mathematics

Article

EdgePose: An Edge Attention Network for 6D Pose Estimation

Qi Feng, Jian Nong and Yanyan Liang *

School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China; 1909853rii30003@student.must.edu.mo (Q.F.); 2009853xii30001@student.must.edu.mo (J.N.)
* Correspondence: yyliang@must.edu.mo

Abstract: We propose a 6D pose estimation method that introduces an edge attention mechanism into the bidirectional feature fusion network. Our method constructs an end-to-end network model by sharing weights between the edge detection encoder and the encoder of the RGB branch in the feature fusion network, effectively utilizing edge information and improving the accuracy and robustness of 6D pose estimation. Experimental results show that this method achieves an accuracy of nearly 100% on the LineMOD dataset, and it also achieves state-of-the-art performance on the YCB-V dataset, especially on objects with significant edge information.

Keywords: 6D pose estimation; edge attention; feature fusion; deep learning; mixed reality

MSC: 68T07; 68T45

1. Introduction

In recent years, 3D vision tasks have emerged as a significant area of research, with 6D pose estimation technology garnering considerable attention due to its wide range of applications, including robotics, autonomous driving, scene understanding, and augmented reality. The problem of estimating the 6D pose, which involves determining the three-dimensional position and orientation of an object, has a rich history of research and has given rise to various methodological approaches. These approaches can be broadly categorized into template-based, correspondence-based, and voting-based methods.

The input data for these methods typically include RGB images, which provide rich texture information, and depth images, which supply geometric details. Depth images have been shown to enhance pose estimation accuracy, leading to a growing trend of using RGB-D data as input for 6D pose estimation tasks. However, a significant challenge in this domain is the requirement for large amounts of training data, which becomes increasingly difficult to obtain as model complexity increases. This has spurred interest in methods that maximize the extraction of useful information from existing data.

In response to these challenges, we propose a novel 6D pose estimation method that introduces an edge-aware attention mechanism within a bidirectional feature fusion network. Our method constructs an end-to-end network model by sharing encoder weights between the edge detection network and the RGB branch of the feature fusion network. This joint training approach ensures that edge features are effectively integrated into the pose estimation process, enhancing accuracy, particularly in challenging scenarios involving complex backgrounds or partial occlusions.

Our method's innovative edge-aware approach, combined with the bidirectional fusion of RGB and depth information, leads to significant improvements in pose estimation accuracy over existing baseline models. We validate the effectiveness of our approach through comprehensive evaluations on the LineMOD and YCB-V datasets, where our method outperforms several state-of-the-art techniques in key metrics, demonstrating its robustness and generalizability.

check for updates

Citation: Feng, Q.; Nong, J.; Liang, Y. EdgePose: An Edge Attention Network for 6D Pose Estimation. *Mathematics* **2024**, *12*, 2607. https://doi.org/10.3390/math12172607

Academic Editors: Heui Seok Lim, Sanghyuk Lee, Yeongwook Yang and Imatitikua Aiyanyo

Received: 30 June 2024
Revised: 17 August 2024
Accepted: 21 August 2024
Published: 23 August 2024

The remainder of this paper is structured as follows: we first review recent advancements in 6D pose estimation, categorizing related work based on input modalities (RGB-based and RGB-D-based) and methodological approaches (template-based, correspondence-based, and voting-based). We then detail our proposed edge-aware 6D pose estimation method, followed by a presentation of our experimental results and conclusions.

2. Related Works

Pose estimation has a wide range of applications, and research includes human pose estimation [1,2], hand gesture estimation [3,4], object pose estimation, etc. We study the 6D pose estimation of objects. 6D pose estimation can be divided into two categories according to the input type: 2D RGB image-based and 3D image-based (depth map and point cloud). Before the popularization of 3D images, most methods were based on RGB images. RGB images provide image information for the target object, but their performance is easily affected by illumination changes, complex backgrounds, and textureless object surfaces.

Since the advent of depth sensors such as Microsoft Kinect, depth maps and point clouds have been widely used in computer vision tasks, and a large number of RGB-D-based 6D pose estimation research works have also emerged (see Table 1).

Table 1. Methods Categorize.

Categories	Correspondence-Based	Template-Based	Voting-Based
RGB-based	[5–19]	[20–25]	[26,27]
RGBD-based	[28–36]	[37,38]	[39–43]

2.1. RGB-Based Methods

2.1.1. Template-Based Methods

The method proposed in [24] is an important representative of template-based methods. This method improves detection speed and accuracy by automatically generating templates from CAD 3D models and using 3D models to obtain the pose estimation of the object. Although the viewpoint sampling and selection method of this method has a greater impact on the results, resulting in slightly lower accuracy than recent methods, as an early attempt, it provides an important direction for subsequent research.

After generating a large number of templates, the method in [25] encodes all the templates to form a codebook instead of directly using the template images as template libraries. This encoding method improves efficiency and accuracy but still relies on the 3D model of the target object, and the codebook contains up to 92,232 entries, which is not conducive to practical applications.

