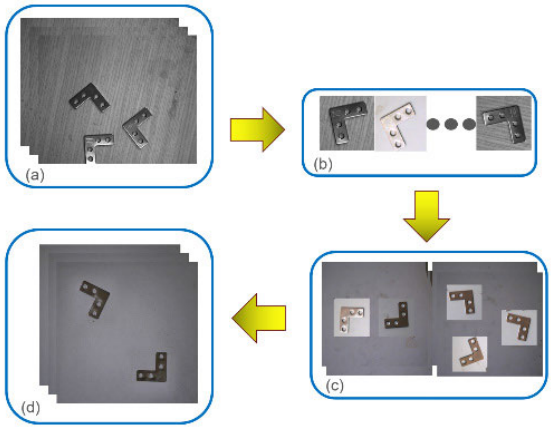

FIGURE 3. Image data generated by traditional image processing algorithms and generative adversarial networks.

B. DATA GENERATION BASED ON IMPROVED CYCLEGAN

Unsupervised learning methods are utilized in the proposed image generation technique to achieve more accurate simulation of target object images captured by cameras. This paper employs an enhanced version of CycleGAN, a deep learning model, to optimize the target object images generated using traditional image processing methods.

CycleGAN (Cycle-Consistent Adversarial Networks) [24] is a deep learning model specifically designed for image translation tasks. It facilitates the transformation of images from one domain to another without the need for paired training data.

Traditional image translation tasks typically necessitate paired image data, where both the source and target domain images are required for training. However, acquiring such paired data poses challenges in real-world scenarios. CycleGAN addresses this issue by leveraging adversarial networks and introducing cycle-consistency loss.

The primary objective of CycleGAN is to learn mappings between two domains, enabling the transformation of images from domain X to domain Y and vice versa. It comprises two generator networks and two discriminator networks. One generator is responsible for the X-to-Y domain transformation, while the other generator handles the reverse transformation.

CycleGAN's key concept revolves around enforcing cycle-consistency throughout the image translation process, ensuring that the translated image can be mapped back to the original domain. This is accomplished by incorporating two cycle-consistency losses, which minimize the pixel-level differences between the generated images and their corresponding original images.

The CycleGAN model comprises two mapping functions: G: X -> Y and F: Y -> X, along with corresponding adversarial discriminators Dy and Dx. Dy encourages G to translate X into images in the style of Y, and vice versa. To further regularize the mappings, the network incorporates two "cycle consistency loss functions" that ensure the style of

the transformed images can be reverted to their original state after inverse transformations, as depicted in Figure 4.

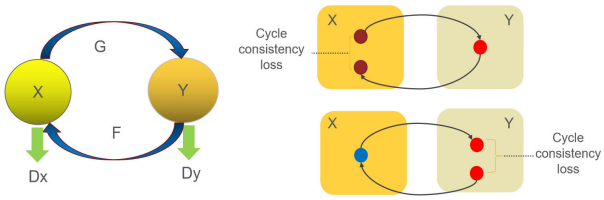

FIGURE 4. Working principle of the cycleGAN network.

To enhance the clarity of generated images and reduce the network's training time, this paper introduces the attention mechanism into the CycleGAN network and presents the Self-Attention CycleGAN network, in conjunction with the newly proposed contour loss function. The main enhancements offered by the proposed network are outlined below:

1) INTRODUCING THE GRADIENT LOSS FUNCTION

The target workpiece image generated by traditional image processing methods often displays pronounced grayscale differences when compared to the background image. To effectively mitigate these grayscale disparities in the target image and enhance the clarity of the generated image, this paper aims to introduce a gradient loss function into the generation network.

Figure 5 illustrates the comparison between the original image and the gradient image obtained from the camera-captured image and the image stitched using traditional methods. The comparison reveals noticeable rectangular boxes in the gradient image of the stitched image. To tackle this issue, this paper introduces an image gradient loss function, designed to address the grayscale disparities between the workpiece and the background in the stitched image while ensuring the clarity of the generated image. The formula for the gradient loss function is as follows:

$$LossT = |Grad(X) - Grad(Y)| \times \alpha \tag{1}$$

In the equation, X denotes the input image, Y represents the output image generated by the network, and represents the weight coefficient of the $LossT$.

After improvement, the loss function of the network is:

$$Loss = Loss_{cycle} + LossT \tag{2}$$

Among them, $Loss_{cycle}$ is the loss function of the original CycleGAN.

2) INTRODUCTION OF MULTI-CHANNEL FUSION ATTENTION MECHANISM

The attention mechanism is a widely employed technique in deep learning for enhancing feature extraction. It enables the model to concentrate on specific parts of the input data that are considered important while selectively disregarding irrelevant or less significant information. The primary objective of the attention mechanism is to enhance the model's capability to capture relevant context or features by assigning varying degrees of importance or weights to different parts of the input.

FIGURE 5. Comparison of edge features between captured images and stitched images.

The attention mechanism operates by computing attention weights for each element or segment of the input. These weights are typically determined through the interaction between the input and the model's learnable parameters. The higher the attention weight assigned to a particular element, the more attention or significance it receives during the model's processing.

