Machine Learning Based Real-Time Industrial Bin-Picking: Hybrid and Deep Learning Approaches

Sukhan Lee*
Intelligent Systems Research Institute
Sungkyunkwan University
Suwon 16419, South Korea
Lsh1@skku.edu

Soojin Lee
Intelligent Systems Research Institute
Sungkyunkwan University
Suwon 16419, South Korea
christie74@skku.edu

***Abstract*— The real-time pick and place of 3D industrial parts randomly filed in a part-bin plays an important role for manufacturing automation. Approaches based solely on the conventional engineering discipline have been shown limitations in terms of handling multiple parts of arbitrary 3D geometries in real-time. In this paper, we present a machine learning approach to the real-time bin picking of randomly filed 3D industrial parts based on deep learning with/without hybridizing conventional engineering approaches. The proposed hybrid approach, first, makes use of deep learning-based object detectors configured in a cascaded form to detect parts in a bin and extract features of the parts detected. Then, the part features and their positions are fed to the engineering approach to the estimation of their 3D poses in a bin. On the other hand, the proposed sole deep learning approach is based on, first, extracting the partial 3D point cloud of the object from its 2D image with the background removed and then transforming the extracted partial 3D point cloud to its full 3D point cloud representation. Or, it may be based on directly transforming the object 2D image with its background removed to the 3D point cloud representation. The experimental results demonstrate that the proposed approaches are able to perform a real-time multiple part bin picking operation for multiple 3D parts of arbitrary geometries with a high precision.**

Keywords*— *Bin Picking, 3D Objects/Parts, Deep Learning Network, 3D Pose Estimation

I. INTRODUCTION

Vision guided automated bin picking, especially, for a pick and place operation of industrial parts arbitrary stacked in a bin, plays an important role for flexible manufacturing in industry 4.0. In spite of the progress in this area to date, currently, vision guided automated bin picking is limited to the number as well as the 3D geometry of parts it can handle. In this paper, we present a hybrid approach of deep learning and engineering as a means of breaking through the current limitations of industrial bin picking toward enabling the real-time pick and place of multiple 3D parts of arbitrary geometries.

Conventional approaches to vision guided automated bin picking has been taking advantage of the progress in computer vision technologies, notably, the technologies related to feature extraction and matching as well as pose estimation. However, conventional computer vision technologies are shown not only inefficient but also ineffective in dealing with 3D in terms of object detection, segmentation and pose estimation. Li et al.[1] proposed 3D object recognition and pose estimation for random bin-picking using the partition viewpoint feature histograms that can be applicable to 3D point cloud data. The authors pre-sampled point cloud by using a voxel grid filter, based on which 3D objects were segmented based on the distance between point clouds. As the post processing, ICP is used to refine object poses. Mitash et al.[2] presented robust 6D object pose estimation with stochastic congruent sets to improve the precision of object semantics segmentation. As mentioned, this conventional computer vision approach to dealing with 3D is rather time consuming and lacking generality. Recent advancement of deep learning approach to computer vision enables a dramatic improvement of many difficulties conventional computer vision approaches have encountered, including object detection, segmentation, feature extraction, etc. Furthermore, the application of deep learning to computer vision is expanding rapidly from 2D to 3D[4-15], as seen by the development of PointNet[4] as well as 3D semantic and instance segmentation. Lee et al.[3] presented 3D pose estimation of bin picking object using deep learning and 3D matching, where they used the deep learning based object detector, YOLO V2, for detecting and recognizing objects, while estimating their poses by ICP.

The proposed approach, first, makes use of deep-learning based object detectors configured in a cascaded form for both detecting parts in a bin and extracting features associated with the individual parts detected. The concatenation of the part label and the feature labels and their positions associated with the part allows the subsequent part net to have its part recognition rate close to 100%, Furthermore, the part features and their positions are to be fed directly into the estimation of the 3D pose of the corresponding part in a bin as well as its degree of occlusion. The experimental results demonstrate that the proposed approach is able to perform a real-time multiple part bin picking operation for multiple 3D parts of arbitrary geometries with a high precision.

II. OVERVIEW OF THE PROPOSED APPROACH

The proposed automated bin picking system consists of the following major components, as illustrated by Fig. 1: 3D camera providing the 2D image and 3D point cloud of the bin, the Cascaded Object Detector formed by a serial connection of object/part detector net and object/part feature detector net both based on YOLO v3, Part Net that corrects errors in object/part labeling by the object detector net by incorporating the part features detected by object/part feature detector net, Geometric Feature Extractor that extracts the geometric features of object/part by applying simple 2D feature extraction to the detected part features, Object to CAD matching that matches the detected part features with the pre-defined CAD model features to estimate object pose, and ICP to refine the object pose precisely.

