IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 December 2023, accepted 22 February 2024, date of publication 1 March 2024, date of current version 7 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3372658

RESEARCH ARTICLE

Category-Level Object Pose Estimation in Heavily Cluttered Scenes by Generalized Two-Stage Shape Reconstructor

HIROKI TATEMICHI[1], YASUTOMO KAWANISHI[2], (Member, IEEE), DAISUKE DEGUCHI[1], (Member, IEEE), ICHIRO IDE[1], (Senior Member, IEEE), AND HIROSHI MURASE[1], (Life Fellow, IEEE)

[1]Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan
[2]Guardian Robot Project, RIKEN, Souraku, Kyoto 619-0288, Japan

Corresponding author: Yasutomo Kawanishi (yasutomo.kawanishi@riken.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (KAKENHI) under Grant Numbers (JP17H00745, JP21H03519).

ABSTRACT In this paper, we propose a method for robust estimation of the pose of an unknown object instance in an object category from a depth image, even if it is occluded. In cluttered scenes, objects are often mutually occluded, and at the same time, objects in a category often have various shapes. For estimating the object pose in such situations, we have previously proposed a Two-stage Shape Reconstructor to extract features by de-occluding the occluded region of a target object and absorbing shape variations in a category. However, the model could not be used except in a situation where the unoccluded mask of the target object is known, and only if the contour of the occluding object is expected to have a linear shape. To cope with this problem, we generalize the previous model to directly extract the feature of a de-occluded object from a depth image containing the object occluded by another object. We also propose a data augmentation method for effectively training the model. Through evaluations on large-scale virtual-world and real-world datasets, we confirm that the proposed method achieves promising results on pose estimation of an unknown occluded object from an observed depth image.

INDEX TERMS Object pose estimation, AutoEncoder, occlusion.

I. INTRODUCTION

Human support robots in daily life have been actively developed in recent years. Such a robot should have the basic functionality of grasping and carrying a target object, such as a mug, according to a request from a user. Object grasping is actively developed in robotics to realize the functionality. It is common for robots to be equipped with an image sensor capturing an RGB/depth image to observe the surroundings. Depth images are especially attracting attention due to their robustness against variations in color and lighting conditions. Object grasping requires not only the detection of the target object but also the accurate estimation of its pose based on such sensor data. While object detection has reached a practical level by recent methods such as Deformable-DETR [1] and YOLOv7 [2], object pose estimation remains a challenging task. It is difficult to estimate the pose of an object occluded by another object, especially for an object with no 3D model available. In this paper, we focus on the pose estimation of such an object.

For estimating the object pose from an image, it is a common practice first to detect an object region in the entire image. If another object occludes the target object, the observable area of the object changes depending on the shape and position of the occluding object (Fig. 1). A previously proposed occlusion-robust object pose estimation method [3] tackles this problem by extracting features by an Augmented AutoEncoder. This AutoEncoder is trained to decode the entire region from an RGB image capturing a target object with a defect, various background clutters, and

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Anisetti.

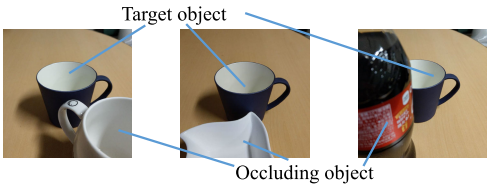


FIGURE 1. Various occlusions change the observable area of a target object.

FIGURE 2. Shape variations within a category.

dynamic changes of the environment, based on a Denoising AutoEncoder. However, this method needs to train various poses of a target object and cannot be applied to an unknown object. This kind of pose estimation task is called *instance-level* object pose estimation [4]. In the task, a specific object with an exact 3D model available is expected to be observed, and its pose can be estimated accurately by comparing the observed shape to the 3D model.

On the other hand, this paper handles variously shaped objects in a category, called *category-level* object pose estimation. In such a task, objects are expected to have various colors and shapes within a category (Fig. 2), and unknown objects with no 3D model available, should be handled. A category-level pose estimation method [5] tackles this problem by extracting features by a Pose-CyclicR-Net based on Deep Convolutional Neural Networks (DCNNs). However, it does not consider a situation when the observed object is severely occluded.

We have previously proposed a method based on a two-stage Encoder-Decoder model that realizes accurate pose estimation of an occluded object in category-level [4]. This model can absorb the shape variations within a category and can accurately estimate the pose of an object even if it is half-occluded. However, as it expects that each input image contains only a segmented object that is occluded simply, i.e., the target is occluded by a simple shaped object, there are two constraints on the input images as follows:

- A segmentation mask of the visible part of the target object should be given.
- The contour of the occluded region is expected to be linear.

