2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

TABLE I
POSE ESTIMATION RESULTS IN SETTING 1

	MAAE [°] ↓			95% MAAE [°] ↓			w/in 5°[%] ↑			w/in 10°[%] ↑		
Elevation angle [°]	10	30	50	10	30	50	10	30	50	10	30	50
PCR-Net [5]	3.56	4.14	13.91	3.07	3.73	7.06	74.4	71.8	50.7	96.6	**97.6**	76.3
AAE [2]	15.55	11.44	9.29	8.63	7.10	5.72	53.8	50.0	52.6	74.4	72.6	79.6
Proposed	**3.02**	**3.50**	**2.04**	**2.67**	**2.91**	**1.78**	**88.6**	**83.1**	**94.7**	**98.3**	96.0	**100.0**

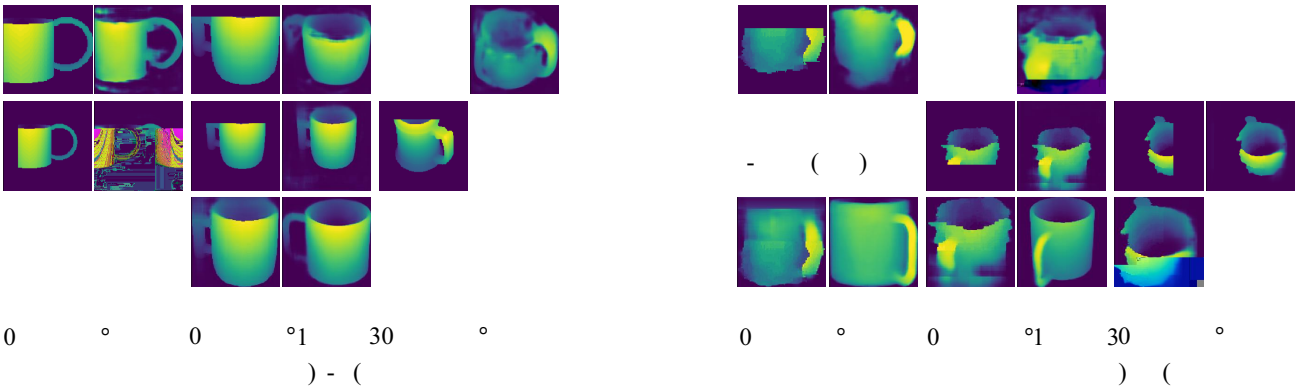


Fig. 9. Visualization results.

C. Metrics

To evaluate the estimated rotation angle around the vertical axis quantitatively, we calculated the absolute angular errors between the pose estimation results and the true poses. The Mean Absolute Angle Error (MAAE) was used as the metric. As the MAAE is often affected by outliers (results with the estimated poses far apart from the true pose), we also used the 95% MAAE as a metric. The metric is calculated while excluding the upper 5% outliers. By considering an application that a robot grasps an object, we also evaluated the ratio of absolute angle error within $5°$ (*w/in* $5°$) and $10°$ (*w/in* $10°$). Besides, we visualized the ratio of absolute angle error within $n°$ ($0 \leq n \leq 20, n \in \mathbb{Z}$) in graphs.

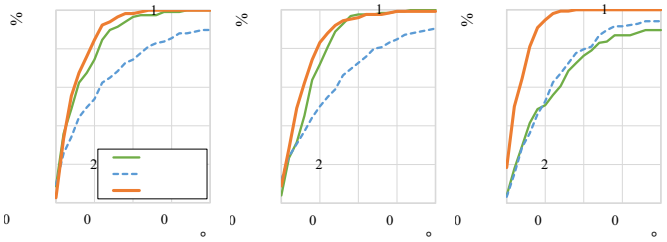
D. Comparative Methods

We evaluated two comparative methods A1 and A2, together with the proposed method.

- Comparative method A1 is based on Pose-CyclicR-Net [5], where features are extracted by a regression model based on a DCNN trained with our dataset. This method is abbreviated as PCR-Net.
- Comparative method A2 is based on Augmented Autoencoder [2], where features are extracted by the model based on Augmented Autoencoder (Fig. 4) trained with our dataset. This method is abbreviated as AAE.

E. Results

1) Setting 1: Fig. 9 shows visualization results using the Encoder-Decoder model. By comparative method A2 (AAE), the target object cannot be de-occluded accurately because such occlusions in the input image are not originally assumed.



(a) Angle: 10° (b) Angle: 30° (c) Angle: 50°

Fig. 10. Absolute angular error in setting 1.

On the other hand, in the first stage of the proposed method, the object can be de-occluded accurately by expanding the background and keeping the position of the unoccluded region of the object. As the second stage of the proposed method, the same pose of the median-shaped object can be reconstructed.