Other related works [20–23] adopt similar template matching methods. Deep-6DPose [20] achieves a direct regression of the pose of 6D objects by decoupling the pose parameters into translation and rotation, making training more feasible. In order to reduce the amount of training data required, DSC-PoseNet [21] effectively alleviates the domain difference between synthetic data and real data by constructing cross-scale self-supervisory signals. In the case of RGB data training, it outperforms the latest methods based on synthetic data training and is on par with fully supervised methods. The PFRL [22] framework performs 6D pose estimation in a Pose-Free manner, designs a reward mechanism based on 2D masks, and uses composite reinforcement optimization rules to effectively learn operation strategies. OSOP [23] uses 2D templates rendered from different perspectives to represent 3D query models for dense feature extraction and matching. Although these methods perform well on synthetic data, there are domain differences between real images and synthetic images, which affects the accuracy of feature extraction.

2.1.2. Correspondence-Based Methods

Correspondence-based methods are a more common class of methods. NOCS [5] jointly estimates the metric 6D pose and size of multiple objects by normalizing the targets to the same space, but the accuracy of the results is too dependent on the results of converting the images to NOCS. Pix2Pose [6] proposes a transformation loss function to handle symmetric objects, uses generative adversarial training to restore occluded parts, and predicts the 3D coordinates of each object pixel through an autoencoder architecture, solving challenges such as occlusion and symmetry, but requires a 3D model with high accuracy. PoseCNN [7] explicitly models the dependencies and independence between pose estimation tasks by decomposing them into different components. Inspired by [7], Convposecnn [44] leverages the pre-trained VGG16 backbone network for feature extraction and uses a fully convolutional architecture to achieve pixel-level quaternion translation prediction, thereby increasing model size and training/inference speed. GDR-Net [8] learns 6D poses in an end-to-end manner from intermediate geometric representations based on dense correspondences by leveraging geometric guidance. SO-Pose [45] is a new deep architecture that regresses 6D poses from two-layer representations of 3D objects. It uses self-occlusion and 2D-3D correspondences to establish a two-layer representation of each 3D spatial object and achieves two-cross-layer consistency. The method in [9] is to recover the pose of a complete 6-DOF object from an RGB image sequence, which helps to capture the pose information of the object more accurately in practical applications. DPOD [10] estimates the dense multi-class 2D-3D correspondence mapping between the input image and the 3D model, thereby achieving end-to-end pose estimation. ZebraPose [11] proposed a novel coarse-to-fine surface encoding method for 6 DoF object pose estimation, assigning a unique descriptor to each 3D vertex. All of these methods rely on the 3D model of the target, which greatly limits the generalization of the model.

Some correspondence-based methods have tried to address this challenge. Some methods use multi-image, some use iterative methods, and some implement the pose estimation of monocular RGB images. DPODv2 [12] introduced NOCS [5] on the basis of DPOP [10], removing the dependence on 3D models. OnePose [13] borrowed the idea of visual positioning. It only needs to perform a simple RGB video scan of the object to build a sparse SfM model of the object and directly query 2D-3D matching through the attention network to achieve pose estimation without relying on CAD models. NeRF-Pose [14] implements a weakly supervised method for 6D object pose estimation in monocular images by first reconstructing the object using implicit neural representation and then training a pose regression network. Reference [15] introduced a new differentiable layer called "Bidirectional Deep Enhanced PnP (BD-PnP)" that can simultaneously satisfy two sets of 2D-3D correspondences. EPro-PnP [16] uses a learnable probabilistic PnP layer to support end-to-end training and incorporate the uncertainty of 2D-3D correspondences. POPE [17] uses pre-trained 2D base models and 3D geometric principles to estimate the relative pose between object cues and target objects, achieving zero-shot object pose estimation in different environments. Gen6D [18] consists of an object detector, a view selector, and a pose optimizer and only requires annotated images of unknown objects to accurately predict poses in any environment. CRT-6D [19] introduces Object Surface Keypoint Features (OSKFs) as a lightweight intermediate 6D pose offset representation for iterative pose refinement. Correspondence-based methods usually use PnP algorithms to solve the final pose, but the timeliness and computational consumption of the algorithm are also challenges.

2.1.3. Voting-Based Methods

Voting-based methods usually appear as an optimization method of template-based or correspondence-based methods. InstancePose [26] uses the compact architecture of convolutional neural networks to improve performance and introduces novel output feature maps such as false object masks, semantic object masks, center vector maps, and 6D coordinate maps to improve the accuracy of 6DoF pose estimation of multiple instance objects in a single RGB image. Reference [27] proposes a segmentation-based 6D object pose estimation framework, which aims to improve the accuracy of rigid object pose estimation in cluttered and occluded scenes, but the pose estimation result depends on the segmentation result. The voting-based method improves the performance of the model but requires a complex network architecture and training process.

2.2. RGBD-Based Methods

With the popularity of depth sensors such as Microsoft Kinect, depth maps have become easy to obtain. Therefore, 6D pose estimation algorithms based on RGB-D images have begun to emerge. Similar to RGB image-based methods, RGB-D-based methods can also be divided into three categories: correspondence-based, template matching-based, and voting-based. Although the development of deep learning has spawned direct regression methods, these methods usually use these three methods indirectly before the regression module. RGB-D-based methods usually outperform the RGB-based methods in the same period, but the computational consumption is also significantly increased.

