JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

A Unified Multi-Task Deep Learning Framework for Robotic Bin-Picking of Planar Objects

The-Thinh Pham[1], Tuan-Khanh Nguyen[2*], Chi-Cuong Tran[3], Quang-Huan Dong[2], Duy-Tan Do[4], Hoang-Vinh-Khang Nguyen[2], Quang-Chien Nguyen[4]

[1]*Can Tho University of Technology, Vietnam*
[2]*Vietnamese-German University, Vietnam*
[3]*National Taiwan University of Science and Technology, Taiwan*
[4]*Ho Chi Minh University of Technology and Engineering, Vietnam*

[*]Corresponding author. Email: khanh.nt@vgu.edu.vn

ARTICLE INFO

Received: 04/05/2026
Revised: 30/06/2026
Accepted: 17/07/2026
Online First: 19/08/2026
Published:

KEYWORDS

Robotic bin-picking;
Pose estimation;
Planar object;
Digital twin;
Deep learning.

ABSTRACT

Automating random bin-picking in industrial robotics, where robots handle diverse and cluttered objects, remains challenging due to the complexity of object detection and pose estimation. While many solutions focus on free-form objects, systems specifically designed for planar objects are lacking. Planar objects pose unique challenges, as the commonly used Point Pair Feature (PPF) approach for free-form objects is ineffective due to their lack of distinctive geometric features. In this study, the proposed framework was implemented and evaluated using USB packs as a representative planar object case study. An innovative approach is introduced for the random bin-picking of planar objects by developing a multi-task model for instance segmentation and keypoint detection in 2D images. A Geometric approach is then employed to estimate the 6D object pose for robotic grasping. Furthermore, a grasp candidate selection strategy is proposed to enable reliable grasping in cluttered industrial environments. Experimental results show that the proposed method achieved mAP50 values of 0.954, 0.800, and 0.926 for bounding box detection, instance segmentation, and keypoint detection, respectively, with a processing time of 2.7 ms. Future work will focus on integrating the framework into a digital twin system to support real-time monitoring, simulation, and optimization of automated manufacturing processes.

DOI: https://doi.org/10.54644/jte.2026.2401

1. Introduction

Automated object manipulation using industrial robots is now a well-established field, focused on developing smarter and more adaptable systems [1]. A key challenge in production lines is random bin-picking, where robots must handle diverse object shapes in cluttered environments to replace manual labor. This task requires systems capable of detecting objects within a scene and accurately estimating their poses [2]. For effective operation, a robust system must consistently interpret its surroundings and manage uncertainties [3]. When objects are randomly placed and partially occluded, high reliability becomes essential.

Despite its critical role in manufacturing, random bin-picking remains a complex problem in image processing and automation, with current solutions often failing to meet the industry's stringent requirements for speed and reliability [4]. Consequently, ongoing research is needed to develop advanced algorithms and comprehensive systems capable of efficiently addressing random bin-picking tasks. Although advancements in robotics have improved precision, many practical challenges persist [5]. After objects are identified in 3D space, accurately grasping them in unstructured environments remains a significant difficulty. Pose estimation plays a crucial role in this process, as it provides essential information about the position and orientation of objects.

Research on pose estimation generally falls into three main categories: feature-based methods, pattern matching approaches, and deep learning techniques. Feature-based methods utilizing 3D data

provide comprehensive solutions for object identification [6], while pattern matching approaches use RGB or RGB-D data to extend 2D information for pose estimation. In addition, supervised machine learning [7], [8] and deep learning models [9] have been widely applied to enhance object recognition and pose estimation performance.

