www.cad-journal.net
Computer-Aided Design AND APPLICATIONS

Robust Object 6D Pose Estimation Under High Dynamic Ambient Light

Lei Lu[1], Jiahe Zhu[2], Wei Pan[3], Haojun Zhang[4] , Zhilong Su[5] , Qinghui Zhang[6] , Wanxing Zheng[7] , Ge Gao[8] and Peng Li[9]

[1] Henan University of Technology, lulei@haut.edu.cn
[2] Henan University of Technology, 15137399020@163.com
[3]Department of R&D, OPT Machine Vision Tech Co., Ltd., vpan@foxmail.com
[4]Henan University of Technology, zhj@haut.edu.cn
[5]Shanghai University, zhilong8845@shu.edu.cn
[6]Henan University of Technology, zqh131@163.com
[7]Henan University of Technology, wanxingzheng@haut.edu.cn
[8]Mech-Mind Robotics Technologies Ltd., gao.ge@mech-mind.net
[9]Henan University of Technology, lipeng@haut.edu.cn

Corresponding authors: Wei Pan, vpan@foxmail.com; Peng Li, lipeng@haut.edu.cn

Abstract. This paper addresses the challenges in object 6D pose estimation caused by illumination changes, proposing an improved Gen6D method for robust operation under high dynamic ambient light conditions. Our approach first utilizes a convolutional neural network (CNN) for 2D object detection. A channel attention mechanism is then integrated to promote inter-channel information exchange and reduce noise, leading to more robust feature representations. We employ a fast image-matching algorithm for initial pose estimation, followed by a 3D CNN to refine the pose. Experimental results on a challenging object pose dataset demonstrate that our approach achieves significantly improved accuracy and robustness under complex and high dynamic lighting conditions.

Keywords: pose estimation, high dynamic ambient light, attention mechanism, convolutional neural network
DOI: https://doi.org/10.14733/cadaps.2026.411-427

1 INTRODUCTION

In recent years, the substantial reduction in the cost of ToF (Time of Flight) cameras, structured light cameras, and LiDAR has significantly advanced applications in autonomous driving [1-4], industrial robotics [5-8], and augmented reality [9-12]. Object 6D pose estimation is one of the key tasks to determine the orientation and location of an object in 3D space. The 6D pose of the object is a rigid transformation from the object coordinate system to the camera coordinate system, which can be described by the 3D translation vector T and rotation matrix R [13-15]. Traditional object pose estimation methods use geometric and feature-based approaches to determine object translation and rotation, relying on matching predefined features or shapes to predict object 6D pose [16-23]. However, traditional methods rely heavily on hand-crafted feature extractors and algorithms, which may suffer from low robustness when faced with variations in object types and environmental conditions. In recent years, deep learning-based 6D pose estimation models have made significant progress[24-26]. These models are able to learn features from large datasets, adapt to complex environments, and improve algorithm stability. However, the accuracy of deep learning-based 6D pose estimation models remain challenged by illumination variations and shadow effects under high dynamic ambient light conditions[27-30]. In response to this need, this paper proposes a robust algorithm based on Gen6D[31] for object 6D pose estimation in environments with high dynamic ambient illumination conditions.

First, the feature extraction ability of traditional Gen6D is analyzed; then, we found that the traditional Gen6D model performs poorly under high dynamic ambient light conditions. To overcome the impact of illumination variations on 6D pose estimation caused by high dynamic ambient light, we introduce an enhanced Gen6D 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.

This paper begins by reviewing the related works on object pose estimation and attention mechanisms. Next, we provide a detailed introduction to our proposed method, VGG-ECA, highlighting the integration of the ECA (efficient channel attention) module within the VGG architecture to address the 6D pose estimation problem under high dynamic ambient light conditions. Finally, we present experimental results that demonstrate the effectiveness of our proposed method. This paper makes the following contributions: (1) We propose integrating an efficient channel attention (ECA) module into the Gen6D pipeline to enhance feature representation; (2) We developed a fast and effective image matching algorithm that improves robustness to varying lighting conditions; (3) We demonstrate, through extensive experiments, significant performance improvements on challenging datasets.

2 RELATED WORKS

2.1 Gen6D

With the widespread application of deep learning technology, object 6D pose estimation based on deep learning has made great progress. For instance, Rad et al. [25] trained a large amount of 3D model data and adjusted relevant parameters more comprehensively, making pose estimation more robust and accurate. Therefore, rigid body pose estimation based on deep learning has gradually become a hot issue in the field of computer vision. Some scholars have carried out continuous research on the problem of rigid body pose estimation. Among them, Wang et al. proposed related instance-level rigid body pose estimation [32] and rigid body pose tracking [33]. Brachmann et al. [34, 35, 36] aim to provide a general pose estimation system and solve rigid body pose estimation problems based on random forests. The team considered the impact of input types in practical applications and changed the research direction from RGB-D-based, the rigid body pose estimate is converted into an RGB-based rigid body pose estimate. Hodan et al. [37] mainly studied the challenges faced by rigid body pose estimation, including the impact of rigid body weak texture and rigid body symmetry, and also proposed the T-LESS data set to solve weak texture rigid body pose estimation. Hu et al. [38,39] mainly study rigid body pose estimation methods based on key points, and take end-to-end implementation as the main direction of their research work. Pham et al. [40, 41] conducted research on the sub-task of rigid body pose estimation, proposed a method for processing point cloud information in 2019, and proposed a method for obtaining 2D-3D corresponding information in 2020. Gen6D is a general model-free 6D pose estimator that does not require high-quality object models, additional depth maps, or object masks during testing, making it particularly suitable for rapid experimental analysis [31]. Gen6D consists of three main stages: the detector, the selector, and the refiner. The detector is responsible for identifying the object's local region and estimating its depth based on the size of this region (translation). The selector then provides a rough estimate of the object's perspective (rotation) by selecting the reference image most similar to the query image. Finally, the refiner takes these initial estimates of translation and rotation and refines them into precise values. However, experiments have shown that the Gen6D deep learning model performs poorly in pose estimation under high dynamic ambient light conditions. In such environments, key factors, including illumination variations, shadow effects, and color distortion, significantly impair the model's pose estimation performance. The research presented in this article serves as an extension of the Gen6D model, wherein we introduce an ECA module into the VGG architecture to enhance object pose estimation under challenging illumination conditions. This improvement allows the Gen6D model to more accurately estimate object poses in complex illumination scenarios.

