Title:
Robust Object 6D Pose Estimation Under High Dynamic Ambient Light

Authors:
Lei Lu, lulei@haut.edu.cn, Henan University of Technology
Jiahe Zhu, 15137399020@163.com, Henan University of Technology
Wei Pan, vpan@foxmail.com, Department of R&D, OPT Machine Vision Tech Co., Ltd.
Haojun Zhang, zhj@haut.edu.cn, Henan University of Technology
Zhilong Su, zhilong8845@shu.edu.cn, Shanghai University
Qinghui Zhang, zqh131@163.com, Henan University of Technology
Wanxing Zheng, wanxingzheng@haut.edu.cn, Henan University of Technology
Ge Gao, gao.ge@mech-mind.net, Mech-Mind Robotics Technologies Ltd.
Peng Li, lipeng@haut.edu.cn, Henan University of Technology

Keywords:
pose estimation; high dynamic ambient light; attention mechanism; convolutional neural network

DOI: 10.14733/cadconfP.2025.132-137

Introduction:
Advancements in ToF cameras, structured light cameras, and LiDAR have significantly expanded applications in autonomous driving, industrial robotics, and augmented reality [1]. Among these, object 6D pose estimation, which determines an object's orientation and location in 3D space, is a critical task. It involves computing a rigid transformation from the object's coordinate system to the camera coordinate system, expressed as a translation vector T and a rotation matrix R. Traditional 6D pose estimation methods rely on geometric and feature-based approaches, which often struggle with robustness under varying environmental conditions. Deep learning-based models have shown promise, leveraging large datasets to adapt to complex scenarios. However, these models are still challenged by high dynamic ambient light conditions, where illumination variations and shadow effects significantly impact accuracy.

To overcome the impact of illumination variations on 6D pose estimation caused by high dynamic ambient light, we introduce an enhanced Gen6D [2] model designed to perform reliably under high dynamic ambient light conditions. Initially, a convolutional neural network (CNN) performs 2D object detection on the color map to identify the object's position and scale. A channel attention mechanism is then added to improve inter-channel information exchange and reduce noise, resulting in more robust feature representations. Additionally, a fast image-matching algorithm estimates the object's approximate rotation by comparing image similarities from the validation and training sets. Finally, a 3D CNN refines the pose by analyzing discrepancies between the initial pose and the ground truth through residual regression. Experiments show that our proposed method effectively addresses the challenges of 6D pose estimation under high dynamic ambient light conditions.

Main Idea:
At first, we evaluate the pose estimation performance of the traditional Gen6D model by artificially creating high-dynamic ambient light scenes and projecting light sources to induce uneven reflectivity on the surfaces of objects. Experiments demonstrate that Gen6D is ineffective in estimating the poses of objects with unevenly reflective surfaces, as shown in Fig. 1. This paper proposes an improved Gen6D method to enhance the performance in these challenging scenarios and makes the following

contributions: (1) We propose integrating an efficient channel attention (ECA) [3] module into the Gen6D [2] pipeline to enhance feature representation; (2) We developed a fast and effective image matching algorithm that improves robustness to varying lighting conditions. The details are presented below.

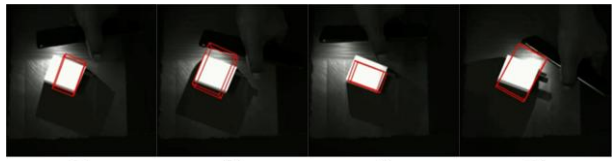

Fig. 1: Gen6D yields insufficient pose estimation results under high dynamic ambient light conditions. Panels (a) to (d) illustrate the outcomes of Gen6D in such environments, highlighting the model's poor performance attributed to the challenges posed by these dynamic lighting conditions.

The overall architecture of the improved model is illustrated in Fig. 2. Our architecture comprises three main components: the object detection module (detector), the rotation estimation module (selector), and the pose refinement module (refiner). To enhance the feature modeling capabilities of the end-to-end learning network without increasing model complexity, we proposed to integrate the ECA [3] module into the VGG [4] backbone. ECA module can enhance the model's focus on local information within the input data, thereby improving local feature extraction and effectively mitigating the impact of high dynamic ambient light. All three modules leverage the VGG-ECA backbone network. The ECA module enables the network to dynamically adjust its attention to different feature channels in the object image, thereby minimizing the interference from redundant information and improving pose estimation performance in high dynamic ambient light conditions. The details are shown below.