If the target object is sufficiently far from the occluding objects, the segmentation mask of the visible part of the target object can be easily estimated based on the difference of depth values. However, in practice, the occluded objects are generally located near the occluding objects in cluttered scenes, which also have various shapes (Fig. 3). In this paper,

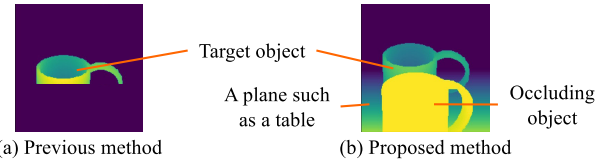


FIGURE 3. Expected input images.

we further improve this previous method [4] for such general scenes.

The main contributions of this work are as follows:

- We extend the first stage of the Two-stage Shape Reconstructor to a generalized one; the Generalized Two-stage Shape Reconstructor, which directly extracts features of an unknown occluded object from a depth image containing another unknown occluding object.
- We propose a data augmentation method suitable for training the Generalized De-occluding AutoEncoder (GDAE).
- We evaluate the proposed method on a large-scale virtual dataset and a real dataset which consist of images with various virtual and real occlusions, respectively.

II. RELATED WORK

A. SENSORY DATA FOR POSE ESTIMATION

Object pose estimation methods for a robot can be categorized into several approaches in accordance with data representation: image-based [3], [4], [5], [6], [7], [8], [9], [10], point cloud-based [11], and voxel-based [12]. The image-based one is the most practical, given that it is commonly mounted on a robot for daily life support. Among them, methods based on an RGB image estimate an object pose from color values. On the other hand, those based on a depth image estimate from the 3D shape represented by depth values to the target object. In comparison with the former methods, the latter methods are not only robust to texture variations of objects, background disturbance, and lighting variations, but also can easily extract the region of the target object if it is far from the surrounding objects. Therefore, in this paper, we make use of the depth image representation.

B. POSE ESTIMATION BASED ON HAND-CRAFTED FEATURES

For image-based object pose estimation, it is common to use features extracted from an observed image [13]. One of the earliest image-based methods is the template matching approach [14]. It searches for the best-matched template based on color values from many templates of a target. The Parametric Eigenspace method [6] reduces the number of templates and embeds the continuous pose change of an object onto a manifold in a low-dimensional subspace represented by a small number of templates. Afterward, various methods based on hand-crafted features were proposed for more effective image recognition [15].

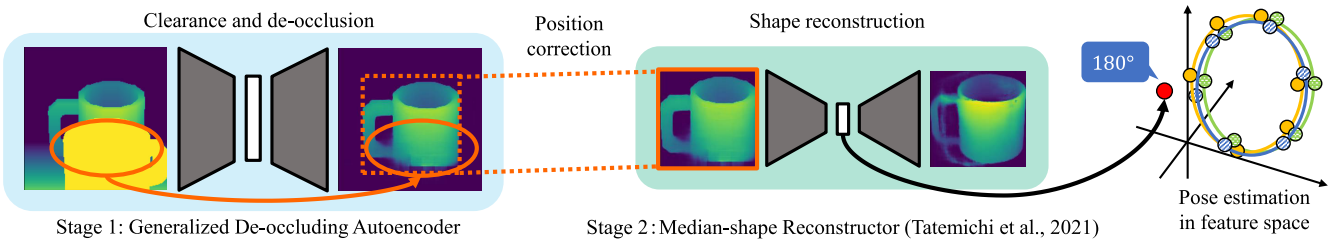


FIGURE 4. Proposed Generalized Two-stage Shape Reconstructor. The input is a depth image containing the object occluded by an arbitrarily shaped object. The Generalized De-occluding AutoEncoder (GDAE) clears the surrounding objects and de-occludes the occluded region of the target object. Then, the complete region of the object is cropped from the de-occluded image. The Median-shape Reconstructor extracts feature vectors from the de-occluded image for pose estimation.

These methods need hand-crafted features representing the property of the target object or the surrounding environment. Additionally, they cannot adapt to shape variations in a category.

C. POSE ESTIMATION BASED ON DEEP LEARNING

In recent pose estimation methods, DCNNs are usually used for extracting more efficient features from an image. Pose-CNN [16], SSD-6D [17], Dense Fusion [18], DualPoseNet [19], and GPV-Pose [20] are some of the accurate instance-level object pose estimation methods. However, these methods are not robust to environmental change including severe occlusions. Since the extracted features depend only on the observable region of the target object, the features are largely affected by occlusions. If the entire region of the occluded object could be predicted, they would work well, regardless to the occlusion. It leads to occlusion-robust pose estimation. To extract such features, Sundermeyer et al. [3] proposed the Augmented AutoEncoder, which generalized the Denoising AutoEncoder [21]. The method extracts a feature from the entire region of the object instance, which is de-occluded from an observed image with some defects in the object, various background clutters, and various lighting conditions.