2.2 Attention Mechanism

Attention mechanisms are currently widely utilized in various fields, including machine translation, speech recognition, image captioning, and image restoration [43]. Their popularity stems from the ability to enhance the model's discriminative power. For instance, in machine translation and speech recognition, attention mechanisms assign different weights to each word in a sentence, allowing neural network models to learn more flexibly. In other words, the attention mechanism enables the model to allocate varying weights to different parts of the input, facilitating the extraction of key information [44]. This capability allows for more accurate predictions without incurring additional computational overhead during model inference.

More specifically, the attention mechanism works by dynamically adjusting the weights of different parts of the input, giving higher priority to the most critical information. This adaptability allows neural networks to better capture long-range dependencies and relationships within the data. For example, in machine translation, the model can attend to the words in a sentence that carry the most significant semantic meaning, regardless of their position within the sentence. Similarly, in speech recognition, attention mechanisms enable the system to focus on the most important phonemes or syllables, even if they are temporally distant from one another. This enhanced focus on relevant information leads to more precise predictions without introducing high additional computational costs during model inference. As a result, the attention mechanism provides a substantial improvement in both the performance and efficiency of neural network-based models.

The Efficient Channel Attention (ECA) module, a specialized technique building on attention mechanisms, improves the performance of deep neural networks in image processing tasks. By reweighting channels, ECA enables the model to prioritize the most informative features while suppressing irrelevant ones, which is crucial for handling complex visual inputs like images with fluctuating lighting, reflections, or shadows [45]. This capability allows the network to adaptively identify the most relevant channels, improving its ability to distinguish between essential and non-essential data. ECA is particularly effective in dynamic environments where environmental conditions are unpredictable, such as in object pose estimation tasks where accurate orientation identification is critical. By focusing the model's attention on the most relevant features, ECA significantly enhances the performance of deep learning models, especially in challenging settings. Building on this, this paper integrates the ECA module into the VGG architecture of Gen6D, boosting its performance for object pose estimation under high dynamic ambient light conditions.

Building on this foundation, this paper proposes the integration of the ECA module into the VGG architecture of Gen6D, significantly enhancing the performance of the Gen6D deep learning model in object pose estimation under high dynamic ambient light conditions. The motivation behind this research is to address a critical challenge: existing methods for object pose estimation often falter under varying illumination conditions. This drives our investigation into utilizing ECA within the Gen6D framework. By doing so, we aim to provide a more robust solution that improves performance across different lighting conditions, thereby offering a significant enhancement to the accuracy and reliability of deep learning models in practical applications.

3 METHODOLOGY

3.1 Problem Statement

Given an RGB image in a high-dynamic ambient light scene, the goal of this paper is to estimate the 6D pose of the object correctly. Since the 6D pose is estimated from the camera image, the 6D pose here refers to the 6D pose from the object coordinate system O to the Rigid transformation of C in the camera coordinate system. The 6D pose of the object can be expressed as a rigid transformation matrix $p \in SE(3)$. Specifically, the rigid transformation matrix p of the 6D pose includes the 3D rotation matrix $R \in SO(3)$ and the 3D transformation matrix $T \in R^3$, $p = \left[R \ | T\right]$, the rotation matrix R is a 3*3 unit orthogonal matrix, which describes the direction of the object coordinate system relative to the camera coordinate system. The dimension of the translation matrix T is 3*1, which describes the origin of the object in the camera coordinate system. Therefore, the dimension of the transformation matrix p of the 6D pose is 3*4. Object 6D pose estimation based on RGB images is known for its high real-time performance and broad applicability. Fully utilizing the color information in RGB images enables efficient feature extraction, making it suitable for dynamic applications such as robotics, augmented reality, and autonomous systems. Object 6D pose estimation from RGB images faces significant challenges, particularly under complex lighting conditions. Illumination changes, shadow effects, occlusion, and sensor noise can distort visual information, obscure key features, or alter object appearance, thereby reducing the robustness and accuracy of the estimation.

High dynamic range environments exhibit substantial illumination variations, including stark contrasts between highlights and shadows. These changes in illumination can alter the appearance of the object in the image, thereby affecting the accuracy of pose estimation. In the transition areas between high light and shadow, shadow effects can alter the surface brightness and texture of the object, impacting the identification and localization of the object by the pose estimation algorithm. The goal of this article is to achieve efficient 6D object pose estimation under high dynamic ambient light conditions while addressing the challenges posed by illumination changes, shadow effects, and color distortion.

3.2 Review of Gen6D

The backbone of Gen6D is the VGG [46] network, which was proposed in 2014 and has gained significant attention in the field of computer vision due to its simple and efficient structure and exceptional performance. VGG employs a uniform 3x3 convolution kernel size and pooling layers, resulting in a deep network architecture capable of effectively extracting abstract features from images. Another notable aspect of the VGG network is its straightforward design, which stacks multiple convolutional and pooling layers to progressively extract high-level semantic information. This is achieved by continuously reducing the size of the feature map while increasing the number of channels. The extracted features are then mapped to category labels through fully connected layers to perform image classification tasks. The VGG network also utilizes the ReLU activation function after each convolutional and fully connected layer to introduce nonlinearity, enhancing the network's expressive power and model accuracy.

Gen6D is primarily divided into three stages: the detector, the selector, and the refiner. The detector is responsible for identifying the object's local area and estimating its depth based on the area size (translation). The selector obtains a rough perspective (rotation) by selecting the reference image most similar to the query image. Finally, the refiner updates the rough estimates of translations and rotations into precise values.

