Multi Task-Guided 6D Object Pose Estimation

Thu-Uyen Nguyen
FPT University
Hanoi, Vietnam
uyennthe176614@fpt.edu.vn

Van-Duc Vu
FPT University
Hanoi, Vietnam
ducvvhe176438@fpt.edu.vn

Van-Thiep Nguyen
FPT University
Hanoi, Vietnam
thiepnvhe173027@fpt.edu.vn

Ngoc-Anh Hoang
FPT University
Hanoi, Vietnam
anhhnhe186401@fpt.edu.vn

Duy-Quang Vu
FPT University
Hanoi, Vietnam
quangvdhe163133@fpt.edu.vn

Duc-Thanh Tran
FPT University
Hanoi, Vietnam
thanhtdhe176812@fpt.edu.vn

Khanh-Toan Phan
FPT University
Hanoi, Vietnam
toanpkhe170983@fpt.edu.vn

Anh-Truong Mai
FPT University
Hanoi, Vietnam
TruongMAHE182474@fpt.edu.vn

Van-Hiep Duong
FPT University
Hanoi, Vietnam
hiepdvhe181185@fpt.edu.vn

Cong-Trinh Tran
FPT University
Hanoi, Vietnam
trinhtche160916@fpt.edu.vn

Ngoc-Trung Ho
FPT University
Hanoi, Vietnam
trunghnhe172033@fpt.edu.vn

Quang-Tri Duong
FPT University
Hanoi, Vietnam
TriDQGCH210221@fpt.edu.vn

Phuc-Quan Ngo
FPT University
Hanoi, Vietnam
QuanNPGCH211110@fpt.edu.vn

Dinh-Cuong Hoang
FPT University
Hanoi, Vietnam
cuonghd12@fe.edu.vn

Abstract

Object pose estimation remains a fundamental challenge in computer vision, with cutting-edge methods relying on both RGB and depth data. Depth information is pivotal, offering crucial geometric cues that enable algorithms to navigate occlusions, fostering a more comprehensive scene understanding and precise pose estimation. However, RGBD-based methods often require specialized depth sensors, which can be costlier and less accessible compared to standard RGB cameras. Consequently, research has explored techniques aiming to estimate object pose solely from color images. Yet, the absence of depth cues poses challenges in handling occlusions, comprehending object geometry, and resolving ambiguities arising from similar colors or textures. This paper introduces a end-to-end multi-task-guided object pose estimation method, utilizing RGB images as input and producing the 6D pose of multiple object instances. While our approach employs both depth and color images during training, inference relies solely on color images. We incorporate depth images to supervise a depth estimation branch, generating depth-aware features further refined through a cross-task attention module. These enhanced features are pivotal for our object pose estimation. Our method's innovation lies in significantly enhancing feature discriminability and robustness for object pose estimation. Through extensive experiments, we demonstrate competitive performance compared to state-of-the-art methods in object pose estimation.

CCS Concepts: • **Computing methodologies → Computer vision**.

Keywords: Pose estimation, robot vision systems , intelligent systems, deep learning, supervised learning, machine vision.

1 Introduction

Object pose estimation holds pivotal importance across diverse applications such as autonomous driving, robotic navigation, manipulation, and augmented reality [1, 4, 10–16, 34, 39]. The current methods can be categorized into two types: those that estimate object poses using only RGB images [3, 37] and those that utilize both RGB and depth (D) images [8, 35]. RGBD-based methods take advantage of depth information to extract additional features or descriptors that are not solely dependent on color. This fusion of RGB and depth features boosts the discriminative power of feature representations, resulting in improved object pose estimation performance. However, one limitation of RGBD-based methods is that they typically require specialized hardware, such as depth sensors, to acquire depth information. This hardware dependency may restrict the applicability of these methods in certain scenarios where such sensors are not available. In recent years, significant advancements in deep learning techniques have prompted researchers to address this issue by utilizing convolutional neural networks (CNNs) on RGB images alone [3, 37]. However, these RGB-based approaches suffer from the lack of geometry information, which limits their performance in challenging situations, including low-contrast scenes, textureless objects, variations in lighting, sensor noise, and occlusions. The absence of depth information in RGB-based methods can lead to the generation of less discriminative features compared to RGBD-based methods. Depth information provides valuable cues regarding the spatial relationships and geometry of objects within a scene, thereby enhancing the discriminative power of feature representations.