Table I shows the pose estimation results in setting 1. The proposed method achieves the best performance on most items. Fig. 10 shows the ratio of absolute angle error within $n°$ ($0 \leq n \leq 20, n \in \mathbb{Z}$) in setting 1. The proposed method is more effective than the comparative method for elevation angle $50°$.

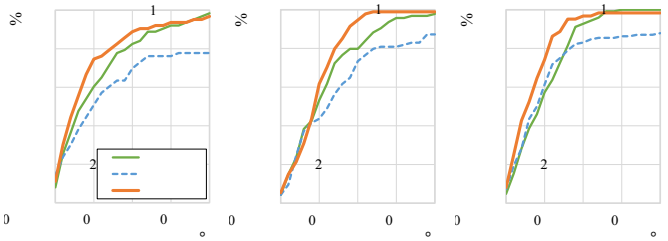
2) Setting 2: In Fig. 9, the de-occlusion of images in the real dataset is more difficult than that of images in a virtual dataset caused by the noisy observation values unique to the real dataset. Actually, by comparative method A2 (AAE), the occluded region is not well de-occluded in the case of the real dataset than that of the virtual dataset. On the other hand, by the proposed method, the Median-shape Reconstructor outputs the median-shaped object in the virtual dataset from the target object in the real dataset. The Encoder-Decoder model can

TABLE II
POSE ESTIMATION RESULTS IN SETTING 2

	MAAE [°] ↓			95% MAAE [°] ↓			w/in 5°[%] ↑			w/in 10°[%] ↑		
Elevation angle [°]	10	30	50	10	30	50	10	30	50	10	30	50
PCR-Net [5]	**5.81**	6.21	5.19	5.08	5.46	4.87	60.3	53.2	57.3	82.5	79.8	92.7
AAE [2]	19.42	17.85	14.18	14.27	10.61	8.03	50.8	43.6	61.3	69.8	73.4	83.1
Proposed	7.61	**5.15**	**4.27**	**3.98**	**4.67**	**3.19**	**74.6**	**61.7**	**74.2**	**88.9**	**94.7**	**96.8**

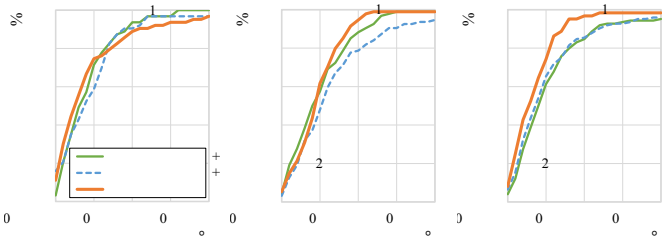
TABLE III
POSE ESTIMATION RESULTS FOR THE SECOND STAGE

	MAAE [°] ↓			95% MAAE [°] ↓			w/in 5°[%] ↑			w/in 10°[%] ↑		
Elevation angle [°]	10	30	50	10	30	50	10	30	50	10	30	50
DAE + PCR-Net	**4.60**	5.39	8.85	4.08	4.83	5.15	71.4	57.4	61.3	**93.7**	88.3	84.7
DAE + AAE	5.47	7.26	7.17	4.32	6.13	4.84	58.7	47.9	65.3	90.5	85.5	83.1
Proposed	7.61	**5.15**	**4.27**	**3.98**	**4.67**	**3.41**	**74.6**	**61.7**	**74.2**	88.9	**94.7**	**96.8**



(a) Angle: 10° (b) Angle: 30° (c) Angle: 50°

Fig. 11. Absolute angular error in setting 2.



(a) Angle: 10° (b) Angle: 30° (c) Angle: 50°

Fig. 12. Absolute angular error for the second stage in setting 2.

reconstruct the object accurately since the variation of the output images is smaller.

Table II shows the pose estimation results in setting 2. The proposed method achieves the best performance on most items also with the real dataset. Fig. 11 shows the ratio of absolute angle error within $n°$ $(0 \leq n \leq 20, n \in \mathbb{Z})$ in setting 2. The proposed method is also effective for the real dataset.

Through this evaluation, we demonstrated the effectiveness of the proposed two-stage Encoder-Decoder model.

V. DISCUSSION

A. *Effectiveness of the Median-shape Reconstructor*

To demonstrate the effectiveness of the Median-shape Reconstructor, we evaluated the second stage of the two-stage Encoder-Decoder model in setting 2. We evaluated two comparative methods B1 and B2, together with the proposed method. For each of the methods, the first stage of the two-stage model was fixed as the Deoccluding Autoencoder. The detail of the comparative methods is as follows:

- Comparative method B1 is based on Pose-CyclicR-Net [5], where the features are extracted by a regression model based on a DCNN trained with the dataset for the Median-shape Reconstructor. This method is abbreviated as DAE + PCRNet.
- Comparative method B2 is based on Augmented Autoencoder [5], where the features are extracted by the Augmented Autoencoder trained with the dataset for the Median-shape Reconstructor. The second stage model outputs the same pose of the target object instead of the same pose of the median-shape object. This method is abbreviated as DAE + AAE.