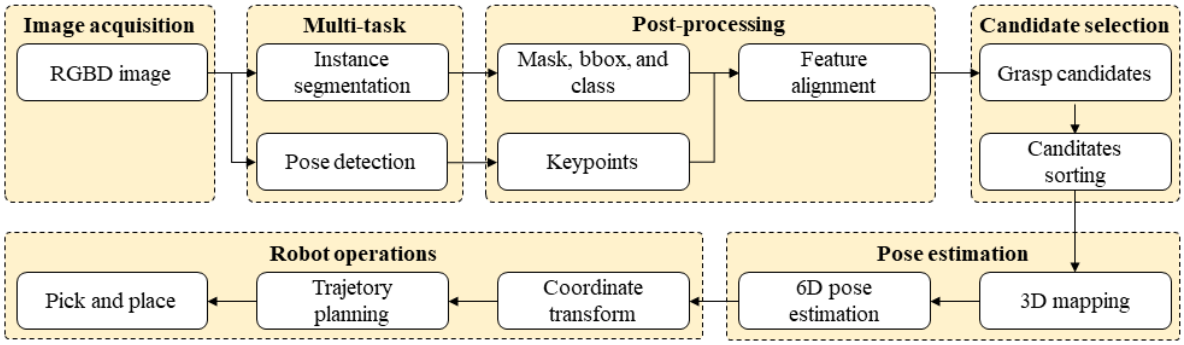

Figure 1. *The overall structure of the system.*

In computer vision, RGB-D data are commonly used in combination with machine learning techniques to improve object recognition and pose estimation. For example, [10] introduced an approach that utilizes RGB-D images with convolutional k-means descriptors for object recognition. In [11], [12] developed a random forest-based method to estimate poses for both textured and texture less objects and later refined the approach to operate using only a single RGB image. Similarly, [13] proposed an instance segmentation network with a dedicated pose estimation branch capable of directly predicting 6D object poses. Progress has also been made in complete bin-picking systems, particularly those based on CAD models, which have demonstrated high success rates in object detection and grasping. In [3], for example, the authors designed a CAD-based system that performed effectively in cluttered environments. Other systems have incorporated RGB-D cameras to enhance detection and pose estimation, such as [14], who employed Kinect sensors with high accuracy, and [15], who utilized multiple cameras in a CAD-based multi-view framework to improve pose estimation.

Recent studies have also explored integrated perception and multi-task learning for robotic grasping and bin-picking. For instance, [16] proposed Fast GraspNeXt, a multi-task network for robotic grasping on edge devices, while [17] introduced MetaGraspNetV2, a bin-picking dataset with annotations for object detection, instance segmentation, 6-DoF pose estimation, object relationships, and grasp detection. Other recent works have focused on related subtasks, including instance-segmentation-based 6D pose estimation, point-cloud-based pose estimation, and category-agnostic segmentation. However, recent studies that directly integrate object detection, instance segmentation, and keypoint-based pose estimation into a unified multi-task framework for robotic bin-picking remain limited. This gap motivates the present study.

While many existing methods focus on free-form objects, a notable gap remains in the literature regarding systems specifically designed for planar objects [18]–[21]. In this study, USB packs are used as a representative planar object case study. Despite their simpler geometries, planar objects present unique challenges; for instance, the commonly used PPF approach is often ineffective due to the lack of distinctive geometric features on flat surfaces. To address these limitations, this study proposes a novel, integrated multi-task architecture optimized for the random bin-picking of planar objects. By unifying object detection, instance segmentation, and pose estimation within a shared backbone and neck framework, the proposed network significantly reduces redundant computations and minimizes inference latency.

The methodology leverages advanced deep learning techniques to categorize and localize objects using 2D image data, generating individual masks and keypoints for each USB pack. Furthermore, the 2D data of the object obtained from the multi-task AI model allows for the construction of an accurate coordinate system that matches the USB pack plane. This approach ensures high precision across all three tasks while maintaining the efficiency required for real-time robotic operations.

2. Proposed Methods

This section describes the proposed solution for random bin-picking of planar objects. The overall structure of the proposal system is shown in Figure 1. It can be seen that the system is divided into five main parts: data preparation and processing, multi-task deep learning model development, candidate selection, pose estimation, and robot operations.

2.1. Data preparation and processing

The proposed system used USB packs to create a representative dataset of planar objects. High-quality 2D images of these objects were collected as foundational input, enabling the training of deep learning models designed to recognize and manipulate planar objects in complex, cluttered environments. The data annotation plays an important role for supervised learning models. In this study, each image was annotated separately with segmentation masks and keypoints using Roboflow. The annotations were then converted into a unified format consisting of the class label, bounding box, keypoints, and segmentation points, as shown in Figure 2. Each object target includes one segmentation mask, one bounding box, and four keypoints. A total of 172 images were collected and divided into training and validation sets using an 8:2 ratio. To alleviate the limitations of the small dataset and reduce the time-consuming process of manual data collection, data augmentation techniques, including contrast adjustment, brightness adjustment, exposure adjustment, noise, and scaling, were applied. As a result, the dataset was expanded to 860 images.