The above methods based on DCNNs all take instance-level pose estimation approaches. In contrast, a category-level pose estimation method from a depth image was proposed by Ninomiya et al. [5]. They proposed Pose-CyclicR-Net, which is a modified DCNN for handling the circularity of object poses by introducing trigonometric functions as outputs. Sun et al. [22] proposed OnePose, which can estimate both instance- and category-level object pose. The model trained a feature matching network between 2D and 3D keypoints.

However, these methods did not explicitly focus on the occlusion of the target object. To handle the occlusions, we have proposed the Two-stage Encoder-Decoder model [4] to realize an occlusion-robust pose estimation at the category level. This method does not only tackle the object occlusion at the category level by extending the Augmented AutoEncoder to a two-stage model, but also tackles the difficulty of shape variations in a category by a Median-shape Reconstructor to reconstruct the representative shape in the category. However, it assumes that the input is a depth image with a segmentation mask of the target object, whose contour of the occluded region is linearly shaped.

III. GENERALIZED TWO-STAGE SHAPE RECONSTRUCTOR

The goal of this paper is to realize a category-level occlusion-robust object pose estimation method in general cluttered scenes considering the various shapes and positions of unknown occluding objects. Here, we assume that the target object category is known, thus we build the pose estimation model for each category. We propose a Generalized Two-stage Shape Reconstructor to extract features from a depth image containing an occluded target object with another object.

Fig. 4 illustrates the proposed Generalized Two-stage Shape Reconstructor. The previously proposed model [4] sets a strong assumption that the input is a depth image with a segmentation mask of the target object, whose contour of the occluded region is linearly shaped. On the other hand, the proposed Generalized Two-stage Shape Reconstructor features the Generalized De-occluding AutoEncoder (GDAE), which is extended to expect an input depth image containing a target object occluded with another object whose contour is not necessarily linearly shaped. This paper proposes a novel framework to construct the GDAE that can directly remove an arbitrarily shaped occluding object by reconstructing the target shape from the occluded image in an end-to-end manner. To achieve this goal, it is necessary to prepare a large number of training images containing variously shaped occluding objects. However, in a real-world scenario, the cost of preparing these training images cannot be ignored. Therefore, the proposed method extends the data augmentation pipeline by introducing an occluded image synthesis process that can handle the target object and the occluding object separately.

In the rest of this section, we describe the pose representation learning method including the training of the proposed GDAE in Section III-A, the proposed data augmentation method in Section III-B, and the pose estimation method using the Generalized Two-stage Shape Reconstructor in Section III-C.

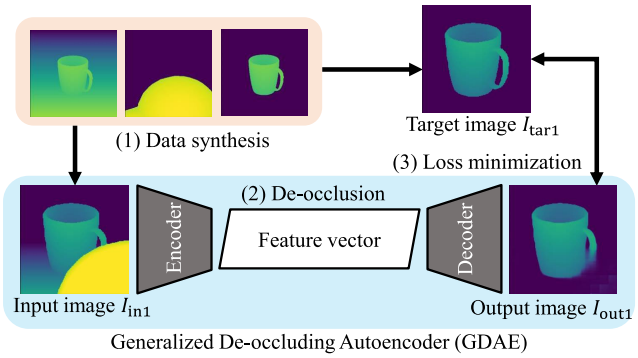


FIGURE 5. Training the Generalized De-occluding AutoEncoder (GDAE).

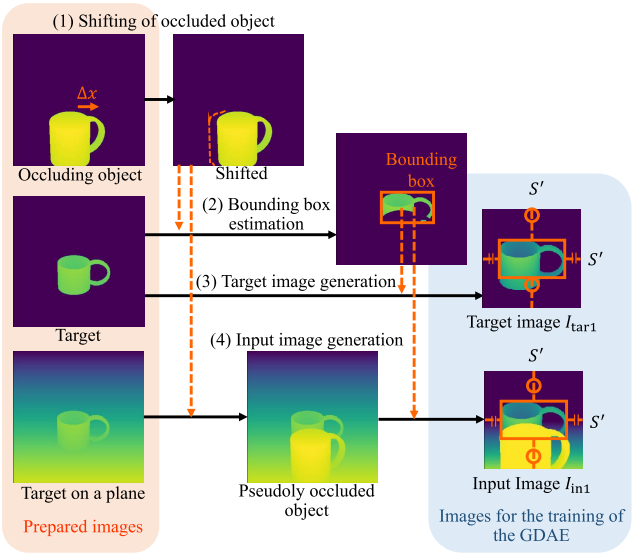


FIGURE 6. Data augmentation of occluded images by synthesizing a target and an occluding object.