However, despite the strength of this approach, Gen6D faces key challenges under dynamic lighting conditions. During image capture in environments with changing light sources, the reflections on the object surface can vary significantly, which impacts the accuracy of the depth and rotation estimates. The detector may struggle to consistently identify the object's local area under such conditions, while the selector might select a reference image that doesn't accurately represent the query due to lighting inconsistencies. The refiner, although effective under stable conditions, may fail to fine-tune the estimates accurately in dynamic lighting, as it depends heavily on the initial rough estimates that are influenced by lighting variations. This makes it difficult to achieve high precision in real-world applications with fluctuating light environments.

We evaluate the pose estimation performance of the 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 Figure 1.

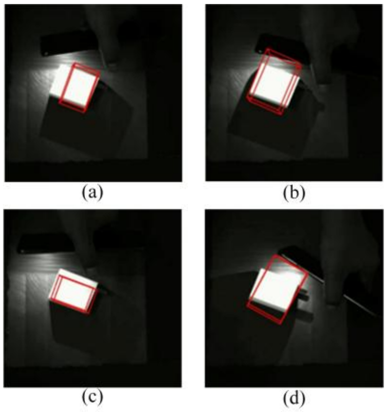


Figure 1: Gen6D yields insufficient pose estimation results under high dynamic ambient light conditions. Panels (a) to (d) illustrate the outcomes of Gen6D in such environments, highlighting the model's poor performance attributed to the challenges posed by these dynamic lighting conditions.

Since Gen6D struggles with estimating object poses under high dynamic ambient light conditions, this article proposes targeted improvements to enhance the model's performance in these challenging scenarios.

3.3 Improved Method

Architecture: The ECA [47] module is integrated into the VGG backbone of Gen6D. ECA enhances 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. The overall architecture of the improved model is illustrated in Figure 2.

ECA: To enhance the feature modeling capabilities of the end-to-end learning network without increasing model complexity, this paper introduces the ECA module. This mechanism enables the network to dynamically adjust its attention to different feature channels in the object image, thereby minimizing the interference from redundant information. By focusing on the most relevant features, the ECA module improves the accuracy and robustness of pose estimation, particularly under high dynamic ambient light conditions.

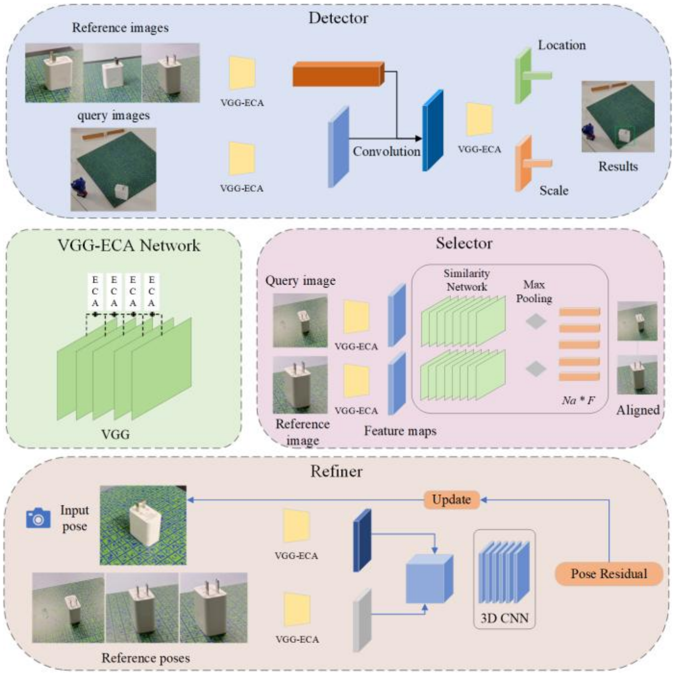


Figure 2: An overview of our proposed method is presented, with the backbone being the VGG-ECA network. Our architecture comprises three main components: the object detection module (detector), the rotation estimation module (selector), and the pose refinement module (refiner). All three modules leverage the VGG-ECA backbone network.

The ECA model is a computational framework based on the attention mechanism, which primarily utilizes attention to encode the input signal, ultimately producing an output with relevant semantic information. The implementation of the attention mechanism involves several key steps: 1. Perceiving and identifying the input information; 2. Screening and filtering the information to select task-relevant data; and 3. Processing and integrating the information to prioritize relevant data, thereby forming specific semantic information.

The attention mechanism ECA is a type of channel attention. This algorithm is an improvement based on the SE [42] algorithm. Although the SE algorithm reduces the complexity of the model through fully connected dimensionality reduction, it disrupts the direct correspondence between weights and channels. By first reducing the dimension and then increasing it, the SE algorithm leads to an indirect correspondence between the weights and the channels. To address the shortcomings of the SE algorithm, the ECA algorithm proposes a one-dimensional convolution method to avoid the negative impact of dimensionality reduction on the data. The structure of the ECA algorithm is shown in Figure 3.

The structure of the ECA algorithm primarily consists of the following three aspects:

1) The feature map is compressed, transforming its size from (N, C, H, W) to (N, C, 1, 1) through global average pooling, thereby facilitating the fusion of global contextual information. This step aligns with the approach used in the SE algorithm. In this context, N represents the batch size, C denotes the number of channels, and H and W refer to the spatial height and width of the feature map, respectively. Specifically, in step 1) of the ECA module, global average pooling is applied across each channel to generate a channel descriptor of shape (N, C, 1, 1), preserving the number of channels while eliminating spatial dimensions to capture global contextual features.