Numerous studies have focused on attention mechanisms, with self-attention, spatial attention, and channel attention being commonly utilized in object detection. Experimental findings have demonstrated that attention modules can enhance the training efficiency and detection accuracy of networks [25]. The self-attention mechanism replaces traditional neighborhood computation by calculating correlations between all positions in the image. These correlations are then utilized as weights to represent the similarity between other positions and the currently calculated position. GcNet (Global-context Networks) represents a further development of the self-attention mechanism. Additionally, CBAM (Convolutional Block Attention Module) is a lightweight convolutional attention module that combines channel attention, which is invariant to spatial dimensions and compresses the channel dimension, with spatial attention, which is invariant to channel dimensions and compresses the spatial dimension. In reference [26], this paper proposes a hybrid attention mechanism that combines self-attention [27], spatial attention, and channel attention. The hybrid attention mechanism is applied to the generation network of CycleGAN to enhance the learning efficiency and generation performance of the network. Figure 6 illustrates the integration of the hybrid attention mechanism into the generation module of the CycleGAN network.

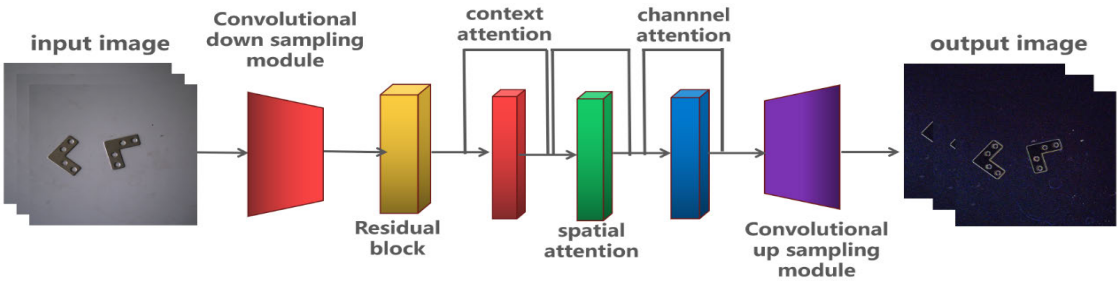

FIGURE 6. Schematic diagram of the cycleGAN generative network structure incorporating a hybrid attention mechanism.

C. INDUSTRIAL OBJECT KEY POINT DETECTION AND 6DOF POSE ESTIMATION BASED ON YOLOV7

This paper presents a method for determining the robot grasping pose of target objects through the utilization of a multi-keypoint with EPnP algorithm for 6DOF pose measurement. The proposed method leverages an enhanced version of the YOLOv7 algorithm for keypoint detection on the surface of the target object.

1) KEY POINT DETECTION ALGORITHM BASED ON YOLOV7

YOLOv7 represents a version of the YOLO series object detection algorithm, designed to strike a balance between model accuracy and inference performance. In a previous study [27], the YOLO algorithm was successfully employed for human pose estimation. In this paper, we extend this approach to the detection of key points on the surface of industrial objects.

Figure 8 illustrates the introduction of six branch detectors in this paper for keypoint detection on the output head of YOLOv7. During the detection of target objects, the proposed method predicts multiple key points on the object's surface.

In the YOLOv7 keypoint detection network, the GAM and SimAM techniques are employed to enhance the original network. Additionally, output headers for key points are incorporated into the network.

Most attention mechanisms overlook the importance of preserving both channel and spatial information in enhancing cross dimensional interactions. Paper [29] proposes a global attention mechanism to improve the performance of deep neural networks by reducing information diffusion and amplifying global interaction representations. The original author followed the sequential design of channel attention and spatial attention in CBAM and redesigned the sub modules. The overall structural design of GAM can be represented as Formula 3, where Mc and Ms represent the channel and spatial attention module, respectively, and $\otimes$ represent the multiplication of corresponding elements. We will add the GAM attention module to the two ELAN modules on the left side of Neck to enhance feature extraction capabilities.

$$\begin{aligned} F_2 &= M_c(F_1) \otimes F_1 \\ F_3 &= M_s(F_2) \otimes F_2 \end{aligned} \tag{3}$$

In response to the prevalent practice of generating 1-dimensional or 2-dimensional parameter weights from the input, paper [30] presents an attention mechanism capable of directly generating 3-dimensional parameter weights without the need for additional parameters. Figure 2(a) illustrates channel attention, where each parameter corresponds to a different channel. Figure 2(b) demonstrates spatial attention, with each parameter corresponding to the same position across all channels. Additionally, Figure 2(c) depicts the SimAM attention principle. In contrast to a simple one-dimensional or two-dimensional parameter structure, the utilization of three-dimensional parameter weights considers the refined features of different feature maps and different elements within the same feature map. Unlike existing channel/spatial attention modules, this module does not require additional parameters to derive 3D attention weights for feature maps. We have incorporated the SimAM parameterless attention module into the four ELAN modules on the right side of the Neck, resulting in a significant reduction in the total number of network parameters.

Furthermore, the network's output header is enriched with keypoint information. Consequently, the network can simultaneously output both the information parameters of the target detection box and the keypoint information.