978-3-9819263-5-4/DATE21/©2021 EDAA

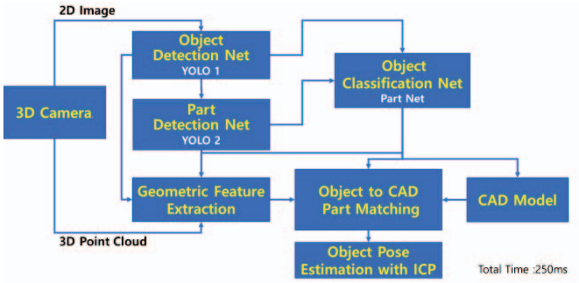


Fig. 1. Hybrid Deep Learning Network Based Automated Bin-Picking System

III. Cascaded Part and Part Feature Detection Network

We use YOLO v3 to first detect objects/parts filed in a bin and then, for individual objects/parts, we apply YOLO v3 again to extract the part features associated with individual objects/parts, as illustrated by Fig. 2 and 3.

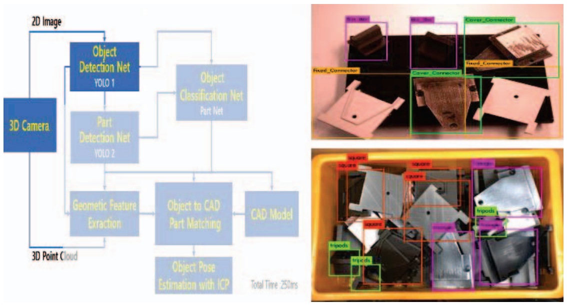


Fig. 2. YOLO v3 Based Object/Part Detector Net and the result of Object/Part Detected

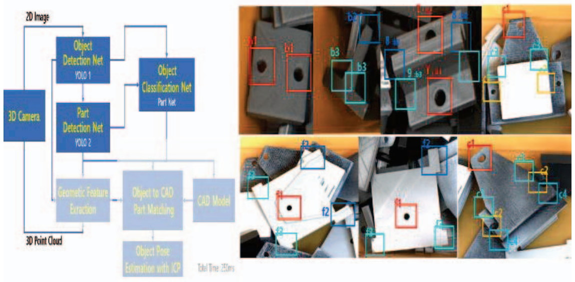


Fig. 3. YOLO v3 Based Object/Part Feature Detector Net and the result of Object/Part Features Detected

IV. Feature Matching and 3D Pose Estimation

A. Conventional Engineering Vision Approach

The part features extracted from the cascaded object/part and part feature detectors are conveniently used to extract the point and line features involved in individual part features that can be assigned with 3D data by incorporating 3D point cloud data from 3D camera. The extracted part features are to be used for 3D pose estimation of the objects/parts in the later process. Refer to Fig. 4 for the extracted part features.

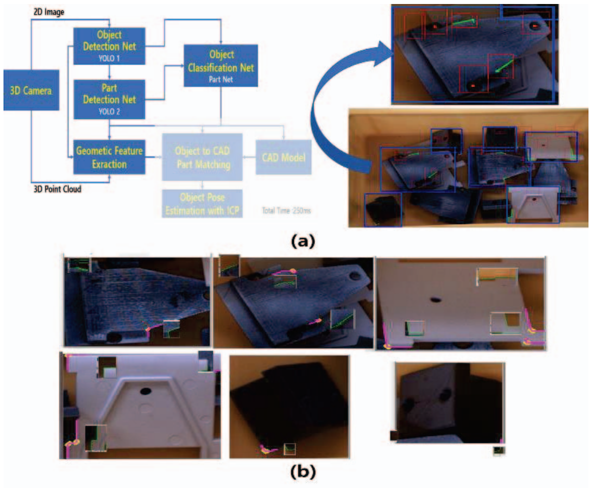


Fig. 4. (a) and (b): Point and Line Feature Extraction Based on the Object/Part Features Detected by Object/Part Feature Detector Net