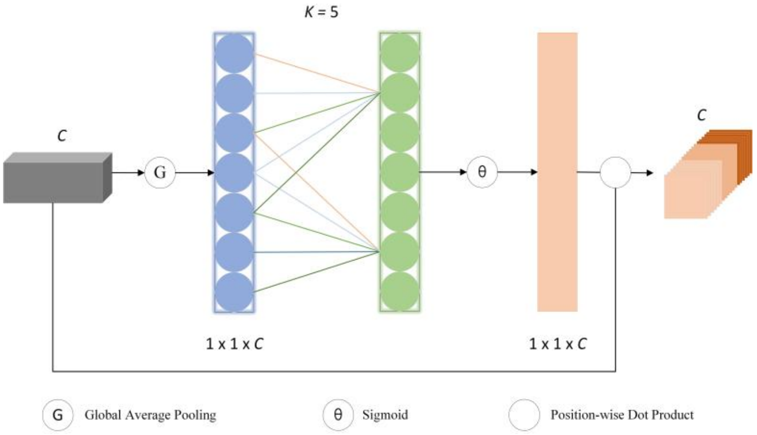


Figure 3: The ECA module enhances useful features and suppresses irrelevant ones by adaptively weighting the features along the channel dimension, thereby improving the network's representation capability and generalization performance.

2) Calculate the size of the adaptive convolution kernel, where C is the number of input channels, b=1, γ=2, and one-dimensional convolution is used to calculate the weight of the channel, and finally, the Sigmoid activation function is used to map the weight between (0-1).

3) The reshaped weight values are multiplied by the original feature map (utilizing Python's broadcasting mechanism) to obtain feature maps with varying weights.

VGG-ECA: 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 dataset, as shown in Figure 4.

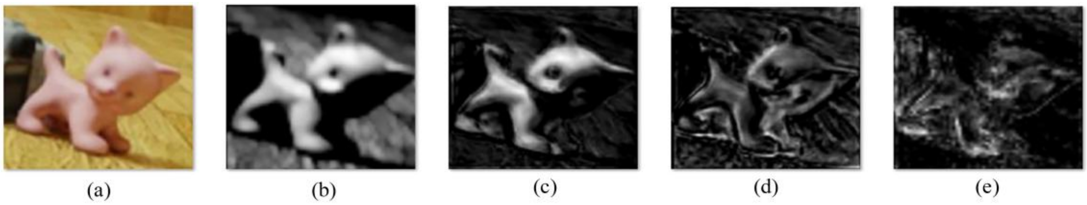

(a) (b) (c) (d) (e)

Figure 4: 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 Figure 5.

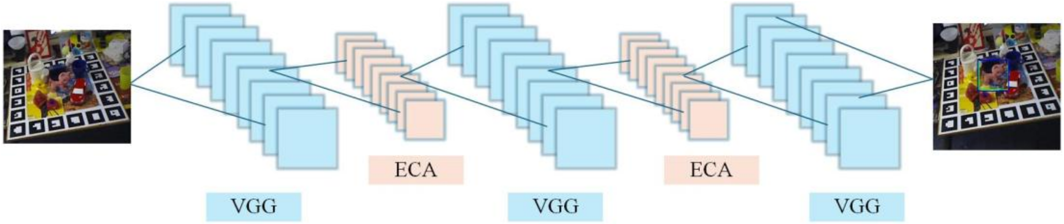


Figure 5: The architecture of VGG-ECA, which integrates the Efficient Channel Attention (ECA) module into the standard VGG network. After the convolutional layers extract spatial features, a global average pooling operation is applied to each channel, producing a compact feature descriptor. The ECA module then computes channel-wise attention weights using a lightweight 1D convolution, enabling the network to emphasize informative features without increasing model complexity. This process enhances the network's ability to capture key features under challenging conditions such as high dynamic ambient lighting, thereby improving the accuracy of object 6D pose estimation.

Compared to Gen6D, we have better performance in the stability of object detection in two-dimensional images under high dynamic ambient light conditions. This paper decomposes the detection problem 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, as shown in Figure 6. 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, 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.

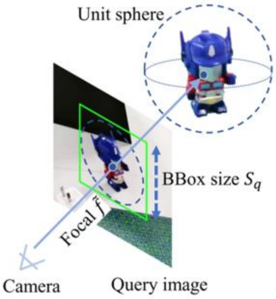


Figure 6: Translation information calculation module.

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.

4 EXPERIMENT

Datasets: 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. The GenMOP dataset consists of 10 objects (as shown in Figure 7), with a range of shapes, from flat objects (such as "scissors") to structural items (such as "mobile phone chargers") [31]. 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.

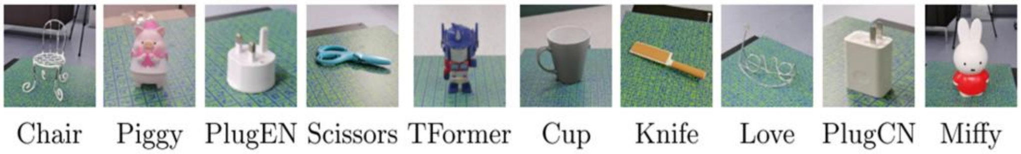


Figure 7: Objects in the GenMOP dataset.

The dataset was collected under various illumination conditions to assess the model's robustness in real-world scenarios. Specifically, two video sequences of the same object were captured in different environments, each sequence containing approximately 200 images. For each sequence, we utilized COLMAP to separately reconstruct the camera poses, ensuring that both sequences were properly aligned. Additionally, key points on the objects were manually marked to facilitate cross-sequence alignment and ensure the accuracy of the evaluation.

Metrics: The metrics used in this study, including Average Distance of Model Points (ADD) 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.

These metrics were chosen because they capture different aspects of pose estimation performance, reflecting both spatial and visual alignment. In real-world applications, achieving a higher ADD value translates to improved spatial precision, which is critical for tasks such as robotic grasping or precise alignment in manufacturing systems. Similarly, a lower Prj-5 error ensures that the estimated pose aligns well with the visual input. By optimizing these metrics, the proposed method demonstrates its practical utility and robustness under diverse conditions.