A. POSE REPRESENTATION LEARNING

The two models in the proposed Generalized Two-stage Shape Reconstructor, which are Generalized De-occluding AutoEncoder (GDAE) and Median Shape Reconstructor, are trained sequentially.

1) GENERALIZED DE-OCCLUSION LEARNING

GDAE is trained to directly extract the feature of an object from a depth image containing another unknown occluding object. Fig. 5 illustrates the training of GDAE. First, the input image I_{in1} and the target image I_{tar1} are generated (Fig. 5 (1)). In this procedure, the target object and the occluding object are randomly selected, and the former is occluded by the latter according to the process introduced in Section III-B. The background of the target object is expanded so that most of the object can be de-occluded in the output image. GDAE decodes the de-occluded image I_{out1} from the input image I_{in1} (Fig. 5 (2)). In the process, surrounding objects are removed from the de-occluded image I_{out1}. For minimizing difference between I_{out1} and I_{tar1} (Fig. 5 (3)), the weights of the GDAE are optimized.

2) DE-OCCLUSION AND POSITION CORRECTION

After training the GDAE, it will decode a de-occluded image similar to the target. A plausible bounding box surrounding the entire region of the target is estimated from the de-occluded image. Then, the target position is aligned so that the bounding box is located at the image center. This procedure tackles the problem of the offset between the true object center and the image center, as mentioned in [4]. The input image I_{in2} generated by this procedure is used for the second stage, that is, the Median-shape Reconstructor, to extract a feature vector.

3) MEDIAN-SHAPE REPRESENTATION LEARNING

The Median-shape Reconstructor is an Encoder-Decoder model proposed in our previous work [4]. We use the model to extract the feature vector $\mathbf{v}$ from the de-occluded target object. This model is trained to encode an input image I_{in2} containing the target object to the feature vector $\mathbf{v}$ and decode an image containing the median-shaped object in the category of the input object.

Before the training, the median-shaped object is selected as the minimum one of the summations of the distance to all of the other object instances in the category. For selecting the median-shaped object, first, we define the distance $d_{i,j}$ between two object instances i and j. Here, we assume that we have images of each object in various poses. The distance $d_{i,j}$ is defined as the sum of distances between images in each pose k of the object instances as $d_{i,j} = \sum_k \|\mathbf{w}_{i,k} - \mathbf{w}_{j,k}\|$, where $\mathbf{w}_{i,k}$ and $\mathbf{w}_{j,k}$ are the image features of object instances i and j in pose k, respectively. Here, the image feature is extracted from each image. The feature is an intermediate activation of a DCNN model based on Pose-CyclicR-Net [5], which is trained with pairs of an image and the pose represented by trigonometric functions $(\cos\theta, \sin\theta)$. By following the definition of *Median* in Statistics, we select an object instance i where the value $D_i = \sum_j d_{i,j}$ is the minimum, as the median-shaped object in the training set. Since the selected median-shaped object robustly simulates the global median shape of the category, it is expected to be also a good reference for any test set.

B. DATA AUGMENTATION FOR GENERALIZED DE-OCCLUSION LEARNING

In our previous work [4], the dataset was generated by a pseudo-occluding process where the target objects are occluded horizontally or vertically. Although this procedure allows us to prepare the dataset easily, the input image is limited to images where the contour of the occluded region is linearly shaped. In this paper, we assume that an object is occluded by an arbitrarily shaped object. Due to the combinatorial explosion, it is difficult to create such an image

in a virtual world, and even more difficult in the real world. To overcome this difficulty, we propose a data augmentation method by collecting images containing a target object and an occluding object separately and superimposing the latter on the former for training the model. Fig. 6 shows the overview of the proposed data augmentation method. The main features of the method are as follows:

- Pseudo-occluding process with the transition of the occluding object for generating various realistic occlusions.
- Training data synthesis by considering the offset of the detected bounding box caused by the occlusion.

1) DATA PREPARATION

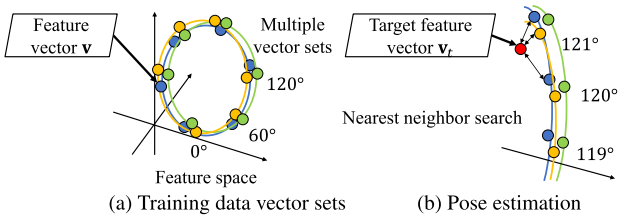
For training the model properly, the training data should contain diverse variations, especially each occluding object and its position. Considering the combinatorial explosion of preparing such a dataset, we segment each image into three parts; a target object, an occluding object, and the background, and prepare them independently. Then, we can generate various images by combining them randomly. We describe how to obtain these images in a virtual world and the real world, in Section IV.