B. Part Matching with CAD Model

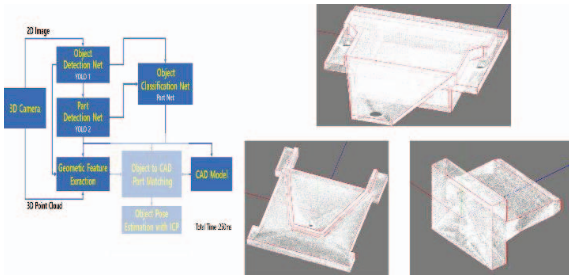


Fig. 5. Point and Line Feature Extraction Based on the Object/Part Features Detected by Object/Part Feature Detector Net

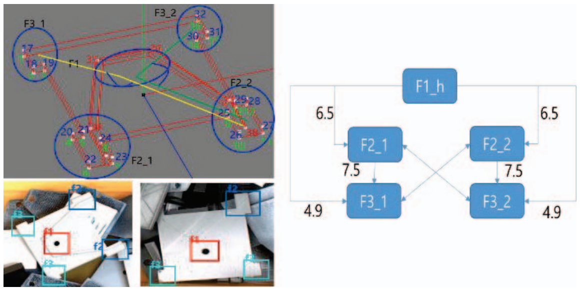


Fig. 6. Matching between the extracted part features and the pre-defined CAD features for pose estimation

Based on the extracted part features, we can approximated estimate the object pose by matching them with the pre-defined CAD features, as illustrated by Figs. 5 and 6.

C. Deep Learning Approach

Instead of the engineering vision approach for object pose estimation described above, we can use deep learning approach to pose estimation. To this end, we developed deep learning based image processing and transformation methods: background removal and Partial-to-Full Point Cloud transformation as well as 2D-to-3D transformation. As shown in Fig. 7, the deep learning based object pose estimation we propose may be based on, first, extracting the partial 3D point cloud of the object from its 2D image with the background removed and then transforming the extracted partial 3D point cloud to its full 3D point cloud representation. Or, it may be based on directly transforming the object 2D image with its background removed to the 3D point cloud representation as shown by a separate branch marked in green color in Fig. 7.

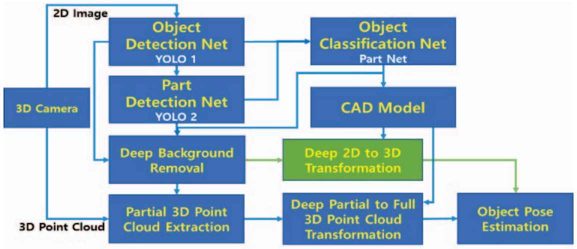


Fig. 7. Recognition and pose estimation of 3D industrial parts solely based on the deep learing approach to partial-to-full 3D point cloud transformation or to 2D-to-3D transformation(green).

C-1. Background Removal of Object/Part Images

The removal of unwanted background is useful for applying deep learning network to extract 3D pose of object/part from 2D images. The CM-GAN[16] of Fig. 8 can be directly used for such background removal by masking the backgrounds from the respective images with unwanted backgrounds. To demonstrate this, we applied the proposed approach to the removal of unwanted backgrounds from the images of industrial parts that are detected from an industrial robotic bin-picking environment, as shown in Fig. 8. For the details of training CM-GAN, refer to [16].

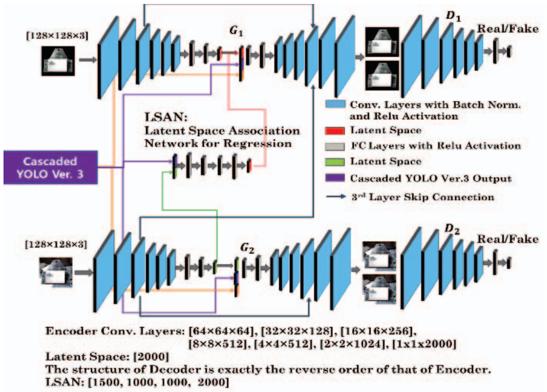

[128×128×3]
G_1
D_1
Real/Fake
LSAN:
Latent Space Association Network for Regression
Cascaded YOLO Ver. 3
Conv. Layers with Batch Norm. and Relu Activation
Latent Space
FC Layers with Relu Activation
Latent Space
Cascaded YOLO Ver.3 Output
3rd Layer Skip Connection
[128×128×3]
G_2
D_2
Real/Fake
Encoder Conv. Layers: [64×64×64], [32×32×128], [16×16×256], [8×8×512], [4×4×512], [2×2×1024], [1x1x2000]
Latent Space: [2000]
The structure of Decoder is exactly the reverse order of that of Encoder.
LSAN: [1500, 1000, 1000, 2000]

Fig. 8. CM-GAN for Background Removal. An Image with Unwanted Background is input to G2 and the corresponding Image without Background is output from G2 through LSAN.