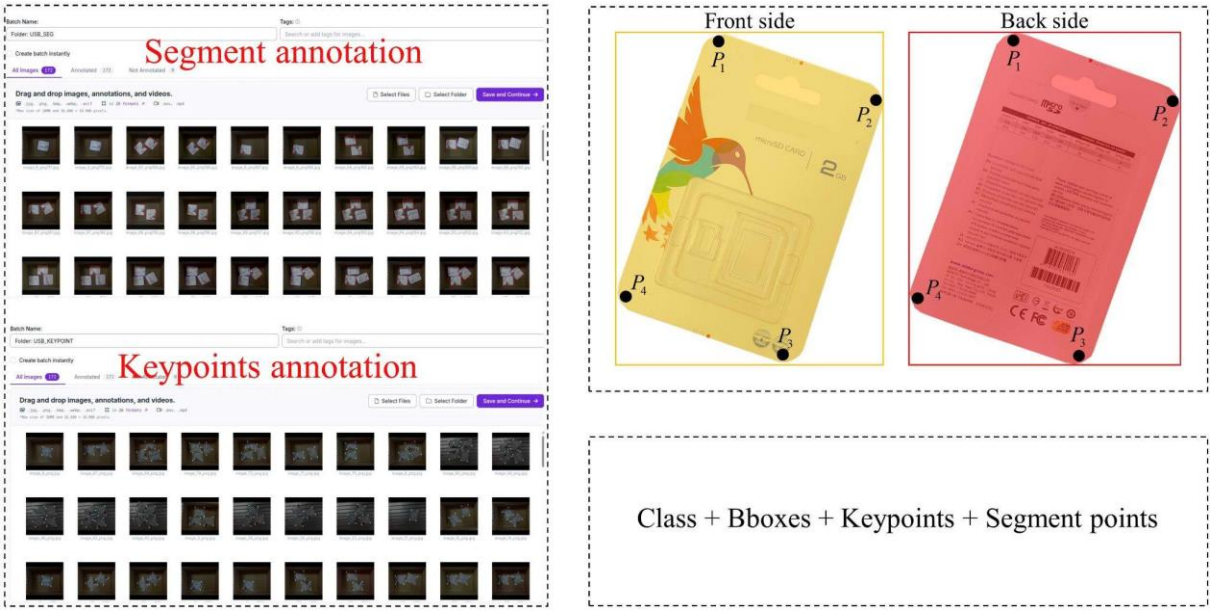

Figure 2. *Label for instance mask and keypoint location.*

2.2. Multi-task deep learning model development

In this study, a standard single-task deep learning architecture is extensively modified into a specialized multi-task framework to meet the rigorous demands of industrial bin-picking. This design builds on the fundamental strengths of conventional models while integrating instance segmentation and keypoint detection into a unified pipeline based on the YOLOv8n architecture [22]. By extending the standard single-target detection logic, the proposed model can simultaneously isolate individual object instances and identify key structural points - a capability essential for managing cluttered industrial scenes where objects, such as planar USB packs, are often overlapping. Following the multi-task approach, the architecture ensures that high-level feature representations are shared across detection, segmentation, and pose estimation heads, resulting in a robust system tailored for the precise localization and categorization of planar geometries as illustrated in Figure 3.

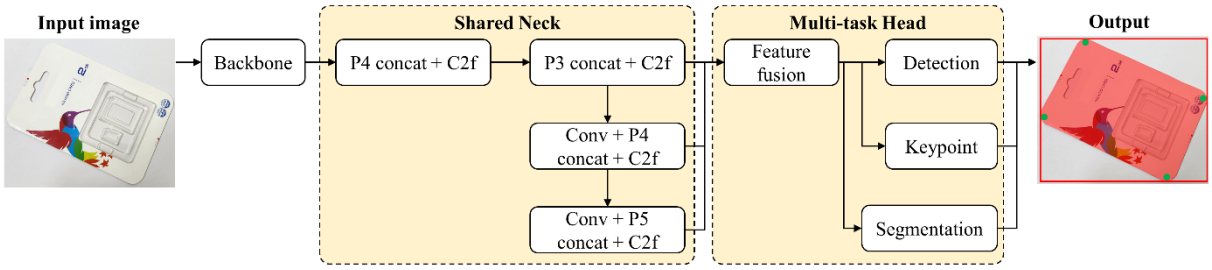

Figure 3. *Multi-task deep learning model implementation.*

Upon finalizing the model architecture, the training process is conducted using the USB pack dataset, driven by a specialized multi-component objective function designed to synchronize three distinct tasks. This training phase involves the iterative optimization of a composite loss, L_{all}, formulated as follows:

$$L_{all} = \lambda_1 L_{box} + \lambda_2 L_{cls} + \lambda_3 L_{dfl} + \lambda_4 L_{kpt} + \lambda_5 L_{kobj} + \lambda_6 L_{seg} \quad (1)$$

where the individual components address specific geometric and semantic requirements:

- Detection: L_{box} utilizing Complete IoU (CIoU) and L_{dfl} (Distribution Focal Loss) optimize bounding box regression, while L_{cls} employs Binary Cross Entropy (BCE) for category prediction.
- Keypoint detection: L_{kpt} calculates keypoint localization errors through Object Keypoint Similarity (OKS), and L_{kobj} (Keypoint Objectness) ensures the visibility of predicted points.
- Instance segmentation: L_{seg} optimizes mask generation by comparing the predicted masks derived from the product of mask coefficients and prototypes-against ground truth data using BCE.

By fine-tuning these parameters, the model enhances its adaptability to varying lighting conditions and complex object orientations, ensuring robust performance across diverse industrial scenarios. Following the training phase, the model undergoes rigorous evaluation to verify its effectiveness in cluttered environments. The evaluation specifically focuses on the precision of the generated masks and keypoints, which are calculated through the coordinated interaction of the multitask heads. This ensures the model's instance segmentation capability in accurately isolating each object from the background, even in high-density stacking. Successful mask generation and keypoint localization, validated by the minimized composite loss, provide a critical foundation for high-precision 3D localization and robotic manipulation.

2.3. Grasp candidate selection

To improve grasping robustness in cluttered industrial scenes, the proposed method first applies instance segmentation to isolate individual object candidates. For each detected instance, mask integrity and keypoint reliability are evaluated to determine whether the object is suitable for top-down grasping. Candidates with incomplete masks, unreliable keypoints, severe occlusion, or unsafe grasp positions are rejected. The remaining candidates are assigned a feasibility score and ranked accordingly, allowing the robot to select the most reliable grasping target.