2) DATA AUGMENTATION

The goal of this procedure is to generate an input image I_{in1} and a target image I_{tar1} for the DGAE from the above-mentioned images. First, for simulating various locations of an occluding object, an occluding object image is randomly shifted along the x-axis. In the procedure, it is shifted by Δx in the x-axis direction (Fig. 6(1)). Second, the observable area of the target and its bounding box is estimated using a target image and a shifted occluding object image (Fig. 6(2)). Third, the target image I_{tar1} is generated by cropping it with the bounding box located at the image center since the input image I_{in1} is expected to be input with the bounding box located at the image center (Fig. 6(3)). Then, an image where the target object is occluded by another object is generated by superimposing an occluding object on a target object on a plane (Fig. 6(4)). Finally, the input image I_{in1} is cropped by the same region referring to the target image I_{tar1}. The image is cropped with the bounding box located at the image center.

C. POSE ESTIMATION

We follow the object pose estimation method described in our previous work [4]. The pose estimator is based on the Nearest Neighbor classifier. In the training phase, a set of feature vectors $\mathcal{V} = \{\mathbf{v}\}$ is extracted from the training depth images of an object in various poses. The vectors in $\mathcal{V}$ continuously vary in the feature space (Fig. 7(a)) according to the pose of the object. In the case of 1D rotation of an object, based on the concept of the Parametric Eigenspace method [6], the feature vectors line up on a 1D manifold in the feature space. From multiple object instances, a set of feature vector sets $\mu = \{\mathcal{V}_1, \mathcal{V}_2, \cdots\}$ is extracted. In Fig. 7(a), the color of a solid



(a) Training data vector sets (b) Pose estimation

FIGURE 7. Nearest neighbor search-based pose estimation.

circle indicates the vectors extracted from the same object instance. Since the Median-shape Reconstructor shrinks the shape variation in a category, the extracted feature vectors are expected to be the same if the poses are the same, even for different objects.

For pose estimation, a feature vector $\mathbf{v_t}$ is first extracted from a target depth image by the Generalized Two-stage Shape Reconstructor. The pose is estimated by Nearest Neighbor regression with feature vectors in the vector sets $\mathcal{V}_1, \mathcal{V}_2, \cdots$ with pose labels. The pose estimator searches for the nearest neighbor vector $\widehat{\mathbf{v}}$ from the set of vectors μ as:

$$\widehat{\mathbf{v}} = \underset{\mathbf{v}\in\mathcal{V}_x, \mathcal{V}_x\in\mu}{\operatorname{argmin}} \|\mathbf{v} - \mathbf{v_t}\| \tag{1}$$

Finally, the pose estimator outputs the pose $\widehat{\mathbf{v}}$, which is the nearest neighbor of the input feature vector $\mathbf{v_t}$.

IV. EXPERIMENTAL EVALUATIONS

To confirm the effectiveness of the proposed method in challenging scenes where the target object is heavily occluded by an unknown object, we evaluated the method using two original datasets. We selected ''mug'' as the target category for the reason that it is one of the asymmetric objects that we can define the pose uniquely, and often appears in home scenes. We selected four categories; ''bottle'', ''cap'', ''mug'', and ''vase'' as occluding object categories that also often appear in such scenes. Additionally, we selected 30° for the elevation angle assuming the viewpoint of human support robots.

To train the model, as is the case with most recent methods [16], [18], it is more realistic to use a large number of computer graphics images and a small number of real-world images rather than capturing an enormous number of the latter with various poses. Therefore, we performed experiments not only in an ideal setting with virtual-world images but also in a realistic setting with both virtual-world and real-world images.

A. DATASETS

We prepared two datasets for the evaluation; a large-scale virtual-world dataset and a real-world dataset.

1) LARGE-SCALE VIRTUAL-WORLD DATASET

Different from the dataset introduced in our previous work [4], we prepared a new large-scale virtual-world dataset that imitates a heavily cluttered scene by realistic occluding

objects. For the proposed data augmentation, three types of virtual settings; a target, an occluding object, and a target on a plane, were rendered. Depth images containing only a target, and a target on a plane were rendered using 3D models of 105 mugs from ShapeNet [23] as the virtual dataset. Depth images were rendered from the 3D model of each mug. For the rendering, a camera virtually rotated around the vertical axis with an interval of 1° while fixing the elevation angle. Images that even humans could not determine the object's pose were discarded. On the other hand, depth images of an occluding object were rendered from a total of 80 3D models; 20 objects for each of the four categories. The objects were virtually shifted along the y-axis in front of the target while fixing the elevation angle of the camera. They were then scaled to 384×384 pixels.