Fig. 9 (b) illustrates typical instances of the part images collected as training and testing samples with unwanted backgrounds. This results in the part dataset consisting of 450 training and 30 testing samples with unwanted backgrounds.

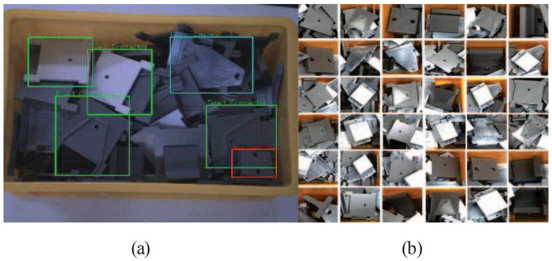

(a) (b)

Fig. 9. Parts detected and segmented from a bin by YOLO v3 for Bin-Picking Operation. (a) A Bin-Picking Image with Parts detected from YOLO v3. (b) Part Database collected from the Output of YOLO v3.

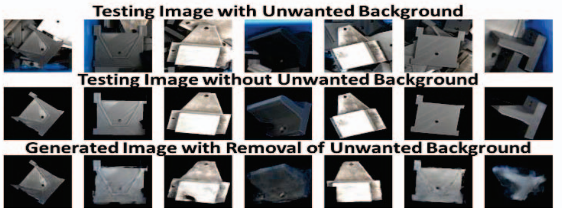

Testing Image with Unwanted Background
Testing Image without Unwanted Background
Generated Image with Removal of Unwanted Background

Fig. 10. The Result of Background Removal Experiment. The third row shows the images generated by CM-GAN after the removal of unwanted backgrounds.

The first and second rows of Fig. 10 illustrate, respectively, the images with and without backgrounds used for the ground truth testing samples. The third row of Fig. 10 shows the output images after the removal of unwanted backgrounds by CM-GAN. They demonstrate the quality and efficiency in the performance of background removal offered by CM-GAN, which may lead to various applications to vision-based robotic tasks for industry.

C-2. Deep Partial-to-Full 3D Point Cloud Transformation

With the background removed bounding box images of objects/parts available from the above step, we can obtain the corresponding partial 3D point cloud representation of parts. This is done by collecting the 3D points corresponding only to those pixels of the part image that do not belong to the removed background, as illustrated by Fig. 11. Note that the 3D point cloud thus obtained represents a partial 3D point cloud representation seen from a particular camera view point. Examples of the partial 3D point cloud representation of parts obtained from their bounding box images are illustrated in Fig. 11.

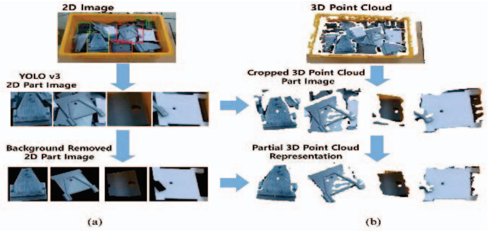


Fig. 11. The partial 3D point cloud representation of individual parts detected from the bin. (a) The unwanted background is removed out from the detected bounding box image of a part. (b) The 3D point cloud of the detected part is extracted from the 3D point cloud of the bin by collecting those 3D points that corresponds only to the foreground pixels of its bounding box image with the unwanted background removed.

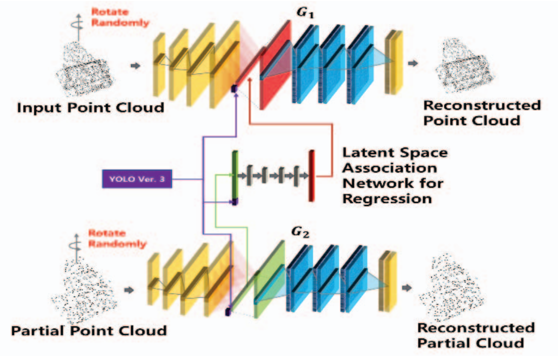
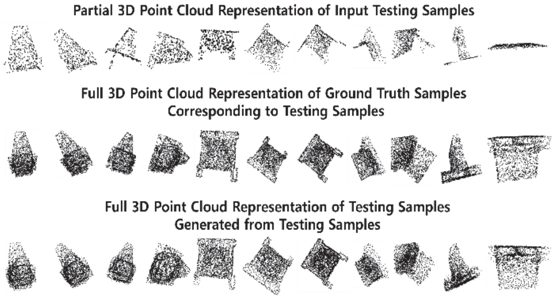

Rotate Randomly
G_1
Input Point Cloud
Reconstructed Point Cloud
YOLO Ver. 3
Latent Space Association Network for Regression
Rotate Randomly
G_2
Partial Point Cloud
Reconstructed Partial Cloud

Fig. 12. Latent space association for partial-to-full 3D point cloud transformation.