This paper aims to bridge the gap between RGB-only and RGB-D-based 6D object pose estimation methods by leveraging monocular depth estimation. Recent advancements in deep learning have demonstrated the potential of monocular depth estimation networks in inferring depth from RGB images. These networks generate depth maps, providing approximate depth information for each pixel in an RGB image. Utilizing this additional depth data enriches the features extracted from RGB images, leading to more accurate pose estimation. These augmented features capture geometric details and enhance the discriminative power of representations, reducing the disparity between RGB-only and RGB-D approaches. At the core of our research lies the introduction of a Multi-Task-Guided 6D Object Pose Estimation framework, integrating semantic segmentation and depth estimation as auxiliary tasks. This comprehensive framework facilitates a deeper understanding of the scene, bolstering the model's adaptability across diverse conditions. We conduct extensive evaluations of our proposed method on widely used public datasets, benchmarking its performance against established state-of-the-art methodologies.

2 Literature Review

Approaches to predicting the 6Dof pose of an object include regression or classification methods that directly estimate pose-related parameters from input images. PoseCNN, a pioneering CNN architecture [37], performs 6D object pose regression from a single RGB image by decomposing the task into distinct components. It estimates the object's 3D translation by localizing its center in the image and predicting the distance between the object and the camera. The 3D rotation estimation is achieved through regression to a quaternion representation. However, direct 3D rotation estimation faces challenges due to the nonlinearity of the rotation space, which can limit CNN generalizability. To tackle this, several approaches [20, 31] discretize the rotation space, transforming the estimation into a classification task by dividing the space into bins. Although this simplifies the problem, it often results in coarse estimates, requiring post-refinement steps. Recently, Trabelsi et al. [33] integrated object classification, initial pose estimation, and iterative refinement into an end-to-end framework using appearance features and flow vectors to enhance accuracy. Despite their intuitiveness, these methods may struggle with generalization in natural scene settings due to the inherently ill-posed nature of the problem. The current more effective strategy involves establishing 2D-3D correspondences and solving for pose-related parameters [25, 27, 30, 32].

Among correspondence-based techniques, keypoint-based methods [27, 32] have demonstrated promising 6Dof pose prediction without extensive post-processing. These methods typically employ deep networks to detect the 2D projections of 3D keypoints, followed by a Perspective-n-Point (PnP) solver for pose estimation [22]. PVNet [25] introduced a voting-based keypoint localization strategy that performs well under occlusion or truncation. It utilizes a CNN to regress pixel-wise vectors representing keypoints and uses these vectors associated with the object's pixels to vote for keypoint locations, facilitating robust recovery of occluded or truncated keypoints. Other works like [30, 38] have also adopted pixel-wise voting schemes to enhance keypoint localization. These methods establish correspondences independently, enforcing consistency post-detection through the PnP algorithm, which is not part of the deep network. To achieve a single-stage process, Hu et al. [18] implemented the PnP step as a deep network, making it trainable end-to-end. However, like two-stage methods, this approach's accuracy depends heavily on keypoint detection quality. Robust and representative feature extraction plays a critical role in keypoint detection performance. The accuracy of feature extraction influences the detection's robustness; hence, enhancing feature quality is pivotal for achieving reliable pose estimation results.