2) REAL-WORLD DATASET

Different from the dataset introduced in our previous work [4], we prepared a real-world dataset that imitates the heavily cluttered scenes by occluding objects actually captured in the real world. For this dataset, the training images and the evaluation images were generated differently. We used an RGB-D image sensor; Microsoft Azure Kinect, fixed 65 cm away from the target object, for capturing both color and depth images. For the training images, first, five different kinds of mugs were placed one by one on a turn-table embedded on a tabletop. All target images were captured while rotating the turn table around the vertical axis with an interval. Images that even a human could not determine the pose of the object were discarded. On the other hand, depth images of an occluding object were captured from a total of twelve actual objects, three objects in each of the four categories. For each object, depth images were captured by shifting the object along the y-axis in front of the target while fixing the elevation angle of the camera. Images containing 1) only the target and 2) only the occluding object, were generated by subtracting the background images containing only the tabletop plane from the captured images. The bounding boxes of the targets were obtained from the images containing only the target.

In contrast, for each evaluation image, a target object and another occluding object were captured together. Each mug in five different shapes was placed on a turn table one by one. The target images were captured while rotating the target object on the turn table around the vertical axis with an interval of 10° and changing the occluding object in each capture. Here, the locations of the mugs in the observation images are unknown in the pose estimation phase. Then, as preprocessing, a mug was first detected from each color image using YOLOv3 [24] and a bounding box was obtained. The depth images were cropped and scaled to 384×384 pixels with the bounding box located at the image center.

TABLE 1. Number of images used for the evaluation.

	Object	Setting 1	Setting 2
For network training	Target	27,100	28,455
	Occluding	2,400	2,760
For pose training	Target	40,650	47,425
	Occluding	2,400	2,760
For evaluation	Target	127	110
	Occluding	600	240

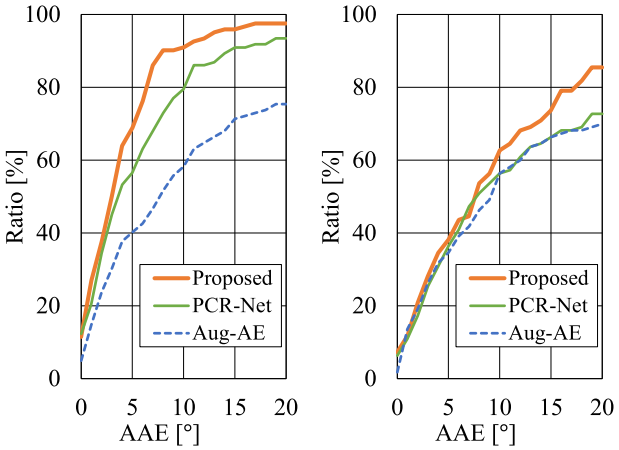


FIGURE 8. Ratios of samples with Absolute Angular Error (AAE) within *n* degrees.

B. EXPERIMENTAL SETTING

We evaluated the performances of the proposed method compared with two comparative methods 1 and 2.

- Comparative method 1 is based on Pose-CyclicR-Net [5], abbreviated as PCR-Net. Feature vectors are extracted as intermediate activations of a DCNN-based regression model trained with our dataset.
- Comparative method 2 is based on Augmented AutoEncoder [3], abbreviated as Aug-AE. Feature vectors are extracted as a vector encoded by the Augmented AutoEncoder trained with our dataset.

The two Encoder-Decoder models in the Two-stage Shape Reconstructor consist of the Encoder and the Decoder, which both are composed of five convolution layers. ReLU is used for each activation function. The intermediate feature map of the Generalized De-occluding AutoEncoder (GDAE) is $12 \times 12 \times 1,024$ dimensions to de-occlude the target object accurately, and the intermediate vector of the Median Shape Reconstructor has 512 dimensions to express each pose sparsely. Mean Squared Error (MSE) was used for the loss function and Adam [25] was used for the optimizer in the training phase.