In calculating ADD, we consider a range of 10% of the object diameter (ADD-0.1d) and use the area under the curve (AUC) value from 0 to 10 cm (ADD-AUC) to measure recall. For the projection error, we analyze recall at a threshold of 5 pixels (Prj-5). The definition of ADD is as follows:

$$ADD = \frac{1}{m}\sum_{v\varepsilon\vartheta} \| (Rv+T)-(R^{*}v+T^{*}) \| \tag{4.1}$$

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:

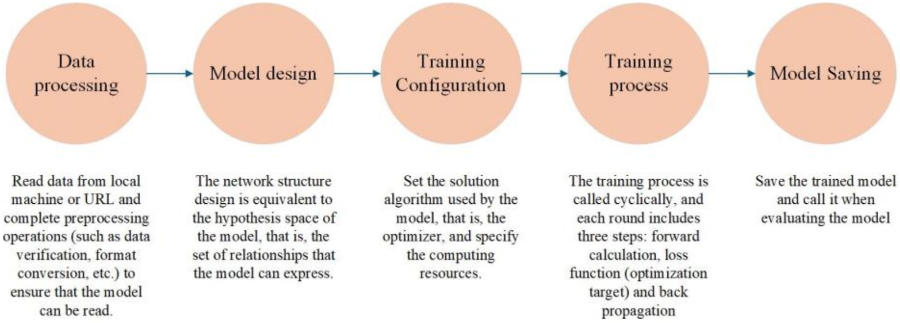
$$di = \| project(Pi, pose_{est}) - project(Pi, pose_{gt}) \| \tag{4.2}$$

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:

$$prj-5=\frac{successful\_samples}{total\_samples} \tag{4.3}$$

A higher Prj-5 index signifies a more accurate pose estimation of the model under 2D projection, indicating superior performance of the algorithm.

Network training: The experiments were conducted using the PyTorch deep learning framework, with a batch size of 8. The Adam optimizer was employed for weight updates, and the model was trained for a total of 3,500 epochs. The initial learning rate was set to 0.001 and was subsequently reduced in accordance with the number of training epochs. The hardware environment for the experiments included an Nvidia GeForce RTX 3070 GPU with 8 GB of VRAM, an Intel Core i7-11800H processor, and 16 GB of RAM, the specific network training process is shown in Figure 8.



Read data from local machine or URL and complete preprocessing operations (such as data verification, format conversion, etc.) to ensure that the model can be read.

The network structure design is equivalent to the hypothesis space of the model, that is, the set of relationships that the model can express.

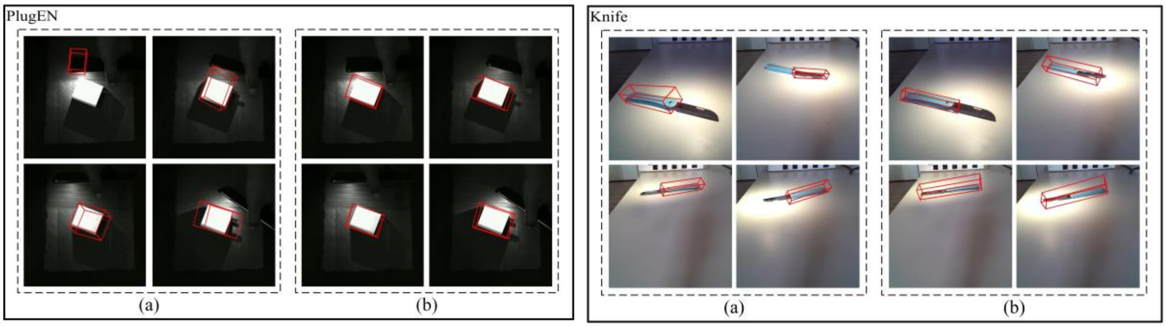
Set the solution algorithm used by the model, that is, the optimizer, and specify the computing resources.

The training process is called cyclically, and each round includes three steps: forward calculation, loss function (optimization target) and back propagation

Save the trained model and call it when evaluating the model

Figure 8: Network training flow chart.

Comparative experiment: This paper presents a object 6D pose estimation model built upon the Gen6D deep learning framework, integrating a convolutional neural network with a channel attention mechanism. 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 RGB images mentioned here were separately collected and assessed under high dynamic ambient light conditions, distinct from the two sequences used for pose reconstruction with COLMAP. The experimental results are visualized as 3D bounding boxes, as illustrated in Figure 9. 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.