The Intersection over Union (IoU) between predicted boxes and ground truth boxes serves as a crucial evaluation metric in object detection. YOLOv7 adopts the $CIOU$ as its loss function. In order to enhance the precision of keypoint prediction on the object's surface, this paper integrates the Geometry Intersection over Union ($GeIOU$) into YOLOv7's loss function, aiming to further improve the accuracy of keypoint detection within the network.

Based on Figure 10, the $GeIOU$ method calculates the minimum bounding polygon formed by the ground truth and predicted key points, and subsequently computes the IOU based on this polygon. The $GeIOU$ is further enhanced by employing the $DIOU$ function. The corresponding formula is presented below:

$$L_{GeIOU} = 1 - IOU + \frac{\rho^2(A, B)}{c^2} \tag{4}$$

IOU represents the intersection over the union of the ground truth area and the predicted area of key points. ρ

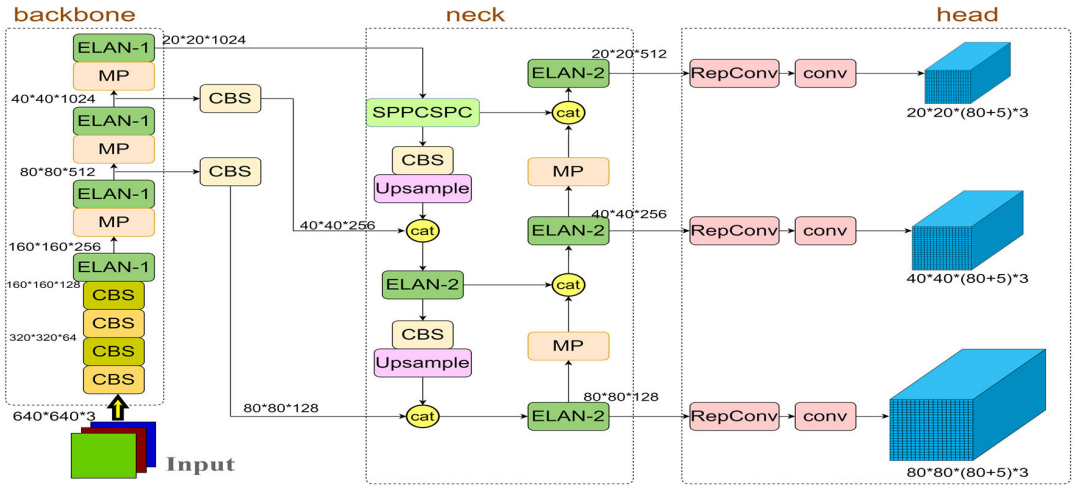

FIGURE 7. Schematic diagram of the YOLOv7 network structure.

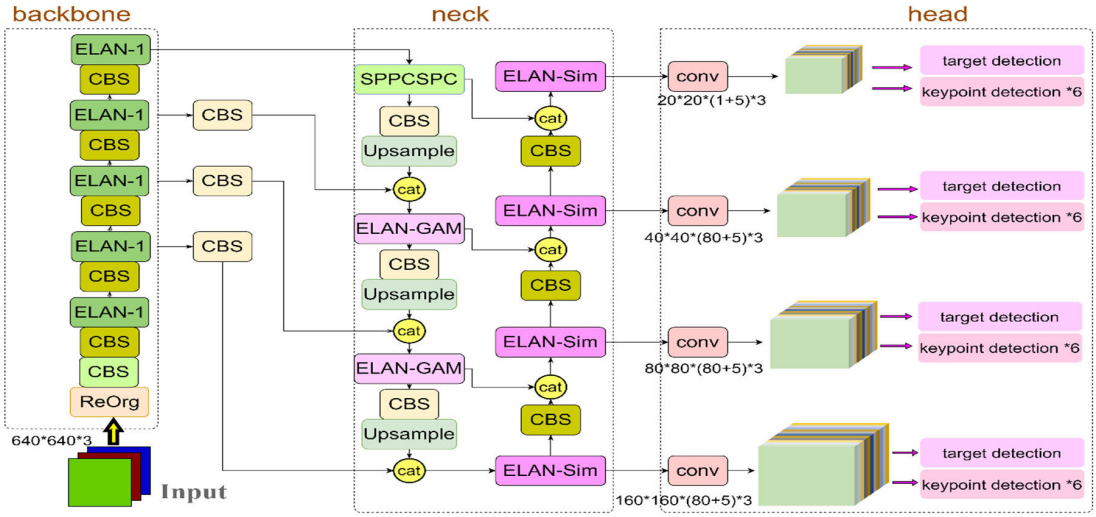

FIGURE 8. Schematic diagram of the YOLOv7 keypoint detection network structure.

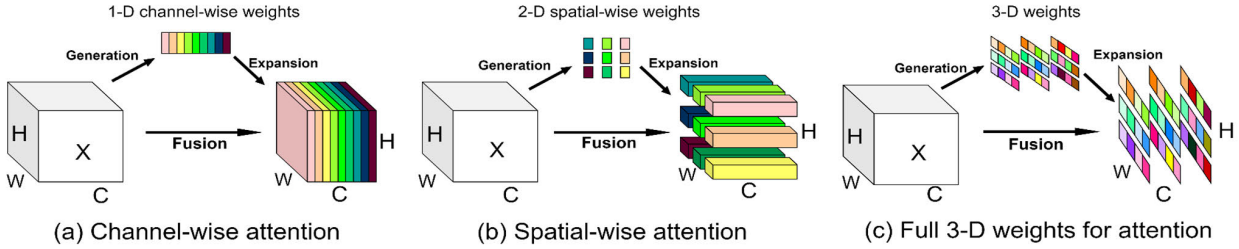

FIGURE 9. SimAM principle.

denotes the Euclidean distance between the predicted A and ground truth center coordinates B, while c represents the diagonal distance of the minimum bounding box that encompasses them.