The mask integrity score evaluates whether the predicted instance mask is complete and reliable enough for top-down grasping. It is defined as:

$$S_{mask,i} = w_a S_{area,i} + w_c S_{compact,i} \quad (2)$$

where $S_{area,i}$ measures whether the mask area is sufficient, $S_{compact,i}$ evaluates shape compactness, and the weighting coefficients w_a and w_c are both set to 0.5.

The keypoint reliability score evaluates whether the predicted keypoints are suitable for estimating a stable grasping pose. It is defined as:

$$S_{kpt,i} = w_v S_{vis,i} + w_p S_{inside,i} + w_g S_{geo,i} + w_q S_{conf,i} \quad (3)$$

where $S_{vis,i}$ represents keypoint visibility, $S_{inside,i}$ checks whether the keypoints lie inside the predicted mask, $S_{geo,i}$ evaluates geometric consistency, and $S_{conf,i}$ denotes the confidence of the predicted keypoints. The weighting coefficients w_v, w_p, w_g, and w_q are assigned equal values of 0.25.

The final feasibility score of each candidate is computed by combining mask integrity and keypoint reliability:

$$S_i = \alpha S_{mask,i} + \beta S_{kpt,i} \quad (4)$$

where α and β are weighting coefficients, both set to 0.5.

2.4. 6D pose estimation

Figure 4. *Geometric calculation of the final 6D target pose from keypoints.*

The object local coordinate frame is constructed using two adjacent edges and the surface normal obtained from their cross product, as shown in Figure 4. The rotation matrix is formed by the orthonormal basis vectors of the object frame, while the translation vector is defined as the object pickup position (P_{offset}) with an optional offset along the surface normal. The 6D Pose is represented by the homogeneous transformation matrix OCT [21], [23], which maps the object's local frame to the camera frame.

$$OCT = \begin{bmatrix} u_x & 0 & 0 & P_{offset,x} \\ 0 & u_y & 0 & P_{offset,y} \\ 0 & 0 & u_z & P_{offset,z} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

where,

$$u_x = \frac{P_2 - P_1}{\|P_2 - P_1\|} \quad (6)$$

$$v_y = P_4 - P_1 \quad (7)$$

$$u_z = \frac{u_x \times v_y}{\|u_x \times v_y\|} \quad (8)$$

$$u_y = u_z \times u_x \quad (9)$$

In this coordinate system, u_x is aligned with the longitudinal edge of the pack to define its primary orientation. The z-axis, u_z is defined as the surface normal of the object's plane; given the planar nature of the USB pack, this vector is computed as the cross product of the two primary edge vectors. Finally,

the y-axis, u_y is derived to ensure a perfectly orthonormal right-handed system, effectively correcting any non-orthogonality resulting from raw measurement noise.

Table 1. *Performance of the Deep Learning Model.*

Side of USB pack set	Bounding box			Mask			Keypoint		
	Pre	Rec	mAP50	Pre	Rec	mAP50	Pre	Rec	mAP50
Back side of USB pack	0.932	0.941	0.937	0.872	0.882	0.798	0.855	0.824	0.892
Front side of USB pack	**0.974**	**0.962**	**0.971**	0.87	0.861	**0.801**	**0.946**	**0.903**	**0.96**

Pre, Rec and mAP50 are short for Precision, Recall, and mean average precision, respectively.

Table 2. *Performance Comparison with Single-Task Deep Learning Models.*

USB pack sets	Bounding box			Mask			Keypoint			Param (M)	GFLOPs
	Pre	Rec	mAP50	Pre	Rec	mAP50	Pre	Rec	mAP50		
YOLOv8n Segmentation	0.982	0.979	0.969	0.892	0.897	0.853	-	-	-	3.26	11.5
YOLOv8n Keypoint	0.961	0.960	0.973	-	-	-	0.920	0.874	0.937	3.08	8.4
Our method	0.953	0.952	0.954	0.871	0.872	0.800	0.901	0.864	0.926	4.84	11.7

Table 3. *Processing Time Comparison Between the Segmentation-Keypoint Pipeline and the Proposed Method.*

Method	Processing time (ms)
Segmentation + Keypoint	4.4
Our method	2.7

2.5. Coordinate system transformation

Grasp planning operation needs the information of the pose of the object in the robot coordinate system. Hence, the relationship between the camera and the end-effector of the robot, shown as Figure 5, is found by hand-eye calibration [23].

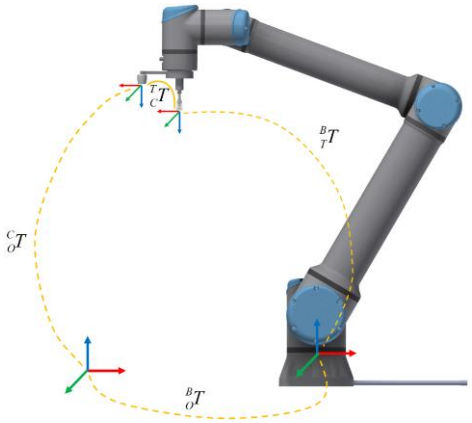

Figure 5. *Relative coordinate of the object with respect to the robot base and camera.*

To locate the object relative to the robot base, following composite transformation can be used:

$$OBT = OCT.CTT.TBT \tag{10}$$

where OBT is transformation matrix between the robot base and the object; OCT is transformation matrix between the camera and the object; CTT is transformation matrix between the camera and tool; TBT is a transformation matrix between the tool and the robot base.