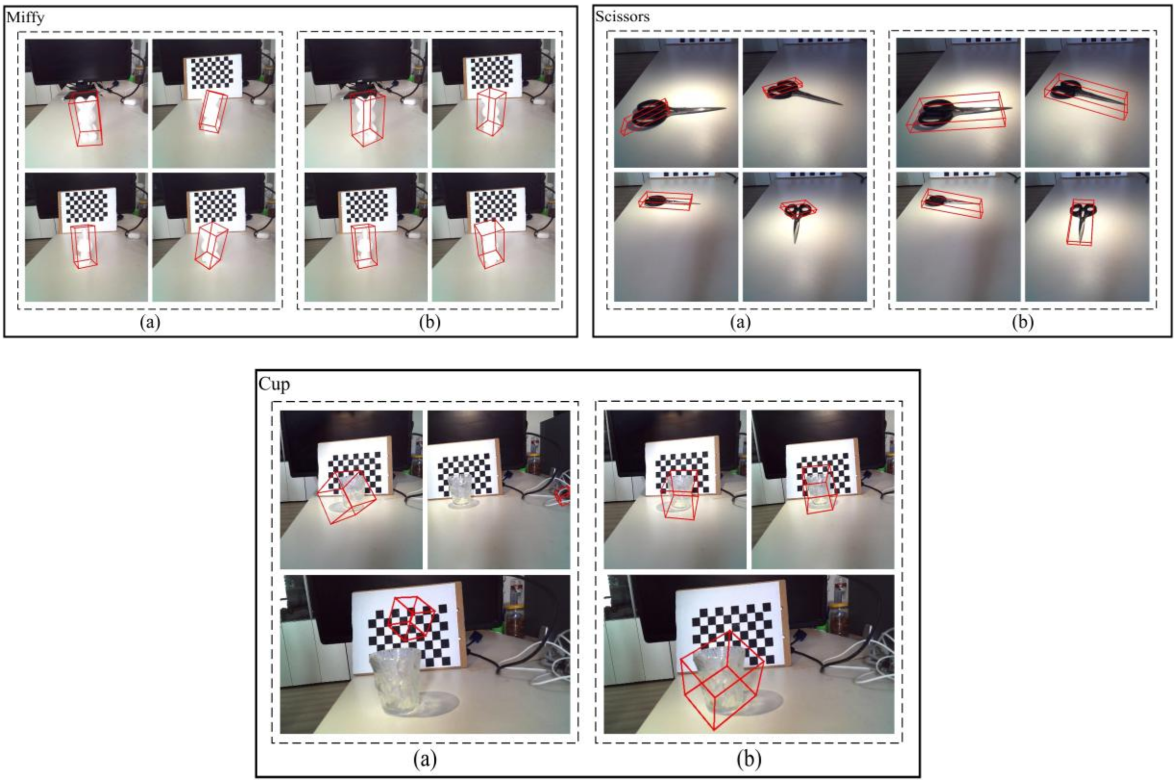


Figure 9: Comparison of pose estimation results: (a) results from the original Gen6D model; (b) results from the proposed method.

The experimental results in part (b) of Figure 9 demonstrate that the introduction of the ECA significantly mitigates the effects of high dynamic ambient light on object pose estimation. This enhancement improves the model's ability to filter key image information and detect important features, ultimately boosting pose estimation accuracy. In contrast, part (a) reveals that the original Gen6D model struggles to accurately estimate the poses of objects like plug_en, knife, miffy, and scissors. Affected by varying ambient light conditions, Gen6D only identifies weakly reflective areas, such as the plastic handles of the knife and scissors, while failing to detect metal parts due to high surface reflectivity. However, as illustrated in part (b), the improved model effectively captures the overall pose of the object under test, reducing the impact of challenging illumination conditions. In addition, strong light exposure in the upper left corner of the image for part (b) of the knife causes the side-viewed blade to appear black, resulting in the improved model's inability to detect the 6D pose of certain blade sections.

For additional rigor, we employed a glass cup instead of a standard mug, further assessing model performance. Part (a) shows that the transparent and high reflective nature of the glass cup leads to failure in object detection by Gen6D. In contrast, part (b) indicates that the enhanced model successfully identifies and detects the pose of the glass cup.

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 Figure 10, 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.

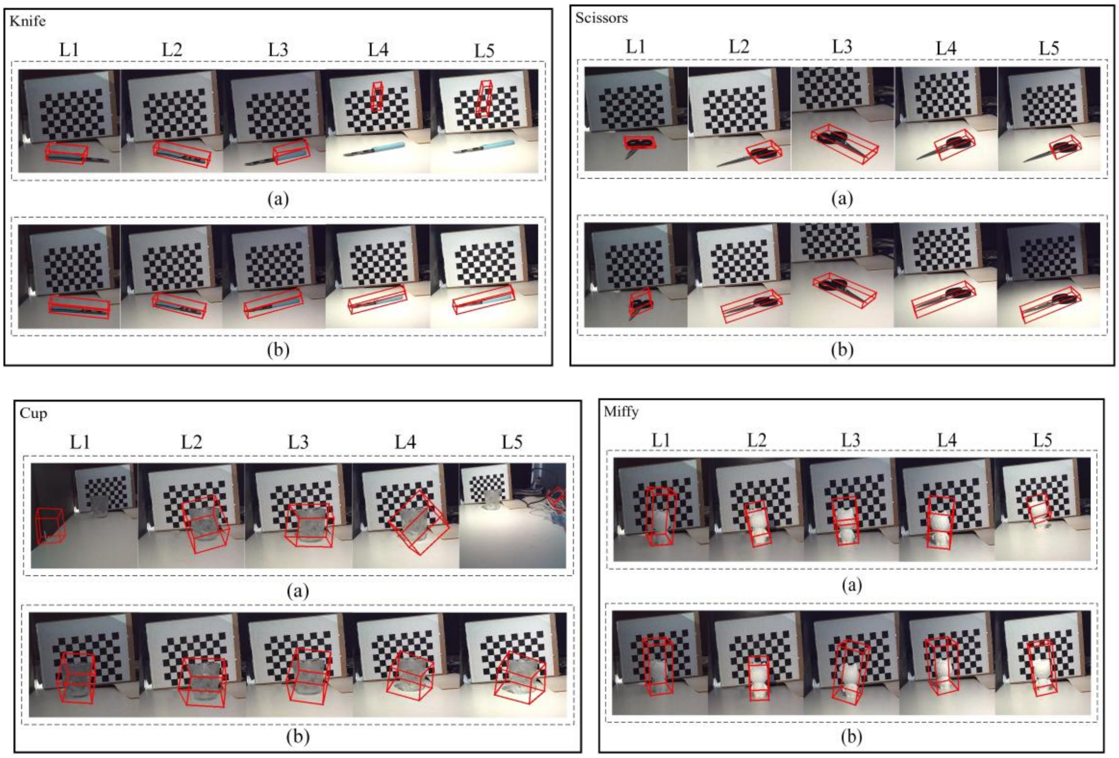


Figure 10: Pose estimation results under five ambient light levels, from dark to bright: (a) results from the original Gen6D model; (b) results from the proposed method.

The images in Figure 10 show a clear comparison between the original and improved models under varying 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, especially in low-light scenarios where the lack of sufficient detail leads to misalignments, and in high-light scenarios where reflections and glare distort the surface features.

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. This is a notable improvement over the original Gen6D model, where reflective surfaces were a significant challenge. However, some failure cases remain, such as in the low-light conditions (L1, L2) for the scissors. Even after model improvement, the reflective properties of the scissors' metal surface continue to cause misalignment in the pose estimation.