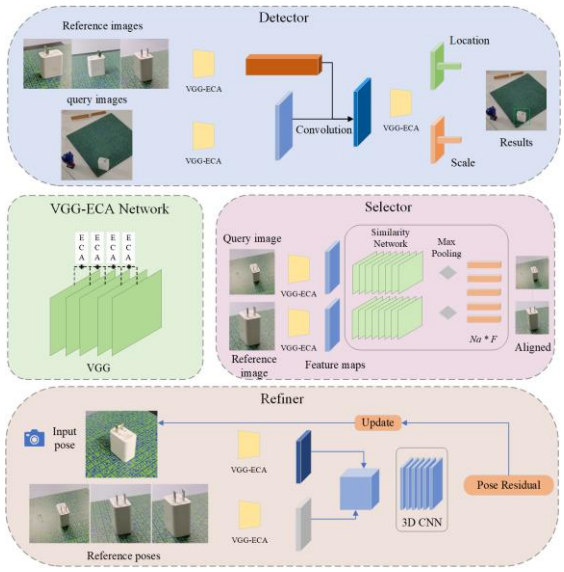

Fig. 2: An overview of our proposed method is presented, with the backbone being the VGG-ECA network.

This paper designs a novel convolutional neural network based on the Gen6D backbone VGG and the ECA module. By embedding the ECA channel attention mechanism into the relatively high-resolution layers of VGG, the network can better focus on both channel and spatial features while integrating all

features of the image. This enhancement enables the network to effectively learn the geometric characteristics of the object. To illustrate the feature extraction capabilities of the VGG convolutional neural network after incorporating the ECA channel attention mechanism, we use the object "Cat" from the LineMOD [5] dataset, as shown in Fig. 3.

Fig. 3: The feature extraction results after embedding the ECA module into the VGG network are presented. Panels (a) to (e) illustrate the two-dimensional feature extraction outcomes of the VGG-ECA network.

Embedding the ECA algorithm into the VGG backbone of the deep learning model enhances the stability of object detection under high dynamic ambient light conditions. In the object detection stage, VGG serves as the backbone. By integrating the ECA attention mechanism into the network layers corresponding to image channels with 64, 128, and 256 filters, the VGG convolutional neural network demonstrates improved effectiveness in extracting features from unevenly reflective images with varying surface materials. This integration enhances the representation of important feature channels while suppressing channel noise. The improved VGG-ECA network structure is illustrated in Fig. 4.

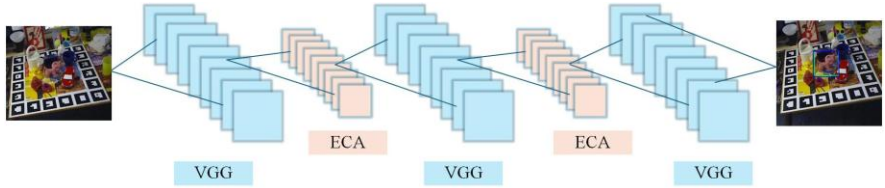

Fig. 4: The VGG-ECA structure, by incorporating the ECA module into VGG, enhances the network's ability to capture important features, particularly in high dynamic ambient light conditions. This enhancement can significantly improve the accuracy of object 6D pose estimation.

Compared to the traditional Gen6D, the VGG-ECA backbone network we proposed demonstrates superior performance across the detector, selector, and refiner modules. In particular, the detector module benefits significantly from the in-depth integration of the VGG-ECA backbone. We divide the detector module, one of the three primary components, into two parts: locating the 2D projection of the object center q and estimating the compact square bounding box size S_q that surrounds the unit sphere.

Specifically, applying depth estimation and object positioning in computer vision, by marking a compact bounding box in the image, which is the circumscribed rectangle of the object, the size of the circumscribed rectangle is the compact square bounding box size S_q , as shown in Fig. 5, and then by changing the position of the camera, we can get different projection positions to calculate the depth of the object. The calculation formula for depth is $d = 2 * \lambda * f / S_q$, where λ is the wavelength of light, f is the virtual focal length, and S_q is the size of the compact bounding box. This projected position and the depth of the object can determine the center position of the object and provide initial translation information for the attitude of the object.

By incorporating the ECA into the VGG architecture, we create a VGG-ECA structure. The VGG-ECA network is employed to extract feature maps from both the query image and the reference images. These feature maps from all reference images serve as convolutional kernels, which are convolved with the feature map of the query image to produce a score map.

To accommodate scale differences, this convolution is executed at N_s predefined scales by resizing the query image to various dimensions. Utilizing the multi-scale score map, this paper regresses a heat map and a scale map, subsequently selecting the position with the maximum value on the heat map as the 2D projection of the object's center. We use the scale values at the same location on the scale map to compute the bounding box size Sq = Sr ∗ s, where Sr is the size of the reference images. Leveraging the detected 2D projections and scales, we calculate the initial 3D translations and extract the object regions for subsequent processing.