3. Experiments and Results

In this section, various experiments are conducted to evaluate the performance of the proposed method on the planar object dataset. First, a set of RGB images is collected and annotated with instance masks, bounding boxes, and keypoints for each planar object. The dataset is then divided into training and validation sets, with data augmentation applied to enrich the dataset. As described in Subsection 2.1, the model that simultaneously performs keypoint detection and instance segmentation is referred to as a multi-task model for this task.

Next, hand–eye calibration is performed to determine the relationship between the camera coordinate system and the robot end-effector coordinate system. Finally, the pose of each planar object is estimated using the keypoints mapping, and the robot executes the picking task based on the transformation matrix defined in Equation (10).

3.1. Multi-task deep learning model

The experiments were conducted on a workstation running Ubuntu 22.04.5 LTS with 32 GB RAM and an Intel Core i9-10900X CPU @ 3.70 GHz. Deep learning model development and training were implemented in Python 3.10 using the PyTorch 2.9.1 framework and accelerated using CUDA 12.8 and cuDNN 9.12.0 on an NVIDIA GeForce RTX 3080Ti GPU with 12 GB of VRAM. To ensure experimental reliability, none of the models in this study used pre-trained weights.

The training set is augmented prior to model training, while the validation set is kept unchanged to assess the model's performance in a realistic setting. The input image resolution is resized to 640 × 640 × 3. The learning rate is set to 0.002, and the Adam optimizer is used for the initial training configuration. The training stops once the loss falls below a threshold. The systematic procedure for the proposed training methodology is detailed in Algorithm 1, which outlines each computational step of the multi-task learning process. For each mini-batch, input images undergo a forward pass through a shared backbone and a FPN/PAN neck to extract multi-scale features. These features are subsequently processed by a unified multi-task head to generate detection boxes, keypoint coordinates, and instance segmentation masks. To ensure effective supervision, a task-aligned assigner is employed to dynamically determine positive samples by aligning prediction quality with ground truth. The training objective is governed by a composite loss function, L_{all}, computed as the weighted sum of detection, keypoint and segmentation components. The entire network is then optimized jointly via backpropagation to refine the shared feature representations for industrial bin-picking tasks.

Algorithm 1: Training algorithm of the proposed multi-task network

Input: Multi-task network F with parameter Θ; training set T; loss weights Λ; threshold thr.
Output: trained network $F(x;\Theta)$
procedure TRAIN(F, T)
  repeat
    Sample a mini-batch (x_s, y_s) from T
    $(feats, \hat{k}, \hat{m}, proto) \leftarrow F(x_s;\Theta)$
    $(\hat{d}, \hat{c}) \leftarrow SplitDet(feats)$
    $\hat{b} \leftarrow DecodeBox(MakeAnchors(feats), \hat{d})$
    $(b^*, c^*, M_{fg}, g) \leftarrow Assign(\hat{b}, \hat{c}, y_s)$
    Compute L_{box}, L_{cls}, L_{dfl}, L_{kpt}, L_{kobj}, L_{seg}
    $L_{all} = \lambda_1 L_{box} + \lambda_2 L_{cls} + \lambda_3 L_{dfl} + \lambda_4 L_{kpt} + \lambda_5 L_{kobj} + \lambda_6 L_{seg}$
    Update Θ by backpropagation
  until $L_{all} < thr$
end procedure
return $F(x_s;\Theta)$

