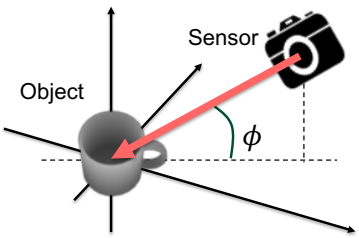

Fig. 12. Rendering environment and the elevation angle.

TABLE I
POSE ESTIMATION RESULTS (MEAN ABSOLUTE ERROR).

(i) By elevation angle ("Mug" class)

Elevation angle	60°	45°	30°	0°
Baseline	11.43	17.63	18.03	21.88
DA	10.37	18.02	17.83	21.99
Proposed (Ω-GAN)	**6.68**	**12.35**	**16.94**	**19.64**

(ii) By object class (Elevation angle 60°)

Object class	Mug	Car	Bike	Chair
Baseline	11.43	23.35	16.02	4.37
DA	10.37	14.58	18.72	3.77
Proposed (Ω-GAN)	**6.68**	**4.76**	**9.05**	**2.66**

B. Comparative Methods

There is a large number of methods for object pose estimation from an image. Currently, deep learning-based approaches are actively developed [30], [31], [32], [33]. Ninomiya et al. [34] proposed the Pose-CyclicR-Net for regression to cyclic objective variables using a convolutional neural network. The network outputs quaternion as the pose representation to handle the pose circularity.

We used a modified version of this Pose-CyclicR-Net as a baseline method. This network receives a depth image and outputs its pose parameter in a $(\cos\theta, \sin\theta)$ format, where θ is the rotation angle from the reference pose, instead of the quaternion representation as our dataset is restricted to the single-axis rotation.

For comparison, we trained the modified Pose-CyclicR-Net using different data as follows:

Baseline: Trained with the original depth images only.
DA: Trained with naïve data augmentation using the original depth images.
Proposed: Trained with the original depth images and generated depth images by the proposed -GAN.

In the case of DA, the original images were randomly shifted at most 10% of the image size and zoomed in the range of $[0.9, 1.1]$ and fed to the network. In the case of Proposed, the original images X_{real} and the generated images X_{fake}, which were randomly generated by the proposed -GAN, were used equally.

The original images X_{real} are annotated with the ground-truth poses Y_o. On the other hand, for each generated image in X_{fake}, the pose parameter $\mathbf{z}_p$ was used as the ground-truth pose. For training the pose estimator with various training samples, the parameters $\mathbf{z}_p$ and $\mathbf{z}_s$ were randomly sampled from the distributions over the pose and the shape manifolds.

The training procedure of the pose estimator is as follows. First, N_o images for a mini-batch were sampled from the original images X_{real}. Then, the same number of images in a batch were randomly generated by the generator G of the proposed -GAN and added to the mini-batch. Here, we applied a median filter to all the generated images to reduce the small noises. The training was repeated for 500 epochs.

C. Pose Estimation Results

The mean absolute error of the pose estimation results considering the pose circularity, *e.g.* the error between 5 and 355 is 10 , are shown in Table I (i).

For all the elevation angles, we confirmed that the proposed method, which is based on training with the images generated by the proposed -GAN, achieved the best performance. This is because the proposed -GAN successfully generated the poses and shapes not included in the training data. As the data captured from the small elevation angles were observed almost from the side, they contained depth images that were difficult to distinguish. Therefore, the pose estimation errors were relatively higher than in other situations.

We also evaluated the pose estimation accuracy for other object classes using the proposed -GAN. As with the "Mug" models, we also prepared several models such as "Car," "Bike," and "Chair" selected from the ShapeNet dataset [26]. They were rendered from the elevation angle of $\phi = 60$. We trained the proposed -GAN for each object class. The evaluation results are shown in Table I (ii). We confirmed that the proposed method is also effective for them.

From the results, we confirmed that the pose estimation results improved even though the generated depth images are not in high quality. If the image generation's quality is improved, we can expect that the pose estimation accuracy would also improve.

VII. CONCLUSION

We proposed the Object Manifold Embedding GAN (-GAN) that generates an image from a distribution in the pose and the shape manifolds. The generator of the proposed -GAN maps the parameters on these manifolds to images. For clearly disentangling these parameters, we also introduced Object Identity Loss to preserve the object instance's shape when only the pose parameter is changed.

We confirmed that the proposed -GAN could generate realistic images according to the pose and shape parameters through evaluation. For evaluating the pose accuracy of the generated images, we trained an object pose estimator with the generated images as data augmentation. We confirmed that the pose estimator trained with the generated images achieves

improved pose estimation accuracy compared to that trained with only the training images and naïve data augmentation, that is, the poses of the generated images are accurate enough for training an object pose estimator.