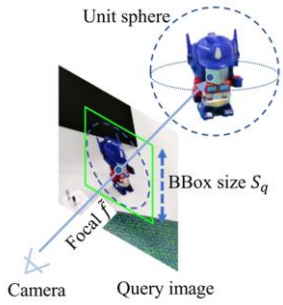

Fig. 5: Translation information calculation.

Experiments and results:

To verify the effectiveness of the proposed method, the model enhanced by ECA was evaluated using objects from the Gen6D dataset, which is known as GenMOP, a general model-free object pose dataset. To better highlight the experimental results of the deep learning model after incorporating ECA, we specifically selected objects with high-reflectivity surface materials, as these materials pose unique challenges in pose estimation due to their susceptibility to variations in ambient lighting, which can distort the object's surface appearance and affect pose accuracy.

The metrics used in this study, including Average Distance of Model Points (ADD) [5] and Projection Error (Prj-5), are widely adopted in 6D pose estimation tasks due to their effectiveness in quantifying pose accuracy. ADD measures the mean distance between the predicted and ground-truth points on the object, providing a direct assessment of pose estimation accuracy in 3D space. Prj-5 evaluates the alignment of the object's 2D projection on the image plane, ensuring that the estimated pose is visually consistent with the ground truth in camera view.

In calculating ADD, we consider a range of 10% of the object diameter (ADD-0.1d) . For the projection error, we analyze recall at a threshold of 5 pixels (Prj-5). The definition of ADD is as follows:

$$ADD = \frac{1}{m}\sum_{v\varepsilon\vartheta} \| (Rv+T) - (R^{*}v+T^{*}) \| \tag{3.1}$$

where v represents the vertex of the object ϑ, R and T represent the predicted pose, and R^* and T^* are the true pose. Prj-5 is a metric used to evaluate the accuracy of pose estimation. It measures the error between the projected 2D representation of the estimated 3D model on the image plane and the true projection. The specific definition is as follows:

$$di = \| project(Pi, pose_{est}) - project(Pi, pose_{gt}) \| \tag{3.2}$$

Where *project(Pi, pose)* is a function that projects the 3D point *Pi* onto the 2D image plane according to a given pose. $pose_{est}$ is the estimated pose and $pose_{gt}$ is the true pose. The value of Prj-5 is calculated by the following formula:

$$prj-5 = \frac{successful\_samples}{total\_samples} \tag{3.3}$$

The enhanced model continues to utilize Gen6D's official dataset, GenMOP, for training. By fine-tuning the loss weights, we aim to achieve optimal performance. To evaluate the effectiveness of this approach, RGB images were collected and assessed under high dynamic ambient light conditions. The

experimental results are visualized as 3D bounding boxes, as illustrated in Fig. 6. Part (a) shows the pose estimation results of the original Gen6D model under high dynamic ambient light, while part (b) displays the outcomes of the improved model incorporating the ECA. The results demonstrate that the inclusion of ECA significantly mitigates the adverse effects of high dynamic ambient light on object pose estimation.

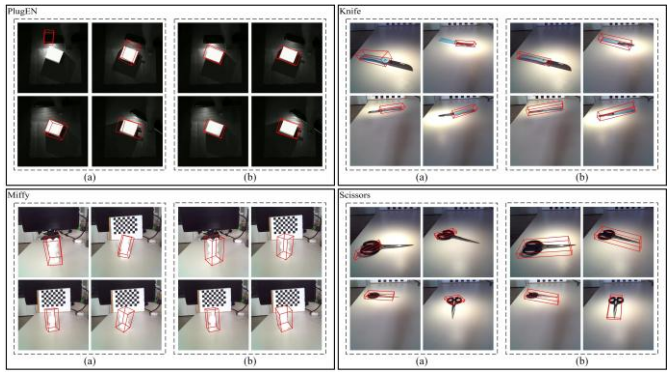

Fig. 6: Comparison of pose estimation results is presented, where part (a) illustrates the outcomes from the original Gen6D deep learning model, while part (b) displays the results from our proposed method. The findings indicate that our approach effectively captures the complete pose of the object, demonstrating notable improvements in detection accuracy.

To reflect illumination variations in high dynamic ambient light scenes, we categorized illumination intensity into five levels, from dark to bright: L1, L2, L3, L4, and L5. The pose estimation effects under these five intensities are visualized in Fig. 7, showcasing knife, scissors, cup, and miffy. Each group contains part (a), representing the pose detection outcomes of the original Gen6D model, and part (b), illustrating the results from the improved deep learning model.