2.2.1. Template-Based Methods

Template matching methods are also widely used in RGB-D images. SAR-Net [37] effectively infers implicit rotation representations through shape alignment between category-level template shapes and instance point clouds. OVE6D [38] introduces a new object view encoding (OVE) that effectively captures the geometric relationship between the object view and its 3D surface, thereby achieving accurate pose estimation.

2.2.2. Correspondence-Based Methods

RGB-D image-based correspondence methods are very common in 6D pose estimation. G2L-Net [28] uses a hierarchical approach to process RGB-D point cloud data, introduces direction-based point-level embedding vector features and a rotation residual estimator, effectively utilizes view information, and improves the accuracy of rotation prediction. The full-flow bidirectional fusion network proposed by FFB6D [29] combines the information of RGB images and depth images and improves the 3D keypoint selection algorithm to achieve better 6D pose estimation. OnePose++ [30] is based on the LoFTR [46] feature matching method and achieves direct 2D-3D correspondence by reconstructing a semi-dense point cloud model without first detecting keypoints in the query image, thereby performing object pose estimation without a CAD model. RBP-Pose [31] combines residual vectors guided by object pose and shape priors and improves shape diversity during training through a nonlinear shape enhancement scheme while retaining the commonality of geometric features, achieving good performance. The framework in [32] is a self-supervised framework based on deep implicit shape representation (DeepSDF [47]) for category-level 6D object pose estimation. FS6D [33] uses a meta-learning strategy to learn a good initialization for pose estimation and then fine-tunes new objects using limited data. Uni6D [34] is a unified framework that bypasses the separate projection step and eliminates potential sources of error in traditional methods. ES6D [35] proposes a symmetry-based loss function that avoids shape-induced uncertainty through primitive grouping. RNNPose [36] fine-tunes the pose by repeated iterations in each rendering cycle and supervises the accuracy of pose estimation using model alignment loss, correspondence loss, and descriptor loss.

2.2.3. Voting-Based Methods

Voting-based methods are usually complementary to the first two methods. Reference [39] proposed a novel discrete continuous rotation regression method that effectively solves the local optimal problem of rotation blur of symmetrical objects. DenseFusion [40] makes full use of the complementary information of RGB and depth data sources to extract pixel-level dense feature embedding for pose estimation through a dense fusion network. PVNet [48] performs keypoint localization through a pixel-level voting network (PVNet) and pose estimation through uncertainty-driven PnP. Compared with directly regressing keypoint coordinates, it obtains superior performance by predicting vector fields and performing RANSAC-based voting. The method proposed in [41] trained a convolutional autoencoder on random local patches to create a codebook of synthetic model view patches and then used these codebooks to vote for 6D object poses during testing. 6-PACK [42] achieves real-time tracking of new instances of known object categories by learning a small set of 3D keypoints to compactly represent objects. PVN3D [43] is a new method based on a deep 3D Hough voting network, which uses modules such as feature extraction, 3D keypoint detection, instance segmentation, and center voting, combined with multitask learning such as training 3D keypoints, semantic segmentation, and center voting to achieve pose estimation.

RGBD-based methods have significant advantages over the RGB-based methods. First, they are able to combine the texture information of RGB images and the geometric information of depth maps to provide a more accurate and robust pose estimation. Second, these methods perform better when dealing with illumination changes, complex backgrounds, and textureless objects. However, RGB-D-based methods also have their disadvantages. Due to the introduction of depth maps, these methods usually require more computing resources and higher computing power, resulting in more time-consuming training and testing processes. However, with the development of high-performance computing hardware, computing power is no longer a bottleneck that limits RGB-D methods. This makes RGB-D methods more feasible in practical applications.

2.3. Datasets

Datasets are very important in machine learning, especially deep learning. In addition to training and testing, datasets are also an important basis for comparisons between algorithms. With the development of 6D pose estimation research, a number of datasets have emerged (see Table 2).

The most commonly used datasets in current works are LineMOD [49] and YCB-V [50]. They all provide objects in real scenes, but the number and categories of objects are not too many.

The T-less datasets [51] are mainly aimed at some textureless and non-color-changing objects in the industry. The data in this dataset are captured in a real and controlled environment, and two 3D models are provided for each object.

The NOCS dataset [5] selects multiple real indoor scenes as the background and uses the plane segmentation algorithm to obtain the pixel-level segmentation of the desktop in the background image. Then, objects are placed in random positions and directions on this plane, and several virtual light sources are added to increase the authenticity.

ShapeNet6D [33], like NOCS, is also a dataset based on ShapeNet [52]. The diversity of samples is increased by adding textures and deforming 3D models in the ShapeNet dataset, and finally, a synthetic method is used to generate sample images.

Table 2. Statistics of different datasets.