3 Methodology

The overall architecture of the proposed Multi Task-Guided Monocular Object Pose Estimation network is shown in Fig. 1. An input RGB image $\mathbf{I}$ is first fed into a shared backbone network. The generated backbone features are then forwarded to task-specific branches which predict a semantic segmentation map and a depth map respectively. We adopt ResNet-50 [7] as the multi-task shared encoder network, generating a feature map $F \in \mathbb{R}^{H \times W \times K}$, where K is the number of channels, H and W are the height and width of the map, respectively. The response maps from the final convolutional layer of the encoder are fed into each task-specific branch for the extraction of pixel-level task-related information.

3.1 Semantic Segmentation

To accomplish semantic segmentation, we employ a standard U-Net architecture coupled with the loss function L_{sem}

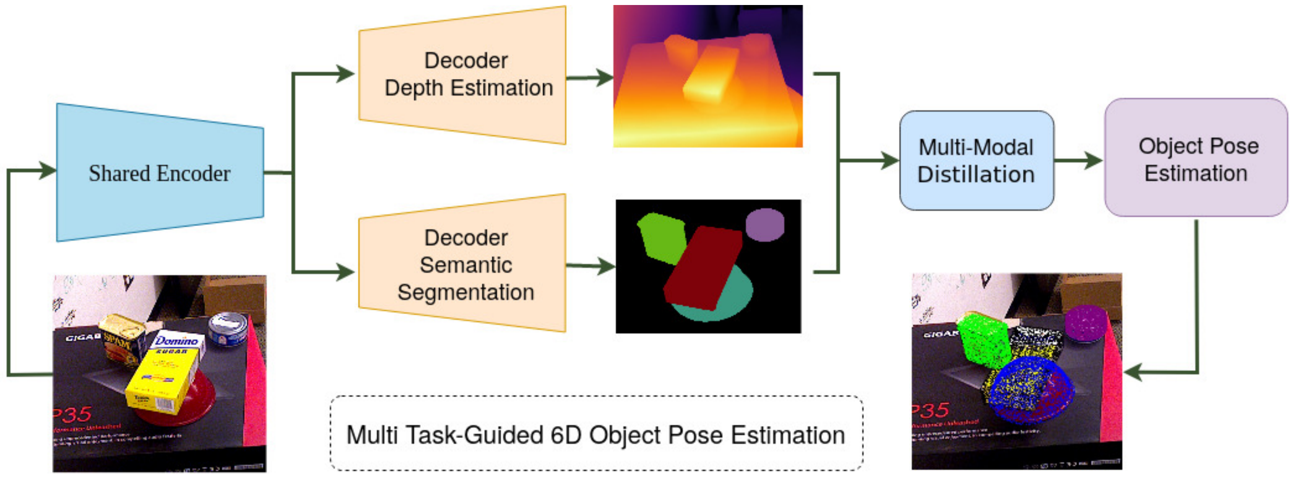


Figure 1. Overview of our network architecture.

[29]. The decoding process within each step involves upsampling the feature map, followed by a 2×2 convolution (up-convolution) that reduces the number of feature channels by half. Subsequent steps include concatenation with the correspondingly cropped feature map from the encoder, followed by two 3×3 convolutions, each followed by a Rectified Linear Unit (ReLU) activation function. The cropping step is essential to account for potential loss of border pixels during each convolution. At the final layer, a 1×1 convolution is utilized to map each 64-component feature vector to the desired number of classes.

3.2 Monocular Depth Estimation

Monocular Depth Estimation (MDE) and Semantic Segmentation are related tasks that involve understanding and interpreting scenes captured by a single camera. Combining these tasks in a multi-task learning framework can be highly beneficial for several reasons. MDE and Semantic Segmentation both require a deep understanding of the visual content within an image. Combining these tasks in a multi-task learning setting allows the model to learn a shared representation that captures both depth information and semantic context simultaneously. By jointly learning depth and semantic information, the model gains a more comprehensive understanding of the scene. For example, understanding not only the distance of objects but also their semantic categories (e.g., whether an object is a car, a person, or a building) provides richer information for various applications.