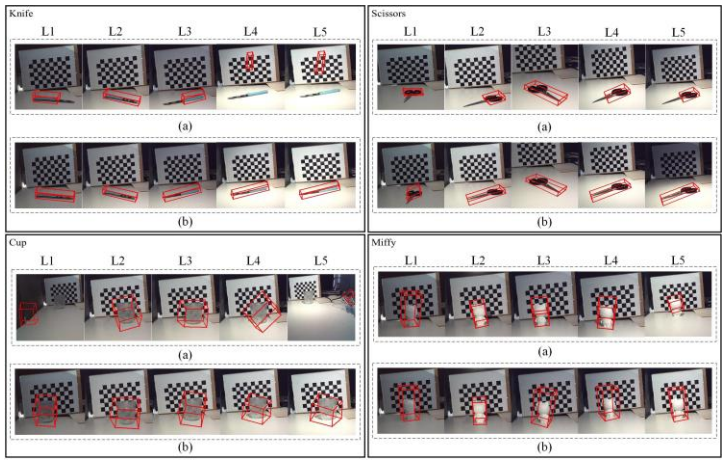

Fig. 7: Pose estimation results under five levels of light, ranging from dark to bright. Part (a) shows the results from the original deep learning model Gen6D, while part (b) displays the outcomes from our proposed method. Across varying ambient light intensities, our approach successfully detects the complete pose of the object, demonstrating enhanced performance in challenging illumination conditions.

Part (a) shows the pose estimation results using the original Gen6D model, where the accuracy is significantly reduced for objects like the knife and scissors. These objects, with their metal components, suffer from poor pose estimation performance under both low (L1, L2) and high (L4, L5) light intensities. The reflective properties of metal surfaces cause inconsistencies in the pose detection.

In contrast, part (b) demonstrates the improved model's performance after integrating the ECA module. The enhancement is particularly evident with the cup, which is made of glass. Under both low and high light intensities, the improved model exhibits more accurate pose estimation, as the ECA module helps mitigate the effect of lighting variations on the reflective glass surface.

Additionally, the pose estimation performance of our method is compared with that of the original deep learning model Gen6D using the official GenMOP dataset. Tab. 1 presents the comparative results, highlighting metrics such as ADD-0.1d and Prj-5. The experimental findings indicate that our method outperforms the original Gen6D model, demonstrating a notable improvement in overall average performance.

	Method	PlugCN	Miffy	Piggy	Scissors	TFormer	Knife	PlugEN	Avg.
ADD-0.1d	Gen6D[2]	24.21	64.29	74.37	31.03	65.87	63.24	23.36	49.48
	Ours	26.89	64.83	76.89	32.34	67.04	65.89	26.46	**51.48**
Prj-5	Gen6D[2]	99.69	99.57	95.48	90.52	99.60	69.73	72.90	89.64
	Ours	99.91	99.79	97.10	93.97	99.95	77.21	75.63	**91.94**

Tab. 1: Performance on the GenMOP dataset.

Conclusions:

This paper proposes a method for 6D object pose estimation under high dynamic ambient light conditions. The method integrates the ECA module into the feature extraction stage of Gen6D to address the challenges posed by varying light sources and their impact on the reflection of object surface materials during pose estimation. The official dataset GenMOP, used by Gen6D, serves as the foundation for training. After incorporating the ECA module, the deep learning model was retrained, and RGB camera images were collected for experimental evaluation. The results demonstrate that the deep learning model, enhanced with the ECA channel attention module, exhibits stable performance in estimating object poses under high dynamic ambient light conditions.

References:

[1] Ma, J.; Zhuo, S.; Qiu, L.; Gao, Y.; Wu, Y.; Zhong, M.; ... Chiang, P. Y.: A review of ToF-based LiDAR, Journal of Semiconductors, 45(10), 2024, 101201. 10.1088/1674-4926/24040015

[2] Liu, Y.; Wen, Y.; Peng, S.; Lin, C.; Long, X.; Komura, T.; Wang, W.: Gen6d: Generalizable model-free 6-dof object pose estimation from RGB images, In European Conference on Computer Vision, Oc tober, 2022, 298-315. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19824-3_18

[3] Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 11534-11542. 10.1109/CVPR42600.2020.01155

[4] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014. https://doi.org/10.48550/arXiv.1409.1556

[5] Brachmann, E.; Krull, A.; Michel, F.; Gumhold, S.; Shotton, J.; Rother, C.: Learning 6d object pose estimation using 3d object coordinates, In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13, 2014, 536-551. Springer International Publishing. https://doi.org/10.1007/978-3-319-10605-2_35