2) 6DOF POSE MEASUREMENT BASED ON EPNP

The EPnP algorithm, proposed by Lepetie and Moreno [10] in 2009, is a highly accurate and efficient pose estimation algorithm. This algorithm eliminates the need for iterative solving, resulting in a time complexity of $O(n)$. It exhibits robustness and requires a minimal number of 3D-2D matching point pairs, typically three pairs for coplanar cases or four pairs for non-coplanar cases, to achieve precise pose estimation. By leveraging the object's pose and the hand-eye relationship between the robot and the vision system, the EPnP algorithm enables the calculation of the robot's grasping pose.

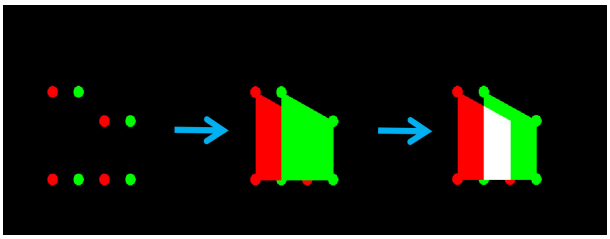

FIGURE 10. GeIOU principle.

When an industrial robot grasps an object, we have:

$$^{U}T_{P}^{1} = {}^{U}T_{T}^{1} * {}^{T}T_{P} \tag{5}$$

where ${}^{U}T_{T}^{1}$ is the grasping pose of the industrial robot. And ${}^{T}T_{P}$ is the pose of the object in the robot's gripper coordinate system.

When an industrial robot takes a picture of an object, we have:

$$^{U}T_{P} = {}^{U}T_{CT}^{2} \times {}^{CT}T_{C} \times {}^{C}T_{P} \tag{6}$$

where ${}^{U}T_{CT}^{2}$ is the robot's pose at the time of capturing the image by the vision system. ${}^{CT}T_{C}$ has been obtained through hand-eye calibration. ${}^{C}T_{P}$ is the pose of the object relative to the vision system, obtained directly from the vision system. By using equations (5) and (6), we can solve for:

$$^{T}T_{P} = [{}^{U}T_{T}]^{-1} \times {}^{U}T_{CT}^{2} \times {}^{CT}T_{C} \times {}^{C}T_{P} \tag{7}$$

Therefore, when an industrial robot takes pictures of objects in different positions, we have:

$$^{U}T_{T}^{n} = {}^{U}T_{CT}^{n} \times {}^{CT}T_{C} \times {}^{C}T_{P}^{n} \times [{}^{T}T_{P}]^{-1} \tag{8}$$

${}^{C}T_{P}^{n}$ and ${}^{U}T_{CT}^{n}$ are the coordinates of the object detected by the vision system, and ${}^{U}T_{T}^{n}$ represents the grasping pose of the target object in the robot's coordinate system.

IV. VALIDATION EXPERIMENTS

A. EXPERIMENTAL ANALYSIS OF CREATING VIRTUAL IMAGES WITH VIRTUAL ENGINE

The Blender virtual engine is utilized to generate 3D models of multiple artifacts, which are then placed in diverse backgrounds. Subsequently, a virtual image of the same dimensions as the one captured by a monocular camera is generated. Figure 11 showcases various virtual images of objects created using the virtual engine. The figure demonstrates that these virtual images can effectively simulate object images captured by the visual system under different lighting conditions, materials, and background settings. By incorporating these virtual images, the method proposed in this paper aims to address the challenges associated with acquiring a limited number of sample images of industrial objects. These virtual images serve as a valuable supplement to the image dataset.

B. EXPERIMENTAL ANALYSIS OF IMAGE STITCHING + GENERATIVE ADVERSARIAL NETWORKS

Figure 12 illustrates a comparison among different generative adversarial networks (GANs) for image generation and image stitching. The first row displays the original images of the generated artifacts obtained through cropping. From the figure, it is evident that there is a noticeable contrast in grayscale values between the artifact region and the background image.

The second row presents the images directly generated by the original CycleGAN. Upon comparing these images with the ones in the first row, it can be observed that the pixel differences between the artifact image and the background image are not effectively eliminated in the images generated by CycleGAN for the same dataset.

The third row showcases the images generated by CycleGAN with the incorporation of SeNet attention modules. Comparing these images with the ones generated by the original CycleGAN, it can be concluded that the SeNet attention modules have limited improvement in the capabilities of CycleGAN.

The fourth row displays the images of the generated artifacts using the proposed gradient loss function in this paper. By comparing these images with the previous ones, it can be observed that the integration of the gradient loss function enhances the optimization effect of CycleGAN for the target artifacts. This demonstrates the effectiveness of the proposed gradient loss function. However, the generated images exhibit the presence of artifact-like structures in regions where no artifacts originally exist, and there is distortion in the restoration of local details of the target artifacts.