Monocular Depth Estimation addresses the challenge of predicting depth information from a single image, a task that inherently lacks unique solutions due to the multitude of 3D scenes that can correspond to a single 2D image. Recent advancements in MDE have leveraged Deep Convolutional Neural Network (DCNN)-based models, highlighting the superiority of deep features over manually crafted ones [23, 26, 28, 40]. Following [6], our approach discretizes continuous depth into intervals, framing the depth network learning as an ordinal regression problem. Our method involves the strategic use of DCNNs to integrate ordinal regression into dense prediction tasks. Specifically, we advocate for a spacing-increasing discretization strategy as opposed to a uniform discretization approach. This choice is motivated by the recognition that uncertainty in depth prediction escalates with the underlying ground-truth depth. Consequently, allowing for a relatively larger error in predicting greater depth values becomes imperative to mitigate the potentially over-strengthened influence of large depth values on the training process. Following the discretization step, the network is trained using an ordinal regression loss. This loss function is designed to incorporate the ordering of discrete depth values, capturing the nuanced relationships between different depth intervals. By emphasizing the ordinal nature of the depth prediction task, our approach enhances the model's ability to understand the inherent hierarchy within the depth values.

3.3 Multi-Modal Distillation

Utilizing task-specific features F_s and F_d, the Multi-Modal Distillation module (MMD) effectively integrates these diverse features by employing attention mechanisms. The initial step involves generating a cohesive set of fused feature maps through the concatenation operation $CONCAT(\cdot)$:

$$F_{fus} = CONCAT(F_s, F_d) \tag{1}$$

To refine the fused features F_{fus}, a channel attention block, denoted as $\mathcal{M}_{ca}$ [17], is incorporated. This block utilizes global average pooling to condense each feature map in F_{fus} into a single pixel, generating a 1D vector of length C. Subsequently, the vector undergoes processing through a Multi-layer Perceptron (MLP) network with a hidden layer and sigmoid activation. This operation is followed by element-wise multiplication with F_{fus}, aiming to recalibrate feature responses by amplifying important channels and suppressing less relevant ones. The resulting output from $\mathcal{M}_{ca}$ is denoted as F^c_{fus}. The overall process within $\mathcal{M}_{ca}$ can be summarized as:

$$\mathcal{M}_{ca}(F_{fus}) = \sigma(MLP(AvgPool(F_{fus}))) \tag{2}$$

$$F^c_{fus} = \mathcal{M}_{ca}(F_{fused}) \otimes F_{fus} \tag{3}$$

The features F^c_{fus} undergo further processing through a spatial attention block $\mathcal{M}_{sa}$ [36]. This spatial attention mechanism serves to identify informative regions and eliminate redundant depth-guided features arising from noise or irrelevant areas. The block initiates with an average-pooling operation to accentuate informative regions, generating a 2D map $F_{avg} \in \mathbb{R}^{W \times H}$. Subsequently, F_{avg} is convolved with a 7×7 filter and normalized using the sigmoid function. The resulting output, denoted as $\mathcal{M}_{sa}(F^c_{fus})$, undergoes element-wise multiplication with the original depth-guided features F_d to derive the enhanced feature representation F_e. The comprehensive attention process can be summarized as:

$$\mathcal{M}_{sa}(F^c_{fus}) = \sigma(f^{7\times7}(AvgPool(F^c_{fus}))) \tag{4}$$

$$F_e = \mathcal{M}_{sa}(F^c_{fus}) \otimes F^c_{fus} \tag{5}$$

where $\otimes$ denotes the element-wise multiplication, σ represents the sigmoid function, and $f^{7\times7}$ denotes a convolution operation using a 7×7 filter.