Additionally, the pose estimation performance of our method is compared with that of the original deep learning model Gen6D using the official GenMOP dataset. Table 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[31]	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[31]	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

Table 1: Performance on the GenMOP dataset.

To evaluate the efficiency of our proposed method, we compare its computation time with the original approach. All experiments were conducted on the following hardware and software environment:

Hardware Configuration: Intel Core i7-11800H CPU, NVIDIA RTX 3070 GPU, 16GB RAM
Operating System: Ubuntu 20.04
Programming Language and Framework: Python 3.8, PyTorch 1.8.1, CUDA 11.1

The comparison of inference time of network models is shown in Table 2.

Model	Inference Time (seconds)	Time Increase (seconds)
Original Model (VGG)	0.4965	-
Modified Model (VGG-ECA)	0.5322	+0.0357

Table 2: Inference time comparison of Network Models.

The modified network with ECA modules resulted in an increase of 35.7 milliseconds (7.19%) in inference time compared to the original model.

5 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.

Future work could focus on extending the proposed method to a broader range of objects and datasets to improve its generalization across different domains. Further investigation into other attention mechanisms, such as spatial attention or hybrid attention models, could also be explored to enhance pose estimation accuracy under complex lighting conditions. Additionally, incorporating real-time performance evaluation and testing the model in real-world industrial applications, such as robotic grasping or automated inspection, would be valuable to assess its practicality and robustness. Finally, addressing the limitations of the current method, such as its dependency on specific dataset characteristics or potential computational complexity, will be important in future research to improve scalability and efficiency.

FUNDING

This research was funded by the National Natural Science Foundation of China [62375078, 62073123]; the Key Research Project Plan for Higher Education Institutions in Henan Province [24ZX011]; and the Cultivation Programme for Young Backbone Teachers in Henan University of Technology; and the key research and development program of Henan Province[231111222100];

and the Open Project of the Institute of Complexity Science, Henan University of Technology[CSKFJJ-2024-3].

Lei Lu, https://orcid.org/0000-0002-3050-6542
Wei Pan, https://orcid.org/0000-0002-0933-2453
Peng Li, https://orcid.org/0009-0001-8437-7408

REFERENCES

[1] Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G.: A survey of deep learning techniques for autonomous driving, Journal of field robotics, 37(3), 2020, 362-386. https://doi.org/10.1002/rob.21918

[2] Levinson, J.; Askeland, J.; Becker, J.; Dolson, J.; Held, D.; Kammel, S.; Thrun, S.: Towards fully autonomous driving: Systems and algorithms, 2011 IEEE intelligent vehicles symposium (IV) , IEEE , 2011, 163-168. https://doi.org/10.1109/IVS.2011.5940562.

[3] Maurer, M.; Gerdes, J. C.; Lenz, B.; Winner, H.: Autonomous driving: technical, legal and social aspects, Springer Nature, 2016. https://doi.org/10.1007/978-3-662-48847-8

[4] Wang, J.; Liu, J.; Kato, N.: Networking and communications in autonomous driving: A survey, IEEE Communications Surveys & Tutorials, 2018, 1243-1274. https://doi.org/10.1007/978-3-662-48847-8.

[5] Bousmalis, K.; Irpan, A.; Wohlhart, P.; Bai, Y.; Kelcey, M.; Kalakrishnan, M.; Vanhoucke, V.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping, 2018 IEEE international conference on robotics and automation (ICRA), IEEE, 2018, 4243-4250. https://doi.org/10.1109/ICRA.2018.8460875

[6] James, S.; Wohlhart, P.; Kalakrishnan, M.; Kalashnikov, D.; Irpan, A.; Ibarz, J.; Bousmalis, K.: Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 12627-12637.

[7] Morrison, D.; Corke, P.; Leitner, J.: Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach, arXiv preprint arXiv:1804.05172.

[8] Tremblay, J.; To, T.; Sundaralingam, B.; Xiang, Y.; Fox, D.; Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects, arXiv preprint arXiv:1809.10790.

[9] Cipresso, P.; Giglioli, I. A. C.; Raya, M. A.; Riva, G.: The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature, Frontiers in psychology, 2018, 9, 2086. https://doi.org/10.3389/fpsyg.2018.02086.

[10] Gattullo, M.; Scurati, G. W.; Fiorentino, M.; Uva, A. E.; Ferrise, F.; Bordegoni, M.: Towards augmented reality manuals for industry 4.0: A methodology, robotics and computer-integrated manufacturing, 2019, 276-286. https://doi.org/10.1016/j.rcim.2018.10.001

[11] Ibáñez, M. B.; Delgado-Kloos, C.: Augmented reality for STEM learning: A systematic review, Computers & Education, 2018, 109-123. https://doi.org/10.1016/j.compedu.2018.05.002

[12] Peddie, J.: Augmented reality: Where we will all live (Vol. 349), Cham: Springer, 2017. https://doi.org/10.1007/978-3-031-32581-6.

[13] He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; Sun, J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 11632-11641.

[14] Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes, arXiv preprint arXiv:1711.00199.

[15] Brachmann, E.; Krull, A.; Michel, F.; Gumhold, S.; Shotton, J.; Rother, C.: Learning 6D object pose estimation using 3d object coordinates, Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Springer Inte

rnational Publishing, Part II 13, 2014, 536-551. https://doi.org/10.1007/978-3-319-10605-2_35

[16] DeMenthon, D. F.; Davis, L. S.: Model-based object pose in 25 lines of code. International journal of computer vision, 1995, 123-141. https://doi.org/10.1007/BF01450852

[17] Nistér, D.: An efficient solution to the five-point relative pose problem, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 756-770. https://doi.org/10.1109/TPAMI.2004.17.

[18] Tsai, R.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, IEEE Journal on Robotics and Automation, 1987, 323-344. https://doi.org/10.1109/JRA.1987.1087109.