The fifth row represents the images generated by the CycleGAN+CBMA module network. It can be observed that after incorporating the CBMA network, the regression of image details is relatively good. However, similar to the case with the gradient loss function, there is a significant amount of interference introduced with metal-like features.

The sixth row displays the images generated by the network proposed in this paper. It can be seen that both the degree of restoration of the target artifact's details and the introduction of impurities are optimized with the method proposed in this paper.

Figure 13 offers a more intuitive demonstration of the generated results based on the improved CycleGAN images. The upper part of the figure displays the stitched images, while the lower part shows the images generated by the proposed method. From the figure, it is evident that the proposed method effectively eliminates the background contours and grayscale differences of the stitched artifacts. This results in image quality that is similar to the original images captured directly by the visual system. Consequently, we assert that the proposed method achieves a commendable restoration effect for industrial objects.

Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) are commonly used metrics for evaluating image quality and similarity. These metrics assess blurriness and

FIGURE 11. Virtual images of the workpiece created at different positions and with different backgrounds by the virtual engine.

FIGURE 12. Comparison of cropped images generated by different generative adversarial networks.

similarity between two images using fixed formulas. In this study, we compare the PSNR and SSIM values between the original images captured by the visual system and the images generated by the improved cycleGAN network. The comparison results are presented in Table 1.

The PSNR metric typically ranges from 0 to 100, while the SSIM metric ranges from 0 to 1. Higher values for both metrics indicate better performance. By comparing the PSNR and SSIM values, it is evident that the proposed method achieves results that closely resemble the original images. Analysis suggests that the original images, captured by the visual system in an industrial environment, exhibit pronounced noise due to factors such as the background and ambient lighting. As a result, the PSNR values in Table 1 may

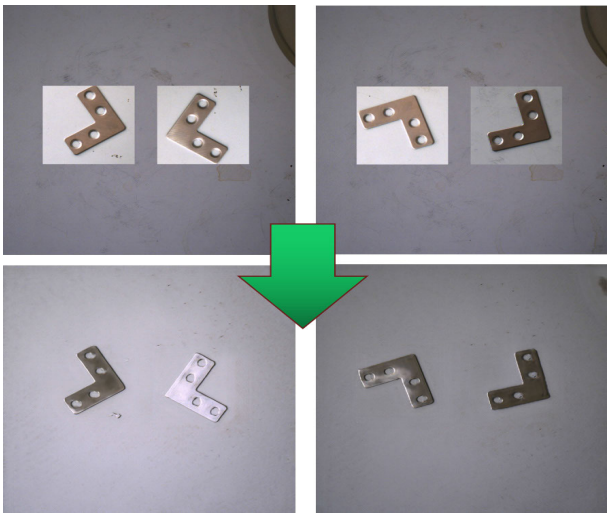

FIGURE 13. Comparison of the effects before and after using generative adversarial networks.

TABLE 1. Comparison of PSNR and SSIM parameters.

	Original image	Our method	cycleGAN	cycleGAN+CBMA
PSNR	15.64	14.21	12.89	14.07
SSIM	0.613	0.56	0.54	0.58

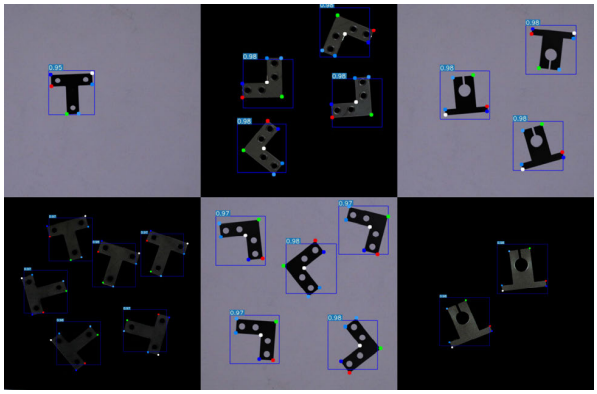

FIGURE 14. Detection result.

be relatively low in several scenarios. However, the PSNR and SSIM values of the various methods closely match those of the original images, indicating a high degree of similarity between the images generated using the proposed image generation method and the images captured by the visual system. This validation further reinforces the rationale behind the proposed method in this study.

C. KEY POINT DETECTION NETWORK EXPERIMENTAL ANALYSIS

The computer configuration used for testing in this experiment includes an Intel i7-10700 CPU and an NVIDIA GeForce RTX 3060 GPU. A Haikang industrial camera with a resolution of 1280×1024 is employed to capture real-world images. The Blender engine is utilized to create 3D models

TABLE 2. Yolov7-improve experimental results.