Table III shows the pose estimation results in setting 2. Fig. 12 shows the ratio of absolute angle error within $n°$ $(0 \leq n \leq 20, n \in \mathbb{Z})$. The Median-shape Reconstructor achieves the best performance except for the case of elevation angle 10°.

Firstly, we compare the proposed method with comparative method B1 (DAE + PCR-Net). The rotation angle represented by trigonometric functions is given when the PCR-Net is trained. On the other hand, the rotation angle and the median-shape is given as images when the Median-shape Reconstructor is trained. The Median-shape Reconstructor is more effective than the PCR-Net since the encoder part of the model can extract the features which are as close as possible to features extracted from an object instance, a median-shaped object.

Secondly, we compare the proposed method with comparative method B2 (DAE + AAE). The only difference between them is that the AAE outputs the same pose of the target object, but the Median-shape Reconstructor outputs the same pose of the median-shaped object.

TABLE IV
MAAE [°] IN MULTIPLE CATEGORIES

Category	Mug	Car	Bike	Chair
PCR-Net [5]	13.91	2.09	14.36	4.52
AAE [2]	9.29	13.48	61.93	12.99
Proposed	**2.04**	**0.60**	**10.37**	**2.44**

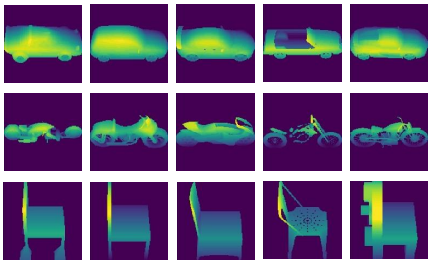


Fig. 13. Dataset with multiple categories.

Through this evaluation, we demonstrated the effectiveness of the proposed approach in which the proposed Encoder-Decoder model outputs an object instance.

B. Evaluation in Multiple Categories

We also evaluated the proposed method in multiple categories other than Mug; Car, Bike, and Chair in setting 1. The elevation angle was set to $50°$. Fig. 13 shows the examples of the dataset.

Table IV shows the Mean Absolute Angular Error (MAAE) in multiple categories. The proposed method achieves the best performance for all categories in the dataset. We confirmed that the proposed method is also effective in category-level.

VI. Conclusion

In this paper, we proposed an occlusion-robust pose estimation method of an unknown object instance in an object category from a depth image. To tackle the difficulty that occlusion causes an offset between the center of the actual object and its observable area, we proposed a two-stage Encoder-Decoder model that separates de-occlusion and feature extraction. To tackle the difficulty that different instances in a category may have different shapes, as the second stage, we proposed a Median-shape Reconstructor that is trained to decode the depth image containing the same pose of a representative shaped object in the category. By evaluating the method with both a large-scale virtual dataset and a real dataset, we confirmed the proposed method achieves good performance on pose estimation of an occluded object from a depth image. The part of the dataset used in the evaluation will be released.

As future work, we need to extend the method so that it can handle 3D axes rotations. For 3D pose estimation, our dataset must include an enormous number of depth images. We also need to improve the method to estimate the pose accurately with a smaller number of training images.

Acknowledgement

Parts of this research were supported by Grant-in-Aid for Scientific Research (JP17H00745).

References

[1] 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 *Proceedings of the 13th International Conference on Computer Vision*, Nov. 2011, pp. 858–865.

[2] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proceedings of the 15th European Conference on Computer Vision*, Part 6, Sept. 2018, pp. 699–715.

[3] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 3364–3372.

[4] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, Jan. 1995.

[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 *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 5, March 2017, pp. 173–178.

[6] M. Li and K. Hashimoto, "Accurate object pose estimation using depth only," *Sensors*, vol. 18, no. 4, pp. 1045:1–17, April 2018.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp. 77–85.

[8] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 1912–1920.

[9] A. V. Patil and P. Rabha, "A survey on joint object detection and pose estimation using monocular vision," Computing Research Repository, arXiv preprint arXiv:1811.10216, Nov. 2018.

[10] R. T. Chin and C. Dyer, "Model-based recognition in robot vision," *ACM Computing Surveys*, vol. 18, pp. 67–108, March 1986.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, Dec. 2012, pp. 1097–1105.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," in *Journal of Machine Learning Research*, vol. 11, Dec. 2010, pp. 3371–3408.

[13] A. X. Chang, T. A. Funkhouser, L. J. 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," Computing Research Repository, arXiv preprint, arXiv:1512.03012, Dec. 2015.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Computing Research Repository, arXiv preprint, arXiv:1804.02767, April 2018.