3.4 6D Object Pose Estimation

Given the enhanced feature $F_e = \{f_i\}$, we predict the object poses using a voting-based module introduced in our previous work [15]. The modules' learning is supervised jointly with a multi-tasks loss:

$$L = \lambda_1 L_{vote} + \lambda_2 L_{pose} + \lambda_3 L_{depth} + \lambda_4 L_{sem} \tag{6}$$

Here, λ_1, λ_2, λ_3 and λ_4 are the weights for each task. The loss consists of a voting loss L_{vote}, a pose loss L_{pose}, a depth loss L_{depth} and a semantic loss $L_{semantic}$. The pose loss function is defined as:

$$L_{pose} = L_t + \alpha L_{rot} + \beta L_{obj} + \gamma L_{sem} \tag{7}$$

Here, α, β, and γ are weights that scale the losses to similar scales. The pose loss includes a translation loss L_t (regression), an $L2$ loss between the output and the ground-truth rotation matrices, an objectness loss L_{obj}, and a semantic classification loss L_{sem}. The objectness loss is a cross-entropy loss for two classes (object or not), and the semantic classification loss is also a cross-entropy loss of semantic classes.

For asymmetric objects, the above loss L_{rot} for rotation is applicable. However, for symmetric objects with multiple correct 3D rotations, the loss L_{rot} is computed as:

$$L_{rot} = \frac{1}{m} \sum_{x_1 \in M} \left\| \min_{x_2 \in M} (\bar{R}x + \bar{T} - \hat{R}x + \hat{T}) \right\| \tag{8}$$

Here, M denotes the set of 3D model points, and m is the number of points. The loss is calculated as the average distance from vertices of the object model in the ground-truth pose to the closest vertices of the model in the estimated pose, ensuring alignment between the two 3D models.

4 Result and Discussion

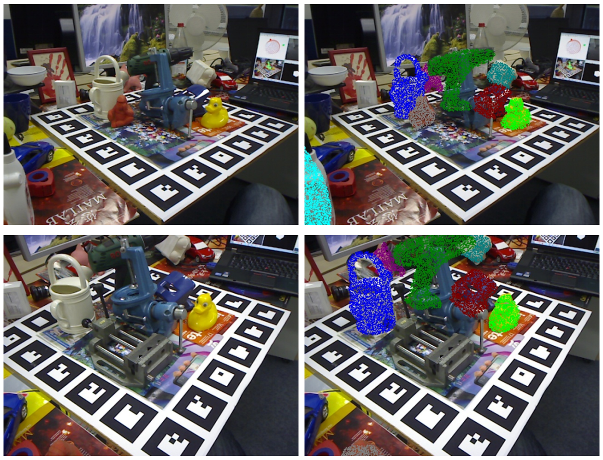
This section encompasses our experimental validation to ascertain the efficacy of the proposed method. The evaluation is conducted on two publicly available datasets: Occluded-LINEMOD [9] and YCB-Video [37]. Implementations were carried out using PyTorch [24] and Python platforms, leveraging a single Nvidia GeForce RTX 2080 Ti 11GB GPU with CUDA and the Linux operating system. For detailed implementations.

4.1 Evaluation Metrics

We assess the performance of 6D object pose estimation using two widely used metrics: ADD(-S) [9, 37] and 2D reprojection error (REP) [2]. The ADD(-S) metric calculates the mean distance between two transformed model points based on the estimated and ground-truth poses. If this distance falls below 10% of the model's diameter, the estimated pose is considered accurate. For symmetric objects, we adopt the ADD-S metric [37], which computes the mean distance based on the closest point distance. Additionally, we gauge the Area Under Curve (AUC) of the ADD(-S) metric by varying the distance threshold, with a maximum threshold of 10 cm [37]. On the other hand, the 2D reprojection error metric, proposed by [2], quantifies the mean distance between the projections of 3D model points using both the estimated and ground-truth poses. A pose is deemed accurate if this distance is below 5 pixels. This metric evaluates the precision of projecting 3D model points onto the 2D image plane.