[19] Schweighofer, G.; Pinz, A.: Robust pose estimation from a planar target, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 2024-2030. https://doi.org/10.1109/TPAMI.2006.252.

[20] Hinterstoisser, S.; Benhimane, S.; Navab, N.: N3m: Natural 3D markers for real-time object detection and pose estimation, 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, 1-7. https://doi.org/10.1109/ICCV.2007.4409004.

[21] Lowe, D. G.: Object recognition from local scale-invariant features, Proceedings of the seventh IEEE international conference on computer vision, IEEE, 1999, 1150-1157. https://doi.org/10.1109/ICCV.1999.790410.

[22] Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.: ORB: An efficient alternative to SIFT or SURF, 2011 International Conference on Computer Vision, IEEE, 2011, 2564-2571. https://doi.org/10.1109/ICCV.2011.6126544.

[23] Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes, Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. Springer Berlin Heidelberg, 2013, 548-562. https://doi.org/10.1007/978-3-642-37331-2_42

[24] Rad, M.; Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth, Proceedings of the IEEE International Conference on Computer Vision, 2017, 3828-3836. https://doi.org/10.1109/ICCV.2017.413

[25] Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; Navab, N.: Ssd-6d: Making rgb-based 3d detection and 6D pose estimation great again, Proceedings of the IEEE International Conference on Computer Vision, 2017, 1521-1529. https://doi.org/10.1109/ICCV.2017.169

[26] Tekin, B.; Sinha, S. N.; Fua, P.: Real-time seamless single shot 6d object pose prediction, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018,292-301. https://doi.org/10.1109/CVPR.2018.00038

[27] Lugo, G.; Hajari, N.; Cheng, I.: Semi-supervised learning approach for localization and pose estimation of texture-less objects in cluttered scenes, Array, 16, 100247, 2022. https://doi.org/10.1016/j.array.2022.100247

[28] Li, C. H. G.; Wu, J. T.: Deep learning approaches for improving robustness in real-time 3D-object positioning and manipulation in severe lighting conditions, The International Journal of Advanced Manufacturing Technology, 2023, 129(9), 3829-3847. https://doi.org/10.1007/s00170-023-12497-5

[29] Bauer, D.; Hönig, P.; Weibel, J. B.; García-Rodríguez, J.; Vincze, M.: Challenges for monocular 6d object pose estimation in robotics, IEEE Transactions on Robotics, 2024. https://doi.org/10.1109/TRO.2024.3433870.

[30] Moonen, S.; Vanherle, B.; de Hoog, J.; Bourgana, T.; Bey-Temsamani, A.; Michiels, N.: Cad2render: A modular toolkit for gpu-accelerated photorealistic synthetic data generation for the manufacturing industry, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, 583-592. https://doi.org/10.48550/arXiv.2211.14054.

[31] 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, European Conference on Com

puter Vision, Cham: Springer Nature Switzerland, 2022, 298-315. https://doi.org/10.1007/978-3-031-19824-3_18.

[32] Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 3343-3352.

[33] Wang, C.; Martín-Martín, R.; Xu, D.; Lv, J.; Lu, C.; Fei-Fei, L..; Zhu, Y.: 6-pack: Category-level 6d pose tracker with anchor-based key points, 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, 10059-10066. https://doi.org/10.1109/ICRA40945.2020.9196679.

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

[35] Krull, A.; Michel, F.; Brachmann, E.; Gumhold, S.; Ihrke, S.; Rother, C.: 6-dof model based tracking via object coordinate regression, Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12, Springer International Publishing, 2015, 384-399. https://doi.org/10.1007/978-3-319-16817-3_25

[36] Brachmann, E.; Michel, F.; Krull, A.; Yang, M. Y.; Gumhold, S.: Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, 3364-3372. https://doi.org/10.1109/CVPR.2016.366

[37] Hodan, T.; Haluza, P.; Obdržálek, Š.; Matas, J.; Lourakis, M.; Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects, 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, 880-888. https://doi.org/10.1109/WACV.2017.103.

[38] Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M.: Segmentation-driven 6d object pose estimation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 3385-3394.

[39] Hu, Y.; Fua, P.; Wang, W.; Salzmann, M.: Single-stage 6d object pose estimation, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, 2930-2939. https://doi.org/10.1109/CVPR42600.2020.00300

[40] Pham, Q. H.; Nguyen, T.; Hua, B. S.; Roig, G.; Yeung, S. K.: JSIS3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 8827-8836.

[41] Pham, Q. H.; Uy, M. A.; Hua, B. S.; Nguyen, D. T.; Roig, G.; Yeung, S. K.: Lcd: Learned cross-domain descriptors for 2d-3d matching, Proceedings of the AAAI conference on artificial intelligence, 2020, (Vol. 34, No. 07, 11856-11864). https://doi.org/10.1609/aaai.v34i07.6859.

[42] Hu, J.; Shen, L.; Sun, G.: Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[43] Ghandi, T.; Pourreza, H.; Mahyar, H.: Deep learning approaches on image captioning, A review, ACM Computing Surveys, 56(3), 2023, 1-39. https://doi.org/10.1145/3617592

[44] Lv, H.; Chen, J.; Pan, T.; Zhang, T.; Feng, Y.; Liu, S.: Attention mechanism in intelligent fault diagnosis of machinery, A review of technique and application, Measurement, 2022, 199, 111594. https://doi.org/10.1016/j.measurement.2022.111594

[45] Xue, M.; Chen, M.; Peng, D.; Guo, Y.; Chen, H.: One spatio-temporal sharpening attention mechanism for light-weight YOLO models based on sharpening spatial attention, Sensors, 21(23), 2021, 7949. https://doi.org/10.3390/s21237949

[46] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image rec ognition, 2014, arXiv preprint arXiv:1409.1556.

[47] Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; & Hu, Q.: ECA-Net: Efficient channel attentio n for deep convolutional neural networks, In Proceedings of the IEEE/CVF conference on c omputer vision and pattern recognition, 2020, 11534-11542.