		σ		
OBJECT-1	Method	real + virtual	real	virtual
	Yolov3	1.662	2.554	1.361
	Yolov4	1.72	2.324	1.687
	Yolov5	2.273	2.3329	1.127
	Yolov7	1.431	2.142	1.058
	Yolov7-tiny	1.917	3.841	1.307
	Yolov7-w6	1.709	2.488	0.938
	Yolov7-improve	1.249	2.4	0.9056
		σ		
OBJECT-2	Method	real + virtual	real	virtual
	Yolov3	2.599	2.184	1.53
	Yolov4	1.957	1.732	1.575
	Yolov5	1.842	3.217	1.3
	Yolov7	1.813	1.66	1.521
	Yolov7-tiny	2.332	2.787	2.433
	Yolov7-w6	1.598	1.63	1.412
	Yolov7-improve	1.335	1.453	1.35
		σ		
OBJECT-3	Method	real + virtual	real	virtual
	Yolov3	1.24	1.374	1.283
	Yolov4	1.664	1.2197	1.394
	Yolov5	1.417	6.246	1.457
	Yolov7	1.26	1.058	1.138
	Yolov7-tiny	1.478	2.094	1.319
	Yolov7-w6	1.149	1.057	1.129
	Yolov7-improve	1.057	1.019	1.081

FIGURE 15. Robot system platform.

of various artifacts.The dataset utilized in this experiment is divided into a training set and a test set. Both sets comprise three different shapes of metal objects. The training set for each object consists of 1200 real images captured by a vision system, as well as virtual object images. On the other hand, the test set consists of 60 real images.

To evaluate keypoint detection, the primary focus is on measuring the deviation between predicted key points and real key points. The accuracy of keypoint detection is determined by calculating the positional error between the predicted and real key points. In this paper, the keypoint

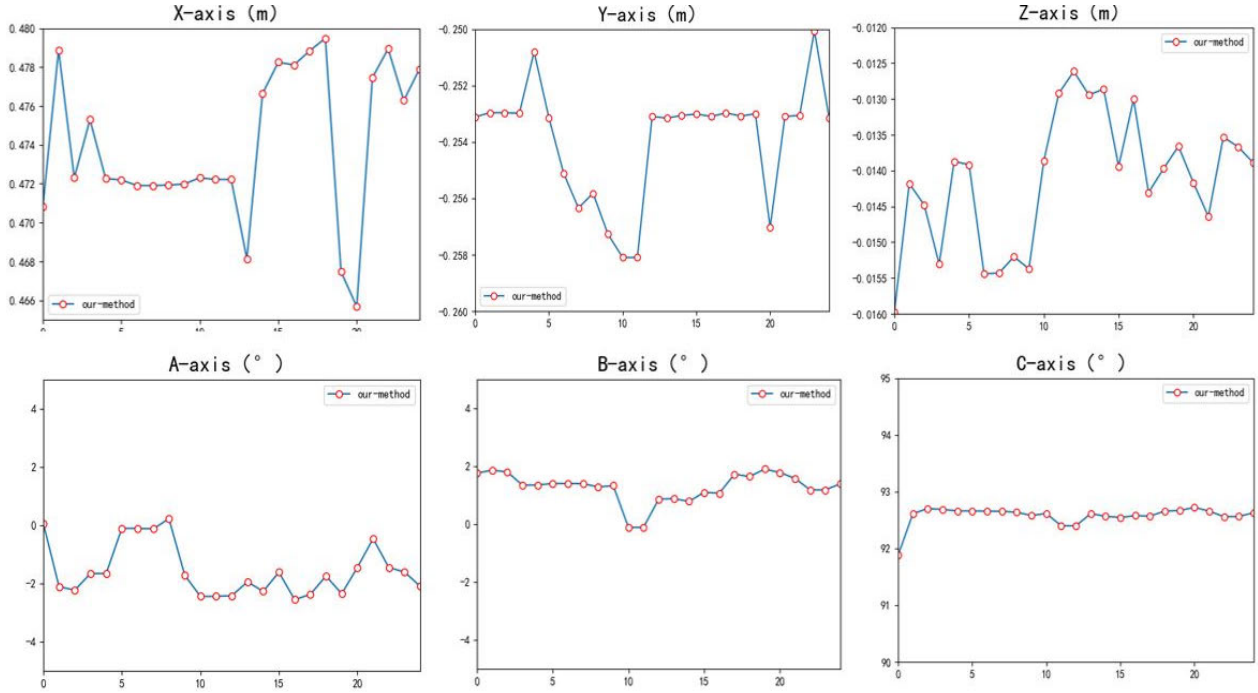

FIGURE 16. Measurement accuracy of the proposed method.

detection error σ is expressed as:

$$\sigma = \frac{\sum_{i=0}^{m}\left(\frac{\sum_{j=0}^{n}\sqrt{(x_j-x)^2+(y_j-j)^2}}{n}\right)}{m} \tag{9}$$

In the equation, x_j and y_j represent the coordinates of the predicted points, while x and y represent the coordinates of the true key points. By calculating the Euclidean distance between the predicted points and the true points as $\sqrt{(x_j-x)^2+(y_j-y)^2}$, we obtain the error distance between the predicted key points and the true key points. The n represents the number of key points in a single image, while m represents the number of test images. The evaluation metric in this study, denoted by σ, represents the average distance error of all key points in the test images.

This study performs a comparative analysis utilizing real images, virtual images, and mixed image datasets. The experimental results are presented in Table 2.

Based on the comparative analysis of the experimental results, several key observations can be made. Firstly, the detection error is highest for real images, followed by real+virtual images, and lowest for virtual images in all three scenarios. This can be attributed to the fact that virtual images, being generated by a virtual engine, exhibit consistent object sizes, leading to improved detection performance.Furthermore, the detection error for real images is higher compared to real+virtual images. This can be attributed to the augmentation of the dataset with virtual images, which allows the detection model to better learn the geometric features of the objects. As a result, the model demonstrates improved monitoring performance when trained on the augmented dataset.