4.2 Evaluation on Occluded-LINEMOD Dataset

Training the Network. Our network is trained from scratch using the Adam optimizer [21], with both training and testing input images set at a resolution of 640×480 pixels. We opt for an end-to-end training approach, employing a batch size of 8 and integrating standard data augmentation techniques. The training initiates with an initial learning rate of 0.001 and spans 220 epochs. At the 80th, 120th, and 180th epochs, we implement a learning rate decay strategy, reducing it by a factor of 0.1 at each step. The convergence of the training process is achieved in approximately 7 hours.



(a) Input **(b)** Predicted Poses

Figure 2. Qualitative results on Occluded-LINEMOD.

(a) Input **(b)** Predicted Poses

Figure 3. Qualitative results on YCB-Video dataset.

Results. Fig. 2 shows qualitative pose estimation results. The evaluation results in Table 1 demonstrate the superior performance of our proposed model compared to state-of-the-art RGB methods on the Occluded-LINEMOD dataset. Across various object categories such as 'Ape,' 'Can,' 'Cat,' 'Driller,' 'Duck,' '*Eggbox**,' '*Glue**,' and 'Holepun,' our method consistently outperforms the benchmark methods in both ADD(-S) and REP metrics. Notably, our model achieves improvements in both ADD(-S) and REP metrics, showcasing significant advancements in accurately estimating object poses, especially for objects with occlusion challenges. For instance, our model excels in the 'Duck' category with an impressive 57.4 ADD(-S) score and a remarkable 66.3 REP score, surpassing other methods by a substantial margin. Similarly, our approach showcases remarkable enhancements in challenging categories like 'Can,' 'Driller,' and 'Holepun,' demonstrating considerable improvements in both ADD(-S) and REP metrics compared to existing methods. The average scores of 60.6 for ADD(-S) and 62.5 for REP underline the consistent and superior performance of our method across diverse object categories, highlighting its effectiveness in handling occlusions and accurately estimating object poses in complex scenarios compared to the state-of-the-art methods on the Occluded-LINEMOD dataset. These results reinforce the effectiveness and robustness of our proposed approach in overcoming challenges posed by occlusions in 6D object pose estimation tasks.

4.3 Evaluation on YCB-Video Dataset

Training the Network. In line with the Occluded-LINEMOD dataset, we maintain a resolution of 640×480 pixels for both training and testing input images. We employ the Adam optimizer, initializing with a learning rate of 0.01. To aid learning, we schedule the decay of the learning rate at epochs 120, 160, 200 with a decay rate set at 0.1 for each scheduled step. Our training procedure involves a batch size of 8 and integrates commonly used data augmentation techniques. The training extends over 240 epochs and converges after approximately 12 hours.

Results. Fig. 3 shows qualitative pose estimation results. The comparison results from Table 2 demonstrate the remarkable superiority of our proposed method over state-of-the-art RGB-based techniques across multiple evaluation metrics on the YCB-Video dataset. Notably, our method achieves outstanding performance in all aspects evaluated: ADD(-S), REP, and AUC. Specifically, in terms of the ADD(-S) metric, our approach achieves a substantial improvement, securing a score of 62.7. This significant lead highlights our method's exceptional accuracy in estimating 6D object poses compared to other approaches. Correspondingly, in the REP metric, our method again demonstrates superiority, recording a score of 53.5. Additionally, in the AUC metric, our approach excels with a score of 91.4, surpassing DGECN and all other compared methods.

4.4 Runtime

Our method's processing speed was evaluated using the YCB-Video dataset to gauge its efficiency. The assessment utilized a single Intel Xeon E-2716G CPU running at 3.7 GHz and an Nvidia GeForce RTX 2080 Ti GPU with 11GB of memory to analyze 640×480 RGB images containing multiple objects. Our method demonstrates a real-time performance. It takes just around 40ms to process all objects within an image, highlighting its efficiency.