In the current work, the objects' rotaiton is restricted to a single-axis rotation; the generator of the -GAN samples a pose parameter from a distribution on a unit circle in the two-dimensional space. We plan to extend this to sample from a unit hypersphere, namely the two- or three-dimensional manifold, to handle more complicated rotations such as around two-dimensional or three-dimensional axes in the future.

ACKNOWLEDGMENT

Parts of this research were supported by MEXT (17H00745), Grant-in-Aid for Scientific Research.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. in Neural Info. Process. Syst. 27*, Dec. 2014, pp. 2672–2680.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Comput. Res. Repos.*, no. arXiv:1411.1784, Nov. 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

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

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *Comput. Res. Repos.*, no. arXiv:1701.07875, Jan. 2017. [Online]. Available: http://arxiv.org/abs/1701.07875

[5] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *Comput. Res. Repos.*, no. arXiv:1703.10717, Mar. 2017. [Online]. Available: http://arxiv.org/abs/1703.10717

[6] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (GANs): A survey," *IEEE Access*, vol. 7, pp. 36 322–36 333, Mar. 2019.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Comput. Res. Repos.*, no. arXiv:1511.06434, Nov. 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," *Comput. Res. Repos.*, no. arXiv:1611.07004, Nov. 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. 15th Int. Conf. on Comput. Vision*, Oct. 2017, pp. 2242–2251.

[10] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," *Comput. Res. Repos.*, no. arXiv:1610.09585, Oct. 2016. [Online]. Available: http://arxiv.org/abs/1610.09585

[11] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Adv. in Neural Info. Process. Syst. 29*, Dec. 2016, pp. 2172–2180.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. 2019 IEEE Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2019, pp. 4401–4410.

[13] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader Networks: Manipulating images by sliding attributes," in *Adv. in Neural Info. Process. Syst. 30*, Dec. 2017, pp. 5967–5976.

[14] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Multi-view 3D models from single images with a convolutional network," in *Proc. 14th European Conf. on Comput. Vision*, vol. 7, Oct. 2016, pp. 322–337.

[15] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang, "HoloGAN: Unsupervised learning of 3D representations from natural images," in *Proc. 17th Int. Conf. on Comput. Vision*, Oct. 2019, pp. 7588–7597.

[16] B. Liu, X. Wang, M. Dixit, R. Kwitt, and N. Vasconcelos, "Feature space transfer for data augmentation," in *Proc. 2018 IEEE Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2018, pp. 9090–9098.

[17] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," *Comput. Res. Repos.*, no. arXiv:1611.02163, 2016. [Online]. Available: http://arxiv.org/abs/1611.02163

[18] C. Xiao, P. Zhong, and C. Zheng, "BourGAN: Generative networks with metric embeddings," in *Adv. in Neural Info. Process. Syst. 31*, Dec. 2018, pp. 2269–2280.

[19] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proc. 2019 IEEE Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2019, pp. 1429–1437.

[20] Z. Shu, M. Sahasrabudhe, R. Alp Güler, D. Samaras, N. Paragios, and I. Kokkinos, "Deforming autoencoders: Unsupervised disentangling of shape and appearance," in *Proc. 15th European Conf. on Comput. Vision*, Sep. 2018, pp. 650–665.

[21] X. Xing, T. Han, R. Gao, S.-C. Zhu, and Y. N. Wu, "Unsupervised disentangling of appearance and geometry by deformable generator network," in *2019 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2019, pp. 10 346–10 355.

[22] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. 2019 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2019, pp. 5738–5746.

[23] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Adv. in Neural Info. Process. Syst. 6*, Dec. 1993, pp. 737–744.

[24] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. 2005 IEEE Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2005, pp. 539–546.

[25] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Department of Computer Science, Columbia University, Tech. Rep. CUCS-005-96, Feb. 1996.

[26] 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," *Comput. Res. Repos.*, no. arXiv:1512.03012, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1512.03012

[27] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *Comput. Res. Repos.*, no. arXiv:1805.08318, May 2018. [Online]. Available: https://arxiv.org/abs/1805.08318

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Adv. in Neural Info. Process. Syst. 29*, Dec. 2016, pp. 2234–2242.

[29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Adv. in Neural Info. Process. Syst. 30*, Dec. 2017, pp. 6626–6637.

[30] 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. 16th IEEE Int. Conf. on Comput. Vision*, Oct. 2017, pp. 1530–1538.

[31] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," in *Proc. 2015 IEEE Int. Conf. on Robot. and Autom.*, May 2015, pp. 1329–1335.

[32] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3D pose estimation," in *Proc. 2015 IEEE Conf. on Comput. Vision and Pattern Recognit.*, Jun. 2015, pp. 3109–3118.

[33] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," *Comput. Res. Repos.*, no. arXiv:1711.00199, Nov. 2017. [Online]. Available: https://arxiv.org/abs/1711.00199

[34] 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 Joint Conf. on Comput. Vision, Imaging and Comput. Graphics Theory and Appl.*, Feb. 2017, pp. 173–178.