Dataset	Modality	N_{cat}	N_{obj}	N_{img}
LineMOD [49]	RGBD	-	13	18,273
YCB-V [50]	RGBD	-	21	133,936
T-less [51]	RGBD	-	30	47,664
NOCS [5]	RGBD	6	1085	300,000
ShapeNet6D [33]	RGBD	51	12,490	800,000

3. Proposed Approach

6D pose estimation can be viewed as a rigid transformation $[\mathbf{R}, \mathbf{t}]$, where $\mathbf{R} \in SO^3$ donates 3D rotation, and $\mathbf{t} \in R^3$ represents translation. This transformation describes the conversion from the object coordinate system to the camera coordinate system. Our approach leverages the robust regression capabilities of neural networks to construct an end-to-end model that combines feature extraction with pose estimation, thus facilitating a direct regression of this transformation.

As shown in Figure 1, our proposed network architecture is composed of two primary components: edge detection and pose estimation. The edge detection part provides edge attention for fusion features by joint training an edge detection network.

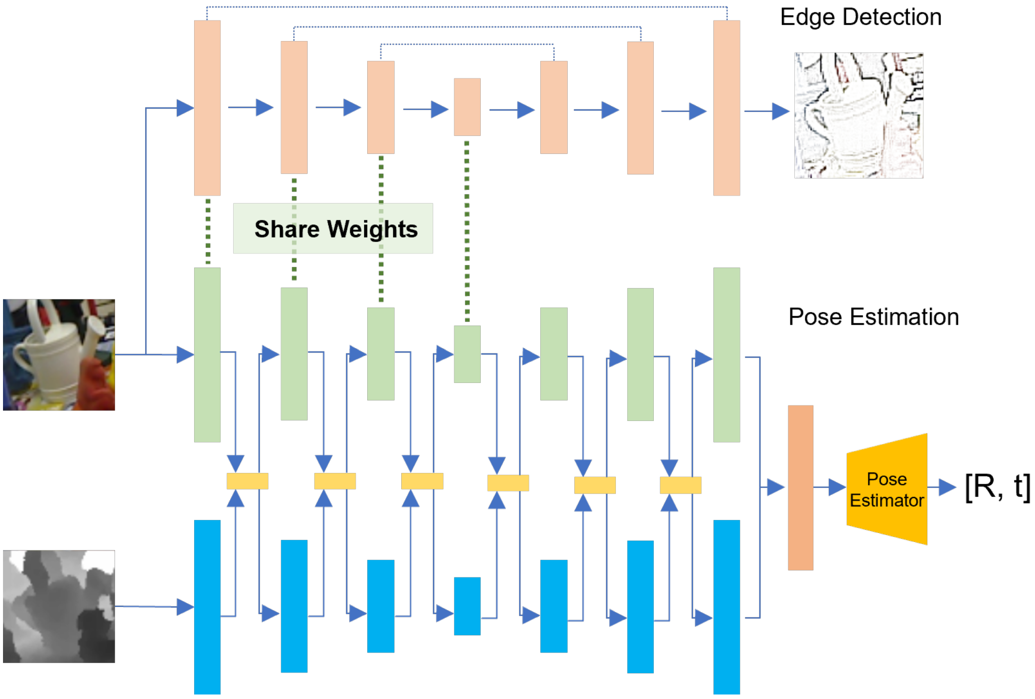

Figure 1. Pipeline of our method.

Initially, an RGB-D image is fed into a Full Flow Bidirectional Fusion Network (FFB) [29], which serves as an efficient backbone for feature extraction and fusion, capturing both semantic and geometric information. During the same time, the RGB image is processed by an edge detection network to extract edge features. The encoder within the edge detection network shares weights with the RGB encoder of the FFB network, ensuring edge attention is brought to feature representation.

During the feature extraction phase, the RGB and depth images undergo separate processing through their respective encoder layers to extract multi-scale features. A bidirectional flow mechanism is employed between these layers to facilitate deep feature fusion.

The fused features are then fed into the pose estimation module, which performs a direct regression of the rotation and translation parameters. The pose estimator integrates multi-scale features from the FFB network and computes the 6D pose parameters $[\mathbf{R}, \mathbf{t}]$ through a regression head. The entire network is trained end-to-end, enabling joint optimization of edge detection and pose estimation, which enhances the overall accuracy and robustness of the 6D pose estimation process.

3.1. Edge Attention

Edge features exhibit robustness, especially for objects with weak surface textures, and play a critical role in vision tasks by enabling the extraction of more detailed information from images without increasing the data volume. To leverage these benefits, we have introduced an edge attention mechanism into our pose estimation network.

The edge attention mechanism is implemented through an edge detection network branch that shares weights with the bidirectional feature fusion network. This integration ensures that the network gets edge attention, thereby enhancing the accuracy of pose estimation. The architecture of the edge detection network mirrors that of the RGB branch of the bidirectional feature fusion network, utilizing a ResNet34 [53] encoder pre-trained on ImageNet [54] and a PSPNet [55] decoder. This consistent architecture facilitates joint training and weight sharing, which is critical for the simultaneous optimization of both networks.