5 Conclusions

In conclusion, this work introduces a robust and efficient approach for 6D object pose estimation, addressing the challenges posed by the absence of depth cues in RGB-only scenarios. By leveraging depth information during training and

Table 1. Quantitative comparison Occluded-LINEMOD dataset with state-of-the-art RGB methods.

Method	PoseCNN		PVNet		Single-Stage		DGECN		Ours	
	ADD(-S)	REP	ADD(-S)	REP	ADD(-S)	REP	ADD(-S)	REP	ADD(-S)	REP
Ape	9.6	34.6	15.8	69.1	19.2	70.3	**50.3**	-	43.5	**71.0**
Can	45.2	15.1	63.3	**86.1**	65.1	85.2	75.9	-	**77.2**	84.4
Cat	0.9	10.4	16.7	65.1	18.9	67.2	26.4	-	**27.4**	**68.0**
Driller	41.4	7.4	65.7	73.1	69.0	71.8	77.5	-	**76.5**	**74.8**
Duck	19.6	31.8	25.2	61.4	25.3	63.6	54.2	-	**57.4**	**63.9**
*Eggbox**	22.0	1.9	50.2	8.4	52.0	12.7	57.8	-	**58.3**	**13.1**
*Glue**	38.5	13.8	49.6	55.4	51.4	56.5	66.9	-	**65.6**	**56.2**
Holepun	22.1	23.1	39.7	69.8	45.6	**71.0**	60.2	-	**61.1**	68.2
Average	24.9	17.2	40.8	61.1	43.3	62.3	58.7	-	**60.6**	**62.5**

Table 2. Quantitative comparison on YCB-Video dataset with state-of-the-art RGB methods.

Method	ADD(-S)	REP	AUC
PoseCNN [37]	21.3	3.7	75.9
PVNet [25]	45.2	47.4	73.4
SegDriven [19]	39.0	30.8	-
Single-Stage [18]	53.9	48.7	-
SO-Pose [5]	56.8	-	90.9
DGECN [3]	60.6	50.3	90.9
Ours	**62.7**	**51.5**	**91.4**

then relying solely on RGB data during inference, our method achieves competitive performance in object pose estimation. Our proposed multi-task-guided framework effectively integrates depth-aware features, enhancing discriminability and robustness crucial for accurate pose estimation. The utilization of depth-supervised training, followed by feature enhancement through cross-task attention, showcases the importance of leveraging complementary information for improving pose estimation accuracy.

The comprehensive experiments conducted on standard benchmarks demonstrate the efficacy of our approach, highlighting its competitive performance against state-of-the-art methods in 6DoF object pose estimation. Furthermore, the real-time operation of our method emphasizes its practical applicability in various real-world scenarios. However, there's room for future exploration in enhancing the robustness of pose estimation in challenging scenarios, such as severe occlusions or varied lighting conditions. Additionally, extending this approach to handle a wider array of object categories and diverse environments could further validate its versatility and generalization capabilities. Overall, this work provides a strong foundation for RGB-based object pose estimation and opens avenues for continued advancements in this crucial domain of computer vision.

References

[1] Pei An, Junxiong Liang, Kun Yu, Bin Fang, and Jie Ma. 2022. Deep structural information fusion for 3D object detection on LiDAR–camera system. *Computer Vision and Image Understanding* 214 (2022), 103295.

[2] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. 2016. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3364–3372.

[3] Tuo Cao, Fei Luo, Yanping Fu, Wenxiao Zhang, Shengjie Zheng, and Chunxia Xiao. 2022. DGECN: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 3783–3792.

[4] Liang-Chia Chen, Dinh-Cuong Hoang, Hsien-I Lin, and Thanh-Hung Nguyen. 2016. Innovative methodology for multi-view point cloud registration in robotic 3D object scanning and reconstruction. *Applied Sciences* 6, 5 (2016), 132.

[5] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. 2021. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 12396–12405.

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2002–2011.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[8] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. 2020. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 11632–11641.