In terms of the detection algorithm, the improved Yolov7 algorithm outperforms the other comparative algorithms in terms of keypoint detection error. This suggests that the proposed method in this study is effective in enhancing the detection performance of objects. To provide a more comprehensive analysis, it would be helpful to have additional details such as the specific values or metrics reported in the experimental results and the corresponding figures in Figure 14.

D. ROBOT GRASPING POSE EXPERIMENTAL ANALYSIS

To further validate the pose measurement accuracy of the proposed algorithm, this paper utilizes an industrial robot grasping platform. The purpose is to assess the algorithm's performance in accurately measuring the pose of objects. The robot grasping platform used in the experiment is depicted in Figure 15.

The robot grasping platform in this study utilizes an MZ07 type 6-joint industrial robot manufactured by Nachi Robotics. This robot has a payload capacity of 7 kg and a repeat positioning accuracy of ±0.03 mm, ensuring precise and accurate movements. To facilitate image acquisition, an MV-CA013-20GC camera is mounted on the robot's end flange. This camera is responsible for capturing images of the objects for further analysis and pose measurement. The parameters for robot hand eye calibration in the camera are shown in Tables 3 and 4

TABLE 3. The parameters of monocular vision.

f/mm	k	Sx	Sy	Cx	Cy	W	H
0.012075	-1673.43	4.8e-06	4.8e-06	656.678	556.742	1280	1024

TABLE 4. Hand-eye calibration parameters.

Average		Max	
Translation error (mm)	Rotational error (°)	Translation error (mm)	Rotational error (°)
1.216	0.133	2.243	0.265

During the testing process, the workpiece remains in a fixed position while the robot is moved to various poses to capture images of the workpiece. The goal is to calculate the pose of the workpiece in the robot's base coordinate system. Ideally, regardless of the robot's pose during image capture, the calculated pose of the workpiece in the robot's base coordinate system should remain the same. However, due to assembly, visual system, and robot errors, the measured values obtained by the system may not be identical. In this paper, the pose measurement accuracy of the visual system is analyzed based on these errors.

Figure 16 presents the results of the robot system's 25 measurements of a fixed-position workpiece. From the figure, it can be observed that the system's measurement errors for the X-axis direction of the workpiece are within ±7 mm, ±4 mm for the Y-axis direction, ±2.5 mm for the Z-axis direction, ±1.5° for the rotation axis A in the X-axis direction, and ±0.5° for the rotation axis C in the Z-axis direction. Considering the measurement field of view of the camera at 166 × 136 mm, the measurement accuracy of the system is calculated to be 4.21% in the X-axis direction and 2.94% in the Y-axis direction. Additionally, based on calibration, it is determined that the camera height is 377.72 mm, and the measurement error percentage is 0.39%. The experimental results demonstrate that the proposed method in this paper maintains a high detection accuracy while ensuring system robustness.

V. CONCLUSION

This paper introduces a novel method for 6DOF pose measurement of robots based on monocular vision. The proposed method leverages image generation techniques and combines generative adversarial networks with attention mechanisms to achieve data augmentation for industrial objects. This approach provides a new solution for augmenting data of industrial small-sample objects, which is beneficial when training pose measurement models.

The method employs an improved keypoint detection network that enables accurate 6DOF pose measurement of the target workpiece. It achieves robust measurement by detecting multiple key points on the surface of the object. This strategy has significant implications for expanding the application range of 6DOF robot systems in the industrial field, reducing visual measurement costs, and enhancing the intelligence of robot systems.

The experimental results demonstrate that the proposed method exhibits high detection accuracy and robustness. Moving forward, the authors plan to explore methods specifically designed for complex environments to further enhance the accuracy of 6DOF pose detection in monocular vision systems for target workpieces.