The edge detection network is trained using edge maps generated by the Canny operator as ground truth. By incorporating edge features, the network gains the ability to better delineate object boundaries, particularly in cluttered or occluded scenes where RGB features alone may be insufficient.

The joint training of the edge detection and pose estimation networks, guided by a combined loss function, ensures that the edge detection network contributes significantly to the overall pose estimation performance. By adding edge attention, our method can more precisely distinguish objects from their backgrounds, even in challenging scenarios. This edge-based approach leads to more accurate localization and translation estimates, ultimately improving the robustness of the pose estimation process.

3.2. Feature Extraction Network

The process of extracting and fusing features from RGB and depth images is critical for accurate pose estimation. Numerous methods for feature fusion have been proposed in previous work, such as DenseFusion [40], Robust6D [39], and FFB6D [29]. Among these, the network structure proposed by FFB6D offers a bidirectional feature fusion network, which performs feature fusion at each layer of encoding and decoding. This approach effectively fills information gaps and improves the quality of both appearance and geometric features.

The bidirectional fusion module, as proposed in FFB6D [29], enables the fusion of appearance and geometry information through two key processes: pixel-to-point fusion and point-to-pixel fusion.

In the pixel-to-point fusion module, RGB features are mapped to point cloud features. Specifically, for each point cloud feature, the corresponding K nearest neighbor pixels on the RGB feature map are identified, and max pooling is used to integrate these pixel features into the fused appearance features. These features are then concatenated with the original point cloud geometric features, and the final fusion features are obtained through a multi-layer perceptron (MLP).

In contrast, the point-to-pixel fusion module maps point cloud features to the RGB feature map. For each pixel feature, the corresponding K nearest neighbor points on the point cloud feature are identified, and these point features are integrated to obtain fused geometric features. These are then concatenated with the original RGB features, and the final fused features are produced through an MLP.

In our feature extraction and fusion process, RGB and depth image features are initially extracted through their respective encoder layers. During this process, the edge attention features are also incorporated into the fusion, resulting in more expressive features that significantly contribute to the accuracy of pose estimation.

3.3. Dataset

We train and test our method on the LineMOD [49] and YCB-V [50] dataset. The LineMOD dataset includes 13 different objects with more than 18,000 images, each with varying shapes, sizes, and textures. It is widely recognized for its challenging scenarios, including cluttered backgrounds, occlusions, and varying lighting conditions. The objects in this dataset are commonly used in industrial and robotic applications, which makes it a suitable benchmark for evaluating pose estimation methods in practical settings.

The YCB-V dataset, which is a subset of the YCB (Yale–CMU–Berkeley) Object and Model Set, is known for its high variability in object types, ranging from simple shapes to more complex geometries. It includes 133,936 images of 21 objects that are commonly encountered in everyday tasks, such as tools, food items, and household objects. The dataset also features scenes with varying levels of occlusion and lighting, providing a robust testbed for assessing the model's performance across a range of conditions.

3.4. Loss Function

Our network is an end-to-end network and consists of two branches, so a multitask loss function is required.

For the pose estimation branch, the network outputs a rotation matrix **R** and a translation vector **t**.

For the network output to be more accurate, the point transformed by the prediction result should be closer to the point transformed by the ground truth.

The network learning process can be understood as minimizing the distance between the pair of above points.

Therefore, the loss function for pose estimation is defined as follows:

$$L^i_{pose} = \frac{1}{N}\sum_{n\in N} \| \, (\bar{R}x_n + \bar{t}) - (\hat{R}_i x_n + \hat{t}_i) \, \|_2 \tag{1}$$

where x is the nth point in point set N on the CAD model corresponding to the image, $[\bar{R}, \bar{t}]$ denotes ground truth pose, and $[\hat{R}_i, \hat{t}_i]$ refers to the pose predicted by the ith point in the point set I sampled in the image.

For a symmetric object, due to the uncertainty caused by its symmetry, we define its loss function as follows, which is used in [7,40]:

$$L^i_{pose} = \frac{1}{N}\sum_{n\in N} \min_{0<n<N} \| \, (\bar{R}x_n + \bar{t}) - (\hat{R}_i x_n + \hat{t}_i) \, \|_2 \tag{2}$$

This loss measures the offset between each point on the estimated pose and the closest point on the ground truth model. In this way, it will not penalize rotations that are equivalent with respect to the 3D shape symmetry of the object [7].

After getting the above loss function, we minimize the sum of the confidence-weight per dense-pixels losses with a confidence regularization:

$$L_{pose} = \frac{1}{N}\sum_i (L^i_{pose} c_i - \omega \log(c_i)) \tag{3}$$

where c_i denotes the confidence of pose predicted by the ith pixel, and ω is a hyperparameter.

For the loss function of another network branch, the edge detection branch, we use binary cross entropy with logits:

$$L_{edge} = -\beta \sum_{E_g t(i,j=1)} \log E_x(i,j) - (1-\beta) \sum_{E_g t(i,j=0)} \log(1 - E_x(i,j)) \tag{4}$$

where (i,j) denotes the location of the pixel on images, $E_g t(i, j = 1)$ denotes pixel (i,j) is on the edge, and β denotes the percentage of non-edge pixel in the whole image.