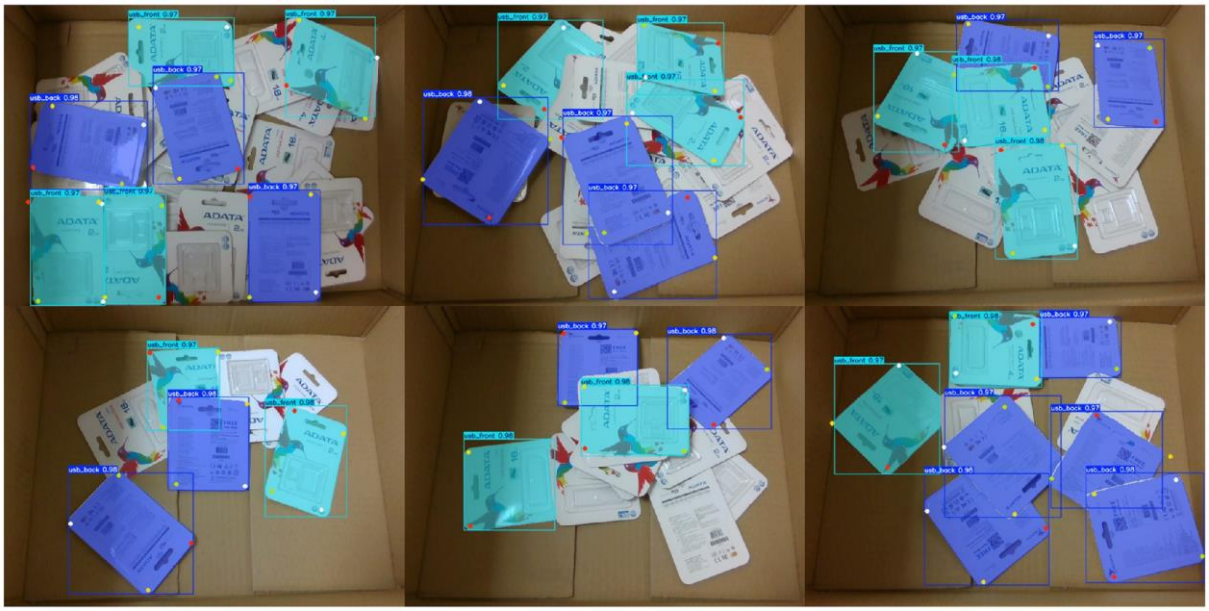

Figure 6. *The output of the multi-task deep learning model.*

Model performance is evaluated using Precision, Recall, and mAP50 metrics, as presented in Table 1. The results indicate that both the front and back sides of the USB pack are detected with high accuracy. Specifically, the front side achieves a bounding box precision of 97.4%, which is higher than that of the back side. In contrast, for mask precision, the back side outperforms the front side. These results demonstrate the model's strong performance on the USB pack dataset. The outputs are illustrated in Figure 6. However, the model performs poorly in highly cluttered practical environments, where unstable lighting and specular reflection from planar surfaces can reduce image quality and affect the accuracy of mask and keypoint detection. Future work will focus on improving robustness by increasing cluttered training samples, applying lighting- and reflection-aware augmentation, and developing occlusion-aware segmentation and keypoint detection strategies.

Table 2 compares the proposed method with single-task YOLOv8n segmentation and YOLOv8n keypoint models. The single-task models achieved slightly higher accuracy because each model optimizes only one task-specific loss, which converges more easily on the current small dataset. In contrast, the proposed method jointly learns bounding box detection, mask segmentation, and keypoint estimation, so the loss function must balance multiple objectives. This may reduce the accuracy of each individual task under limited training data.

However, as shown in Table 3, the proposed method is more efficient than using two separate models sequentially to obtain both segmentation and keypoints. The sequential Segmentation + Keypoint pipeline requires 4.4 ms, whereas the proposed unified method requires only 2.7 ms. Therefore, although the proposed method has slightly lower accuracy, it provides faster inference and a more compact multi-task solution for real-time robotic applications.

3.2. Grasp candidate selection and pose estimation

Experimental results demonstrate that the proposed top-down strategy effectively generates a prioritized list of graspable candidates and estimates a 6D pickup pose for each valid object. Candidates possessing valid instance masks and reliable keypoints are prioritized for pose estimation, whereas those with missing or low-confidence keypoints are excluded to prevent uncertain grasping maneuvers, as detailed in Algorithm 2. The 2D spatial posture of the object is represented by the red and green vectors in Figure 7, which are generated from the predicted keypoints. These points subsequently facilitate the calculation of the 3D transformation matrix, with the resulting posture visualized in Figure 4.

Algorithm 2: Grasp candidate selection

Input: Input image I; predicted bounding boxes $\hat{B}$; instance masks $\hat{M}$; keypoints $\hat{K}$.
Output: Ranked list of feasible top-down grasping candidates G
Initialize candidate list $G \leftarrow \emptyset$
for each detected instance i **do**
  Extract bounding box $\hat{b_i}$, mask $\hat{m_i}$, and keypoint $\hat{k_i}$
  Compute $S_{mask,i} = w_a S_{area,i} + w_c S_{compact,i}$
  Compute $S_{kpt,i} = w_v S_{vis,i} + w_p S_{inside,i} + w_g S_{geo,i} + w_q S_{conf,i}$
  if $S_{mask,i} < \tau_m = 0.91$ or $S_{kpt,i} < \tau_k = 0.94$ **then**
    continue
  end if
  Compute $S_i = \alpha S_{mask,i} + \beta S_{kpt,i}$
  if $S_i \geq \tau_s = 0.92$ **then**
    Add (i, p_i, θ_i, S_i) to G
  end if
end for
Sort G in descending order according to S_i
return G

Table 4. *Quantitative Evaluation of Surface Normal Orientation Error and Translation Error Using Predicted Keypoints on the USB Pack Dataset.*

Side of USB pack set	Normal orientation error (˚)	Translation error (mm)
Back side of USB pack	3.05 ± 0.62	3.7 ± 0.72
Front side of USB pack	2.97 ± 0.81	3.2 ± 0.50

Table 4 presents the annotation-based pose consistency analysis for evaluating the influence of keypoint localization errors on pose estimation. The back side of the USB pack achieved a normal orientation error of 3.05 ± 0.62° and a translation error of 3.7 ± 0.72 mm, while the front side achieved a normal orientation error of 2.97 ± 0.81° and a translation error of 3.2 ± 0.50 mm. These results indicate that the keypoint-derived surface normal estimation remained relatively stable, with an average orientation error of approximately 3°. Although planar USB packs are sensitive to small keypoint deviations, the proposed method still provided acceptable pose consistency under the current experimental conditions.