REFERENCES

[1] G. Wan, F. Li, B. Liu, S. Bai, G. Wang, and K. Xing, "A novel robotic 6DOF pose measurement strategy for large-size casts based on stereo vision," *Assem. Autom.*, vol. 42, no. 4, pp. 458–473, Jul. 2022.
[2] D. Shin, "Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm," *J. Inf. Sci.*, vol. 49, no. 1, pp. 18–31, Feb. 2023.
[3] G. Wan, F. Li, W. Zhu, and G. Wang, "High-precision six-degree-of-freedom pose measurement and grasping system for large-size object based on binocular vision," *Sensor Rev.*, vol. 40, no. 1, pp. 71–80, Jan. 2020.
[4] G. Wan, G. Wang, K. Xing, Y. Fan, and T. Yi, "Robot visual measurement and grasping strategy for roughcastings," *Int. J. Adv. Robotic Syst.*, vol. 18, no. 2, pp. 715–720, 2021.
[5] R. Ding, M. Cheng, Z. Han, F. Wang, and B. Xu, "Human-machine interface for a master-slave hydraulic manipulator with vision enhancement and auditory feedback," *Autom. Construct.*, vol. 136, Apr. 2022, Art. no. 104145.
[6] Z. Zhou, L. Li, A. Fursterling, H. J. Durocher, J. Mouridsen, and X. Zhang, "Learning-based object detection and localization for a mobile robot manipulator in SME production," *Robot. Comput.-Integr. Manuf.*, vol. 73, Feb. 2022, Art. no. 102229.
[7] Y. Ma, K. Du, D. Zhou, J. Zhang, X. Liu, and D. Xu, "Automatic precision robot assembly system with microscopic vision and force sensor," *Int. J. Adv. Robotic Syst.*, vol. 16, no. 3, May 2019, Art. no. 172988141985161.
[8] M. Ulrich, C. Wiedemann, and C. Steger, "Combining scale-space and similarity-based aspect graphs for fast 3D object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1902–1914, Oct. 2012.
[9] C. Wu, L. Chen, Z. He, and J. Jiang, "Pseudo-Siamese graph matching network for textureless objects' 6-D pose estimation," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2718–2727, Mar. 2022.
[10] W. Zou, D. Wu, S. Tian, C. Xiang, X. Li, and L. Zhang, "End-to-end 6DoF pose estimation from monocular RGB images," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 87–96, Feb. 2021.
[11] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1530–1538.
[12] J. Kang, W. Liu, W. Tu, and L. Yang, "YOLO-6D+: Single shot 6D pose estimation using privileged silhouette information," in *Proc. Int. Conf. Image Process. Robot. (ICIP)*, Negombo, Sri Lanka, Mar. 2020, pp. 1–6.
[13] Y. Zhang, Y. Zhou, H. Pan, B. Wu, and G. Sun, "Visual fault detection of multi-scale key components in freight trains," 2022, *arXiv:2211.14522*.
[14] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.
[15] Z. Cao, Y. Sheikh, and N. K. Banerjee, "Real-time scalable 6DOF pose estimation for textureless objects," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2441–2448.
[16] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab, "Dominant orientation templates for real-time detection of texture-less objects," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 13–18.
[17] A. Crivellaro, M. Rad, Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit, "A novel representation of parts for accurate 3D object detection and tracking in monocular images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4391–4399.

[18] Y. Konishi et al., "Fast 6D pose estimation from a monocular image using hierarchical pose trees," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 9905, 2016, pp. 398–413.
[19] Z. He, Z. Jiang, X. Zhao, S. Zhang, and C. Wu, "Sparse template-based 6-D pose estimation of metal parts using a monocular camera," *IEEE Trans. Ind. Electron.*, vol. 67, no. 1, pp. 390–401, Jan. 2020.
[20] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2938–2946.
[21] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*.
[22] C. Capellen, M. Schwarz, and S. Behnke, "ConvPoseCNN: Dense convolutional 6D object pose estimation," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020.
[23] G. Billings and M. Johnson-Roberson, "SilhoNet: An RGB method for 6D object pose estimation," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3727–3734, Oct. 2019.
[24] S. Niu, B. Li, X. Wang, S. He, and Y. Peng, "Defect attention template generation cycleGAN for weakly supervised surface defect segmentation," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108396.
[25] J. Liang, F. Xu, and S. Yu, "A multi-scale semantic attention representation for multi-label image recognition with graph networks," *Neurocomputing*, vol. 491, pp. 14–23, Jun. 2022.
[26] J. Zhang, L. Xing, Z. Tan, H. Wang, and K. Wang, "Multi-head attention fusion networks for multi-modal speech emotion recognition," *Comput. Ind. Eng.*, vol. 168, Jun. 2022, Art. no. 108078.
[27] J. E. Arco, A. Ortiz, N. J. Gallego-Molina, J. M. Górriz, and J. Ramírez, "Enhancing multimodal patterns in neuroimaging by Siamese neural networks with self-attention mechanism," *Int. J. Neural Syst.*, vol. 33, no. 4, Apr. 2023, Art. no. 2350019, doi: 10.1142/s0129065723500193.
[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7464–7475.
[29] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.
[30] L. Yang, R. Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

GUOYANG WAN received the Ph.D. degree from Dalian Maritime University, in 2021. He is currently a Lecturer with Anhui Polytechnic University. He has published several SCI and EI retrieval papers. His main research interests include machine vision and robot vision guidance.

JINCHENG CHEN received the bachelor's degree from Huizhou University. He is currently pursuing the master's degree with Anhui Polytechnic University. His research interests include deep learning and defect detection.

JIAN ZHANG received the B.Sc. degree from the Anhui Institute of Information Technology, in 2021. He is currently pursuing the M.Sc. degree with Anhui Polytechnic University. His main research interests include object detection and deep learning.

BINYOU LIU received the Ph.D. degree from the China University of Technology. He is currently a Professor with Anhui Polytechnic University. He has published several SCI and EI retrieval papers and won many scientific and technological awards. His main research interests include advanced control theory and application and high-speed image tracking.

HONG ZHANG received the master's degree from the Hefei University of Technology, in 2003. She is currently an Associate Professor with the School of Electrical Engineering, Anhui Polytechnic University. She has published more than ten papers. Her research interests include electronic measurement and sensor technology.

XIUWEN TAO is currently pursuing the master's degree with Anhui Polytechnic University. She has published an EI paper. Her research interest includes digital image processing.