Finally, we combine the above loss functions to get the final loss function:

$$Loss = L_{pose} + \lambda L_{edge} \tag{5}$$

where λ is a hyperparameter for balance.

3.5. Evaluation Matrics

The average distance metrics ADD and ADDS are widely used for the performance evaluation of the 6D pose estimation. The ADD metric computes the average Euclidean distance between the corresponding 3D points on the object model when transformed by the ground truth pose $[\bar{R}, \bar{t}]$ and the estimated pose $[\hat{R}, \hat{t}]$. Mathematically, it is defined as follows:

$$ADD = \frac{1}{N} \sum_{p \in O} \parallel (\hat{R}p + \hat{T}) - (\bar{R}p + \bar{T}) \parallel \tag{6}$$

where $p \in O$ refers to the points on the model, and N refers to the total number of points.

For symmetric objects, the ADD metric can be problematic because the transformation might lead to multiple correct correspondences, resulting in an artificially inflated error. To address this, the ADDS metric is used, which computes the average distance between a point on the object model transformed by the estimated pose and the closest point on the model transformed by the ground truth pose:

$$ADDS = \frac{1}{N} \sum_{p_1 \in O} \min_{p_2 \in O} \parallel (\hat{R}p_1 + \hat{T}) - (\bar{R}p_2 + \bar{T}) \parallel \tag{7}$$

We report the area under the accuracy threshold curve obtained by varying the distance threshold (ADDS and ADD(S) AUC) in the YCB-V dataset. In the LineMOD datasets, we report the accuracy of the distance less than 10% of the diameter of the objects (ADD-0.1d).

3.6. Results

We evaluated our proposed method on the YCB-V and LineMOD datasets, benchmarking its performance against several state-of-the-art methods. The quantitative results are summarized in Tables 3 and 4.

Performance on the YCB-V Dataset. As shown in Table 3, our method outperforms existing approaches, achieving high accuracy on most objects. In particular, for objects with distinct edge information, our method leverages the edge attention mechanism to improve the accuracy of pose estimation. For example, our method achieves 97.9% accuracy on the 'sugar box' object and 96.9% on 'scissors', outperforming the best-performing baseline models such as PVN3D and FFB6D. However, for objects with symmetric or less pronounced edge features, the improvement is less significant. For example, the 'bowl' object achieves an accuracy of 92.0%, which is comparable to the results of other methods.

Overall, the results on the YCB-V dataset demonstrate that our method provides a substantial improvement in pose estimation accuracy, particularly for objects where edge features play a critical role in defining the shape and orientation.

Performance on the LineMOD Dataset. Table 4 presents the results of our method on the LineMOD dataset, where it outperforms the competing methods. Our approach achieves state-of-the-art results for several objects, highlighting the robustness of our method. The addition of the edge attention mechanism significantly improves the performance, especially on challenging objects with complex geometries like 'ape' and 'duck'.

These results validate the effectiveness of incorporating edge information into the pose estimation process, enabling our method to achieve state-of-the-art performance on both the YCB-V and LineMOD datasets. Performance gains are particularly evident for objects where edge features are crucial, confirming the utility of our approach in enhancing the precision of 6D pose estimation tasks.

Table 3. Quantitative evaluation results on YCB-V dataset.

Objects	PoseCNN [7]	PVN3D [43]	Uni6d [34]	DeepIM [56]	FS6D [33]	FFB6D [29]	Ours
master chef can	50.9	80.5	70.2	71.2	36.8	80.6	80.7
cracker box	51.7	94.8	85.2	83.6	24.5	94.6	95.3
sugar box	68.6	96.3	94.5	94.1	43.9	96.6	97.9
tomato soup can	66.0	88.5	85.4	86.1	54.2	89.6	88.7
mustard bottle	79.9	96.2	91.7	91.5	71.1	97.0	96.5
tuna fish can	70.4	89.3	79.0	87.7	53.9	88.9	88.9
pudding box	62.9	95.7	89.8	82.7	79.6	94.6	94.5
gelatin box	75.2	96.1	96.2	91.9	32.1	96.9	92.7
potted meat can	59.6	88.6	89.6	76.2	54.9	88.1	87.6
banana	72.3	93.7	93.0	81.2	69.1	94.9	96.1
pitcher base	52.5	96.5	94.2	90.1	40.4	96.9	96.7
bleach cleanser	50.5	93.2	91.1	81.2	44.1	94.8	95.4
bowl	69.6	90.2	95.5	81.4	0.9	96.3	92.0
mug	57.7	95.4	93.0	81.4	39.2	94.2	94.4
power drill	55.1	95.1	91.1	85.5	19.8	95.9	95.5
wood block	31.8	90.4	94.3	81.9	27.9	92.6	92.8
scissors	35.8	92.7	79.6	60.9	27.7	95.7	96.9
large marker	58.0	91.8	92.8	75.6	74.2	89.1	89.4
large clamp	25.0	93.6	95.9	74.3	34.7	96.8	95.9
extra large clamp	15.8	88.4	95.8	73.3	10.1	96.0	97.1
foam brick	40.4	96.8	96.1	81.9	45.8	97.3	97.7

Table 4. Quantitative evaluation results on LineMOD.