[9] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. 2013. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11.* Springer, 548–562.

[10] Dinh-Cuong Hoang, Liang-Chia Chen, and Thanh-Hung Nguyen. 2016. Sub-OBB based object recognition and localization algorithm using range images. *Measurement Science and Technology* 28, 2 (2016), 025401.

[11] Dinh-Cuong Hoang, Achim J Lilienthal, and Todor Stoyanov. 2020. Object-RPE: Dense 3D reconstruction and pose estimation with convolutional neural networks. *Robotics and Autonomous Systems* 133 (2020), 103632.

[12] Dinh-Cuong Hoang, Achim J Lilienthal, and Todor Stoyanov. 2020. Panoptic 3D mapping and object pose estimation using adaptively weighted semantic information. *IEEE Robotics and Automation Letters* 5, 2 (2020), 1962–1969.

[13] Dinh-Cuong Hoang, Anh-Nhat Nguyen, Van-Duc Vu, Duy-Quang Vu, Van-Thiep Nguyen, Thu-Uyen Nguyen, Cong-Trinh Tran, Khanh-Toan Phan, and Ngoc-Trung Ho. 2023. Grasp Configuration Synthesis from 3D Point Clouds with Attention Mechanism. *Journal of Intelligent and Robotic Systems* 109, 3 (2023), 71.

[14] Dinh-Cuong Hoang, Johannes A Stork, and Todor Stoyanov. 2022. Context-aware grasp generation in cluttered scenes. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 1492–1498.

[15] Dinh-Cuong Hoang, Johannes A Stork, and Todor Stoyanov. 2022. Voting and attention-based pose relation learning for object pose estimation from 3d point clouds. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8980–8987.

[16] Dinh-Cuong Hoang, Todor Stoyanov, and Achim J Lilienthal. 2019. Object-rpe: Dense 3d reconstruction and pose estimation with convolutional neural networks for warehouse robots. In *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 1–6.

[17] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[18] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. 2020. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2930–2939.

[19] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. 2019. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3385–3394.

[20] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. 2017. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*. 1521–1529.

[21] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*.

[22] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. 2009. EPnP: An accurate O (n) solution to the PnP problem. *International journal of computer vision* 81 (2009), 155–166.

[23] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. 2021. Deep learning for monocular depth estimation: A review. *Neurocomputing* 438 (2021), 14–33.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[25] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. 2019. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4561–4570.

[26] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. 2023. iDisc: Internal Discretization for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21477–21487.

[27] Mahdi Rad and Vincent Lepetit. 2017. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*. 3828–3836.

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022).

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

[30] Chen Song, Jiaru Song, and Qixing Huang. 2020. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 431–440.

[31] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. 2018. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*. 699–715.

[32] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. 2018. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 292–301.

[33] Ameni Trabelsi, Mohamed Chaabane, Nathaniel Blanchard, and Ross Beveridge. 2021. A pose proposal and refinement network for better 6d object pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2382–2391.

[34] Van-Duc Vu, Dinh-Dai Hoang, Phan Xuan Tan, Van-Thiep Nguyen, Thu-Uyen Nguyen, Ngoc-Anh Hoang, Khanh-Toan Phan, Duc-Thanh Tran, Duy-Quang Vu, Phuc-Quan Ngo, et al. 2024. Occlusion-Robust Pallet Pose Estimation for Warehouse Automation. *IEEE Access* (2024).

[35] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. 2019. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3343–3352.

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[37] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2018. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *Robotics: Science and Systems (RSS)*.

[38] Xin Yu, Zheyu Zhuang, Piotr Koniusz, and Hongdong Li. 2020. 6dof object pose estimation via differentiable proxy voting loss. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[39] Haoruo Zhang and Qixin Cao. 2019. Holistic and local patch framework for 6D object pose estimation in RGB-D images. *Computer Vision and Image Understanding* 180 (2019), 59–73.

[40] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4106–4115.