We evaluated the methods in the following two settings: 1) Evaluation on the large-scale virtual-world dataset, and

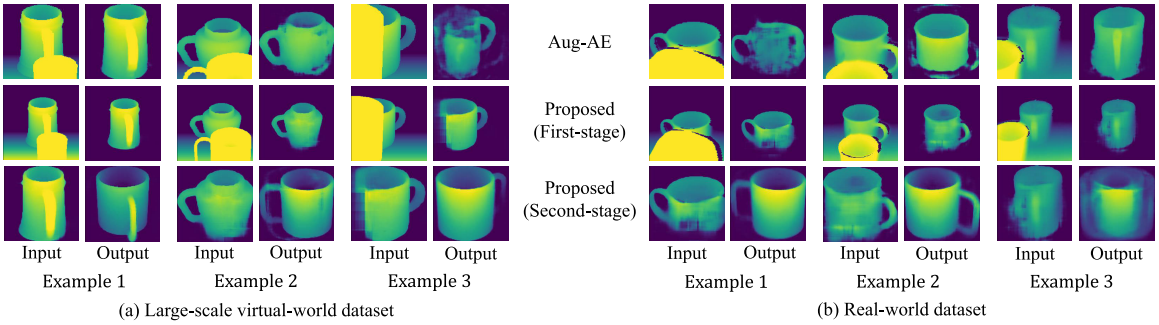


FIGURE 9. Visualization results of the target objects.

2) Evaluation on the real-world dataset. Table 1 shows the overview of the datasets. In Setting 1, we trained the models and evaluated the methods on the virtual-world dataset. For the Generalized Two-stage Shape Reconstructor training, we used depth images from the 3D models of 100 mugs as target images, and those of 80 objects as occluding object images. We trained the Reconstructor with the proposed data augmentation that randomly selects occluding objects and their positions. For the pose estimator training, we used depth images from the 3D models of 30 mugs among 100 mugs. In total, we used 40,650 generated images by the proposed data augmentation. For evaluation, we rendered 180 images from the five mugs and the 20 unknown occluding objects, which were not used in the training phase. We used 127 images out of 180, excluding images whose poses could not be estimated visually.

In Setting 2, we trained all the models in the virtual-world and the real-world datasets. For the Generalized Two-stage Shape Reconstructor training, we added five real mugs as the target objects and the twelve occluding objects to the dataset in Setting 1. For the pose estimator training, we similarly added the five real mugs. For the evaluation, we used 110 images captured from the five unknown real mugs and eight unknown occluding objects, which were not used in the training phase.

C. EVALUATION METRICS

To evaluate the difference between the object pose and the estimated rotation angle around the vertical axis, we calculated Absolute Angular Error (AAE) between the pose estimation result and the corresponding true pose, described in our previous work [4]. Based on AAE, we used three metrics; Mean Absolute Angular Error (MAAE), 95% Mean Absolute Angular Error (95MAAE), and the ratio of AAE within n° (w/in n° where $0 \leq n \leq 20$). 95MAAE is less susceptible to outliers (results with the estimated poses far apart from the true pose) compared to MAAE.

D. RESULTS

First, Table 2 shows the pose estimation results in Settings 1 and 2. The proposed method achieves the best performance

TABLE 2. Pose estimation results.

(a) Setting 1

Method	MAAE ↓	95MAAE ↓	w/in 5° ↑	w/in 10° ↑
PCR-Net	9.11°	5.82°	**64.6%**	78.7%
Aug-AE	20.93°	13.25°	48.8%	63.8%
Proposed	**7.43°**	**5.48°**	53.5%	**80.3%**

(b) Setting 2

Method	MAAE ↓	95MAAE ↓	w/in 5° ↑	w/in 10° ↑
PCR-Net	20.48°	13.79°	36.4%	56.4%
Aug-AE	28.49°	20.42°	34.5%	56.4%
Proposed	**14.25°**	**9.79°**	**38.2%**	**62.7%**

on most items. Next, Fig. 8 shows the w/in n° in Settings 1 and 2. The results demonstrate that the proposed method is effective on both virtual and real images.

We discuss the effectiveness of the proposed GDAE by comparing the de-occlusion results between GDAE and Aug-AE. Fig. 9 shows examples of the de-occlusion results of Aug-AE, GDAE, and the Median-shape Reconstructor (the second stage of the proposed method). Since Aug-AE could not estimate the offset between the true object center and the image center, it could not de-occlude the occluded region accurately when the target object was heavily occluded. Meanwhile, GDAE could remove the surrounding object and de-occlude the occluded region accurately by considering the offset in the training. In Addition, in most cases, the Median-shape Reconstructor could reconstruct the median-shaped object accurately from the de-occluded images.