Objects	PVNet [48]	PoseCNN [7]	DPOD [10]	Pix2Pose [6]	HybridPose [57]	Robust6D [39]	Ours
ape	43.6	21.6	53.3	58.1	63.1	85.0	99.7
benchvise	99.9	81.8	95.3	91.0	99.9	95.5	100.0
cam	86.9	36.6	90.4	60.9	90.4	91.2	100.0
can	95.5	68.8	94.1	84.4	98.5	95.1	99.1
cat	79.3	41.8	60.4	65.0	89.4	93.6	100.0
driller	96.4	63.5	97.7	76.3	98.5	82.6	100.0
duck	52.6	27.2	66.0	43.8	65.0	88.1	100.0
eggbox	99.2	69.6	99.7	96.8	100.0	99.9	98.9
glue	95.7	80.0	93.8	79.4	98.8	99.6	99.4
holepuncher	81.9	42.6	65.8	74.8	89.7	92.6	100.0
iron	98.9	74.9	99.8	83.4	100.0	95.9	100.0
lamp	99.3	71.1	88.1	82.0	99.5	94.4	99.8
phone	92.4	47.7	74.2	45.0	94.9	93.5	99.9

4. Conclusions

In this work, we introduce an edge attention mechanism into the bidirectional feature fusion network to enhance pose estimation performance. By sharing the weights of the edge detection encoder with the RGB branch encoder in the feature fusion network, the model pays more attention to edge information. Our approach achieves state-of-the-art results on the YCB-V and LineMOD datasets.

However, there is room for future improvements. Given the complexity of our model, computational efficiency could be further improved using techniques such as quantization, pruning, and distillation. Additionally, real-world data often present more variability than the datasets used in this work. Improving generalization to unseen objects in real-world scenarios through domain adaptation and transfer learning is another important direction for future research.

Author Contributions: Methodology, Q.F.; Validation, Q.F.; Formal analysis, Q.F.; Investigation, J.N.; Resources, Q.F. and J.N.; Writing—original draft, Q.F.; Writing—review & editing, J.N. and Y.L.; Supervision, Y.L.; Funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Zhuhai Industry-University-Research Project (No. 2220004002411), National Key Research and Development Plan under Grant 2021YFE0205700, Science and Technology Development Fund of Macau project 0070/2020/AMJ, 00123/2022/A3, 0096/2023/RIA2.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; Wang, X. Deep dual consecutive network for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 525–534.
2. Peng, Q.; Zheng, C.; Chen, C. A Dual-Augmentor Framework for Domain Generalization in 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 16–21 June 2024; pp. 2240–2249.
3. Chen, X.; Wang, G.; Guo, H.; Zhang, C. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing* **2020**, *395*, 138–149. [CrossRef]
4. Wang, Y.; Zhao, P.; Zhang, Z. A deep learning approach using attention mechanism and transfer learning for electromyographic hand gesture estimation. *Expert Syst. Appl.* **2023**, *234*, 121055. [CrossRef]
5. Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; Guibas, L.J. Normalized object coordinate space for category-level 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2642–2651.
6. Park, K.; Patten, T.; Vincze, M. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7668–7677.
7. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv* **2017**, arXiv:1711.00199.
8. Wang, G.; Manhardt, F.; Tombari, F.; Ji, X. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16611–16621.
9. Kaskman, R.; Shugurov, I.; Zakharov, S.; Ilic, S. 6 dof pose estimation of textureless objects from multiple rgb frames. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 612–630.
10. Zakharov, S.; Shugurov, I.; Ilic, S. Dpod: 6d pose object detector and refiner. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1941–1950.
11. Su, Y.; Saleh, M.; Fetzer, T.; Rambach, J.; Navab, N.; Busam, B.; Stricker, D.; Tombari, F. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6738–6748.
12. Shugurov, I.; Zakharov, S.; Ilic, S. Dpodv2: Dense correspondence-based 6 dof pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7417–7435. [CrossRef] [PubMed]
13. Sun, J.; Wang, Z.; Zhang, S.; He, X.; Zhao, H.; Zhang, G.; Zhou, X. Onepose: One-shot object pose estimation without cad models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6825–6834.
14. Li, F.; Yu, H.; Shugurov, I.; Busam, B.; Yang, S.; Ilic, S. NeRF-Pose: A First-Reconstruct-Then-Regress Approach for Weakly-supervised 6D Object Pose Estimation. *arXiv* **2022**, arXiv:2203.04802.