Figure 7. *The results of candidate selection and pose estimation.*

4. Conclusions

In this research, a multi-task deep learning framework was developed and evaluated for a robotic bin-picking system using USB packs as a representative planar object case study. Unlike traditional methods that rely on 3D CAD models for mapping and estimation, this approach reconstructs the 6D pose directly from four predicted keypoints per object. To handle cluttered industrial scenes, instance segmentation is employed to isolate individual objects, while both mask integrity and keypoint reliability are utilized to classify and rank feasible top-down grasping candidates. While the results demonstrate the fundamental effectiveness of this approach, performance in high-density clutter is currently constrained by the training dataset's size. This limitation can lead to inaccurate keypoint predictions, subsequently degrading pose estimation accuracy. Expanding the dataset with more diverse scenarios is expected to significantly enhance the model's robustness and reliability.

Furthermore, inherent symmetry in USB pack geometry and the presence of low-texture surfaces often introduce significant ambiguities, leading to potential errors in pose estimation. To mitigate these issues, the number of keypoints defined on USB pack can be increased, thereby providing more geometric constraints to resolve orientation uncertainty. Within the proposed multi-task framework, this is achieved by refining the keypoint head and the corresponding L_{kpt} and L_{kobj} loss components to supervise a denser set of local features, ensuring more robust spatial alignment even in challenging visual conditions.

As future work, the proposed framework will be extended toward integration with our digital twin framework [24], [25]. By incorporating accurate real-time pose estimation, the digital twin could enable virtual physical synchronization, continuous monitoring, and optimization of robotic manipulation processes.

Acknowledgments

This research is funded by the Vietnam Ministry of Education and Training under grant number B2024-VGU-03.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] W. Abbeloos and T. Goedemé, "Point pair feature based object detection for random bin picking," in *Proc. 13th Conf. Comput. Robot Vis. (CRV)*, 2016.
[2] W. C. Chang and C. H. Wu, "Eye-in-hand vision-based robotic bin-picking with active laser projection," *Int. J. Adv. Manuf. Technol.*, vol. 85, nos. 9–12, pp. 2873–2885, 2016.
[3] M. Y. Liu *et al.*, "Fast object localization and pose estimation in heavy clutter for robotic bin picking," *Int. J. Robot. Res.*, vol. 31, no. 8, pp. 951–973, 2012.
[4] C. Martinez, H. Chen, and R. Boca, "Automated 3D vision guided bin picking process for randomly located industrial parts," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, 2015.
[5] C. M. Lin *et al.*, "Visual object recognition and pose estimation based on a deep semantic segmentation network," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9370–9381, 2018.
[6] J. Vidal *et al.*, "A method for 6D pose estimation of free-form rigid objects using point pair features on range data," *Sensors*, vol. 18, no. 8, Art. no. 2678, 2018.
[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[8] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, no. 1, pp. 3–24, 2007.
[9] Z. Q. Zhao *et al.*, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019.
[10] M. Blum *et al.*, "A learned feature descriptor for object recognition in RGB-D data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2012.
[11] E. Brachmann *et al.*, "Learning 6D object pose estimation using 3D object coordinates," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014.
[12] E. Brachmann *et al.*, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.
[13] T. T. Do *et al.*, "Deep-6DPose: Recovering 6D object pose from a single RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.
[14] C. H. Wu, S. Y. Jiang, and K. T. Song, "CAD-based pose estimation for random bin-picking of multiple objects using an RGB-D camera," in *Proc. 15th Int. Conf. Control, Autom. Syst. (ICCAS)*, 2015.

[15] Y. K. Chen *et al.*, "Random bin picking with multi-view image acquisition and CAD-based pose estimation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, 2018.
[16] A. Wong *et al.*, "Fast GraspNeXt: A fast self-attention neural network architecture for multi-task learning in computer vision tasks for robotic grasping on the edge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2023.
[17] M. Gilles *et al.*, "MetaGraspNetV2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 3, pp. 2302–2320, Jul. 2024.
[18] D. Liu *et al.*, "Point pair feature-based pose estimation with multiple edge appearance models (PPF-MEAM) for robotic bin picking," *Sensors*, vol. 18, no. 8, Art. no. 2719, 2018.
[19] T. T. Le and C. Y. Lin, "Bin-picking for planar objects based on a deep learning network: A case study of USB packs," *Sensors*, vol. 19, no. 16, Art. no. 3602, 2019.
[20] B. Munkhtulga, K. Gou, and C. Lodoiravsal, "6D pose estimation of transparent objects from a single RGB image for robotic manipulation," *IEEE Access*, vol. 10, pp. 114897–114906, 2022.
[21] X. Liu *et al.*, "A Pose Estimation Approach Based on Keypoints Detection for Robotic Bin-picking Application," *2021 China Automation Congress (CAC)*, Beijing, China, 2021, pp. 3672-3677, doi: 10.1109/CAC53003.2021.9727987.
[22] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLOv8*, 2023.
[23] J. Ha, "Probabilistic framework for hand-eye and robot-world calibration AX = YB," *IEEE Trans. Robot.*, vol. 39, no. 2, pp. 1030–1047, 2023.
[24] Q. H. Dong, T. K. Nguyen, C. C. Tran, T. T. Pham, D. T. Do, H. V. K. Nguyen, and Q. C. Nguyen, "Effectiveness of digital twin framework for collaborative robotic manipulation," *J. Tech. Educ. Sci.*, vol. 20, no. 4, 2025.
[25] Q. H. Dong, T. T. Pham, K. Nguyen, C. C. Tran, H. H. Tran, and D. T. Do, "CAD-guided 6D pose estimation with deep learning in digital twin for industrial collaborative robot manipulation," *EAI Endorsed Trans. AI Robot.*, vol. 4, 2025, doi: https://doi.org/10.4108/airo.9676.