V. CONCLUSION

In this paper, we proposed an occlusion-robust pose estimation method from a depth image for an unknown instance in an object category. To tackle this problem, our previous work proposed a model to extract features by de-occluding the occluded region of the target object and absorbing shape variations in a category. However, this method could not be used except in a situation where the segmentation mask of the visible part of the target object is known and the occluding contour is linearly shaped. Thus,

we extended this method to the Generalized Two-stage Shape Reconstructor that directly extracts features of an unknown occluded object from a depth image containing another unknown occluding object. Besides, we proposed a data augmentation method for training the Generalized Two-stage Shape Reconstructor effectively. Through the evaluation on large-scale virtual-world and real-world datasets, we confirmed that the proposed method successfully estimated the pose of an unknown occluded object from an observed depth image. In the future, we will extend the proposed method to 3D axis rotation and demonstrate object grasping using the proposed method and implement real-wrold applications.

REFERENCES

[1] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. 9th Int. Conf. Learn. Represent.*, May 2021, pp. 1–16.

[2] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[3] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proc. 15th Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 712–729.

[4] H. Tatemichi, Y. Kawanishi, D. Deguchi, I. Ide, A. Amma, and H. Murase, "Median-shape representation learning for category-level object pose estimation in cluttered environments," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4473–4480.

[5] H. Ninomiya, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, N. Kobori, and Y. Nakano, "Deep manifold embedding for 3D object pose estimation," in *Proc. 12th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, vol. 5, Mar. 2017, pp. 173–178.

[6] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *Int. J. Comput. Vis.*, vol. 14, no. 1, pp. 5–24, Jan. 1995.

[7] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 858–865.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[9] F. Tombari, S. Salti, and L. D. Stefano, "Unique signatures of histograms for local surface description," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 356–369.

[10] Y. Kawanishi, D. Deguchi, I. Ide, and H. Murase, "Object manifold embedding GAN for image generation by disentangling parameters into pose and shape manifolds," in *Proc. 28th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4473–4480.

[11] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

[12] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[13] A. V. Patil and P. Rabha, "A survey on joint object detection and pose estimation using monocular vision," in *Proc. Int. Joint Conf. Metall. Mater. Eng.*, vol. 227, Apr. 2019, pp. 1–11.

[14] R. T. Chin and C. R. Dyer, "Model-based recognition in robot vision," *ACM Comput. Surveys*, vol. 18, no. 1, pp. 67–108, Mar. 1986.

[15] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Found. Trends Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2007.

[16] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. 14th Conf. Robot., Sci. Syst.*, vol. 19, Jun. 2018, pp. 1–10.

[17] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1530–1538.

[18] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3338–3347.

[19] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "DualPoseNet: Category-level 6D object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3540–3549.

[20] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "GPV-Pose: Category-level object pose estimation via geometry-guided point-wise voting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6771–6781.

[21] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, Jul. 2008, pp. 1096–1103.

[22] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "OnePose: One-shot object pose estimation without CAD models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6815–6824.

[23] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

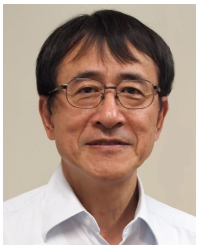
[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

HIROKI TATEMICHI received the B.Eng. and M.Inf. degrees from Nagoya University, Japan, in 2019 and 2021, respectively. In 2021, he joined Fujitsu Ltd. He is currently working as a Solution Engineer for government agencies. His research interest includes object pose estimation for robotic manipulation. He is a member of IEICE.

YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Postdoctoral Fellow with Kyoto University, in 2012. He moved to Nagoya University, Japan, as a Designated Assistant Professor, in 2014. In 2015, he became an Assistant Professor, and in 2020, he became a Lecturer there. Since 2021, he has been a Team Leader with the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and computer vision for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter. He is a member of IIEEJ and a Senior Member of IEICE.

DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow with Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor with the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor with Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. His research interests include object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.

ICHIRO IDE (Senior Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from The University of Tokyo, in 1994, 1996, and 2000, respectively. He became an Assistant Professor with the National Institute of Informatics, Japan, in 2000, and an Associate Professor with Nagoya University, Japan, in 2004, where he has been a Professor, since 2020. He was a Visiting Associate Professor with the National Institute of Informatics, from 2004 to 2010, an invited Professor with Institut de Recherche en Informartique et Systèmes Aléatoires (IRISA), France, in 2005, 2006, and 2007, and a Senior Visiting Researcher with ISLA, Instituut voor Informatica, Universiteit van Amsterdam, The Netherlands, from 2010 to 2011. His research interests include the analysis and indexing to authoring and generation of multimedia contents, especially in large-scale broadcast video archives and social media, mostly on news, cooking, and sports contents. He is a Senior Member of IEICE and IPS Japan, and a member of ACM, JSAI, and ITE.



HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. He has been a Professor with Nagoya University, since 2003, and a Professor Emeritus, since 2021. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He got a Medal with Purple Ribbon from the Government of Japan, in 2012. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of IAPR, IPSJ, and IEICE.