15. Lipson, L.; Teed, Z.; Goyal, A.; Deng, J. Coupled iterative refinement for 6d multi-object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6728–6737.
16. Chen, H.; Wang, P.; Wang, F.; Tian, W.; Xiong, L.; Li, H. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 2781–2790.
17. Fan, Z.; Pan, P.; Wang, P.; Jiang, Y.; Xu, D.; Wang, Z. POPE: 6-DoF Promptable Pose Estimation of Any Object in Any Scene with One Reference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 16–21 June 2024; pp. 7771–7781.
18. 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 Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 298–315.
19. Castro, P.; Kim, T.K. Crt-6d: Fast 6d object pose estimation with cascaded refinement transformers. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 5746–5755.
20. Do, T.T.; Cai, M.; Pham, T.; Reid, I. Deep-6dpose: Recovering 6d object pose from a single rgb image. *arXiv* **2018**, arXiv:1802.10367.
21. Yang, Z.; Yu, X.; Yang, Y. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3907–3916.
22. Shao, J.; Jiang, Y.; Wang, G.; Li, Z.; Ji, X. PFRL: Pose-Free Reinforcement Learning for 6D Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 11454–11463.
23. Shugurov, I.; Li, F.; Busam, B.; Ilic, S. Osop: A multi-stage one shot object pose estimation framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6835–6844.
24. 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. In Proceedings of the Asian Conference on Computer Vision, Daejeon, Republic of Korea, 5–9 November 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 548–562.
25. Sundermeyer, M.; Marton, Z.C.; Durner, M.; Brucker, M.; Triebel, R. Implicit 3d orientation learning for 6d object detection from rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 699–715.
26. Aing, L.; Lie, W.N.; Chiang, J.C.; Lin, G.S. Instancepose: Fast 6dof pose estimation for multiple objects from a single rgb image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2621–2630.
27. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-driven 6d object pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3385–3394.
28. Chen, W.; Jia, X.; Chang, H.J.; Duan, J.; Leonardis, A. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 4233–4242.
29. He, Y.; Huang, H.; Fan, H.; Chen, Q.; Sun, J. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3003–3013.
30. He, X.; Sun, J.; Wang, Y.; Huang, D.; Bao, H.; Zhou, X. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 35103–35115.
31. Zhang, R.; Di, Y.; Lou, Z.; Manhardt, F.; Tombari, F.; Ji, X. Rbp-pose: Residual bounding box projection for category-level pose estimation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 655–672.
32. Peng, W.; Yan, J.; Wen, H.; Sun, Y. Self-supervised category-level 6D object pose estimation with deep implicit shape representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 22 February–1 March 2022; Volume 36, pp. 2082–2090.
33. He, Y.; Wang, Y.; Fan, H.; Sun, J.; Chen, Q. FS6D: Few-Shot 6D Pose Estimation of Novel Objects. *arXiv* **2022**, arXiv:2203.14628.
34. Jiang, X.; Li, D.; Chen, H.; Zheng, Y.; Zhao, R.; Wu, L. Uni6d: A unified cnn framework without projection breakdown for 6d pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; Volume 1, p. 5.
35. Mo, N.; Gan, W.; Yokoya, N.; Chen, S. Es6d: A computation efficient and symmetry-aware 6d pose regression framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6718–6727.
36. Xu, Y.; Lin, K.Y.; Zhang, G.; Wang, X.; Li, H. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 14880–14890.
37. Lin, H.; Liu, Z.; Cheang, C.; Fu, Y.; Guo, G.; Xue, X. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6707–6717.
38. Cai, D.; Heikkil, J.; Rahtu, E. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 6803–6813.

39. Tian, M.; Pan, L.; Ang, M.H.; Lee, G.H. Robust 6d object pose estimation by learning rgb-d features. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–4 June 2020; pp. 6218–6224.
40. 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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3343–3352.
41. Kehl, W.; Milletari, F.; Tombari, F.; Ilic, S.; Navab, N. Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 205–220.
42. Wang, C.; Martín-Martín, R.; Xu, D.; Lv, J.; Lu, C.; Fei-Fei, L.; Savarese, S.; Zhu, Y. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–4 June 2020; pp. 10059–10066.
43. 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 11632–11641.
44. Capellen, C.; Schwarz, M.; Behnke, S. ConvPoseCNN: Dense convolutional 6D object pose estimation. *arXiv* **2019**, arXiv:1912.07333.
45. Di, Y.; Manhardt, F.; Wang, G.; Ji, X.; Navab, N.; Tombari, F. SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12396–12405.
46. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8922–8931.
47. Park, J.J.; Florence, P.; Straub, J.; Newcombe, R.; Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 165–174.
48. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. Pvnet: Pixel-wise voting network for 6dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4561–4570.
49. Hinterstoisser, S.; Holzer, S.; Cagniart, C.; Ilic, S.; Konolige, K.; Navab, N.; Lepetit, V. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 858–865.
50. Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; Dollar, A.M. The ycb object and model set: Towards common benchmarks for manipulation research. In Proceedings of the 2015 International Conference on Advanced Robotics (ICAR), Istanbul, Turkey, 27–29 July 2015; pp. 510–517.
51. 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. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 880–888.
52. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. Shapenet: An information-rich 3d model repository. *arXiv* **2015**, arXiv:1512.03012.
53. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
54. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
55. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
56. Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; Fox, D. Deepim: Deep iterative matching for 6d pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 683–698.
57. Song, C.; Song, J.; Huang, Q. Hybridpose: 6d object pose estimation under hybrid representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 431–440.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.