The-Thinh Pham received his Bachelor degree in Mechatronic Engineering from Can Tho University of Technology, Vietnam in 2019. He received his Master degree in Mechanical Engineering from National Taiwan University of Science and Technology, Taiwan in 2023. He is currently pursuing a Ph.D. degree with Mechanical Engineering, National Taiwan University of Science and Technology. His current research interests include computer vision, deep learning and robotics.
Email: ptthinh@ctuet.edu.vn. ORCID: https://orcid.org/0009-0009-4426-0457

Tuan-Khanh Nguyen received the B.S. and M.S. degrees in electronic telecommunication from Can Tho University, Vietnam in 2010 and Ho Chi Minh University of Technology, Vietnam in 2014, respectively. He got the PhD degree in Electronic and Computer Engineering at National Taiwan University of Science and Technology in 2023. He is currently working at the Vietnamese-German University as a Junior Lecturer in Semiconductor and Microsystems Engineering. His research interests include semiconductors, radio-frequency biomedical sensors, noncontact vital-sign radar sensors, and microwave circuits and modules.
Email: khanh.nt@vgu.edu.vn. ORCID: https://orcid.org/0000-0002-5162-4417

Chi-Cuong Tran received the B.S. and M.S. degrees in Automation and Control Engineering from Can Tho University, Vietnam, in 2015 and 2017, respectively. He received his Ph.D. degree in Mechanical Engineering from National Taiwan University of Science and Technology (NTUST), in 2023. He is currently working as a postdoctoral researcher at National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include intelligent robotics, intelligent automation, computer vision, and machine learning.
Email: tccuong09@gmail.com. ORCID: https://orcid.org/0009-0004-7698-0970

Quang-Huan Dong is a postdoctoral researcher in Business Information Systems at the Vietnamese-German University, Vietnam. He obtained a Bachelor's degree in Information Technology from the University of Science, Vietnam National University Ho Chi Minh City, in 2008. He completed a dual Master's degree in Business Information Systems jointly offered by the Vietnamese-German University, Vietnam, and Heilbronn University, Germany, in 2015. He received his PhD from the TUM School of Engineering and Design, Technical University of Munich, Germany, in 2023, where he conducted his research at the Institute of Automation and Information Systems. Prior to his academic career, he worked as a software engineer, technical lead, and project manager in the software industry for more than eight years. His research focuses on factory automation, software engineering, and digital twins.
Email: huan.dq@vgu.edu.vn. ORCID: https://orcid.org/0000-0002-3085-3464

Duy-Tan Do received the B.S. degree in Electronics and Telecommunications from the Ho Chi Minh City University of Technology (BK-HCM), Vietnam, in 2010, the M.S. degree in Wireless Communications from the Kumoh National Institute of Technology, South Korea, in 2013, and the Ph.D. degree in Electronic and Telecommunication Engineering from the Autonomous University of Barcelona, Spain, in 2019. He is currently with the Department of Electronics and Information Engineering, HCMC University of Technology and Engineering (HCM-UTE), Vietnam, as an associate professor. His main research interests include resource allocation optimization for wireless networks and advanced coding for reliable systems.
Email: tandd@hcmute.edu.vn. ORCID: https://orcid.org/0000-0003-4570-0441

Hoang-Vinh-Khang Nguyen received his B.S. degree in Electrical Engineering and Information Technology (EEIT, now Electrical and Computer Engineering) from the Vietnamese-German University (VGU) in 2015 and his M.S. degree in Mechatronics and Sensor Systems Technology (MST) from VGU in 2018. As a recipient of the DAAD scholarship, he spent one year at Karlsruhe University of Applied Sciences, Germany, where he conducted research for his master's thesis. His research interests include biomedical signal processing for rehabilitation applications and exoskeleton robot control.
Email: khang.nhv@vgu.edu.vn. ORCID: https://orcid.org/0000-0003-0256-2237

Quang-Chien Nguyen received the B.S. degree in Automation and Control Engineering from Ho Chi Minh City University of Technology and Engineering (HCM-UTE), Vietnam in 2023. He is currently pursuing a Master's degree in Automation and Control Engineering at HCM-UTE. His research interests include robotics, mobile robot, intelligent control and motion control.
Email: 2431101@student.hcmute.edu.vn. ORCID: https://orcid.org/0009-0008-1332-7746