As the next step, we introduce the dual associative point AEs, which is CM-GAN without discriminators, as shown in Fig. 12, where the two point AEs, G1 and G2, take, respectively, partial and full 3D point cloud representations of parts as their inputs. The dual associative point AEs is to learn the transformation from the partial to the full 3D point cloud representations of parts through LSAN, similar to the CM-GAN for image completion and masking. The LSAN is to transform the global features learned by the encoder of partial 3D representations to those by the full 3D representations as the input to the decoder of full 3D representations. Note that we chose the Bias-Induced Point AE [17] for the implementation of dual associative point AEs. However, any of the point AEs currently available, including the original Point Net [4] and Folding Net [5], would equally serve the purpose. In Fig 12, instead of attaching a classification network to G1 and G2 to learn the part class, we used the part class information obtained from the CM-GAN for background removal and concatenated it to the inputs of the G1 and G2 decoders as well as of LSAN.


Partial 3D Point Cloud Representation of Input Testing Samples

Full 3D Point Cloud Representation of Ground Truth Samples Corresponding to Testing Samples

Full 3D Point Cloud Representation of Testing Samples Generated from Testing Samples

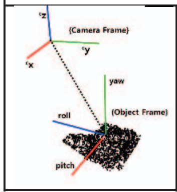
Fig. 13. Full 3D point cloud representation of industrial parts generated by dual associative point AEs based partial-to-full transformation.

TABLE Ⅰ. AVERAGE CHAMFER DISTANCE(CD) ERROR OF FULL 3D POINT CLOUD REPRESENTATION OF RECONSTRUCTED BY DUAL ASSOCIATIVE POINT AEs BASED PARTIAL-TO-FULL TRANSFORMATION.

	Average Chamfer Distance (CD) Error
Partial-to-Full Transformation by Dual Associative Point AEs (m^2)	0.0098

Fig. 13 illustrates the partial to full 3D point cloud transformation based on the dual associative point AE of Fig. 12. The partial 3D point cloud representation of industrial parts, as shown in the first row of Fig. 13, is shown to successfully generate the corresponding full 3D point cloud representation at the third row of Fig. 13, where the partial 3D point cloud representation is obtained by applying the background removed part images to the 3D point cloud representation of a bin picking scene. Table Ⅰ shows the quantitative evaluation of the accuracy involved in the partial to full 3D point cloud transformation based on the dual associative point AEs. The accuracy is measured by the average Chamfer distance between the reconstructed and the ground truth point cloud, where 40,000 training and 16,000 testing data, collected at various positions and orientations of the industrial parts in the bin, are used for evaluation. The full point cloud representation of parts reconstructed from the partial point clouds are then input to a separate classification network to obtain their position and orientation with reference to the camera frame. The classifier is configured with an encoder having the same structure as that used in the dual associative point AEs plus a fully-connected network on top of the latent space of the encoder. The output of the classifier consists of 23 classes, Roll 5, Pitch 5 and Yaw 13, for orientation and 25 classes, X 10, Y 10 and Z 5, for position. Note that the number of classes are determined in such a way as to compromise the precision with the complexity in representation and training. As such, we consider the possible range of the position and orientation variations in the bin to come up with the class layout. In particular, the range of position is defined in terms of the possible discrepancy of the center of partial point clouds to the ground truth. This comes up with the 15(Roll and Pitch) and 30(Yaw) degree separations for orientation and the 0.78(X), 0.79(Y) and 1.1(Z) cm separations for position. Table Ⅱ shows the accuracy of estimating the pose of industrial parts in the bin based on the average MSE for the 16,000 testing data. Refer to the left-most figure in the table for the definition of the coordinate system used in the experiment. Fig. 14 illustrate the 3D position and orientation of the parts in the bin, using their CAD models, that are obtained by the CM-GAN based 3D pose estimation process described above.

TABLE Ⅱ. AVERAGE MEAN SQUARE ERROR(MSE) OF GENERATED FULL 3D POINT CLOUD REPRESENTATION OF INDUSTRIAL PARTS POSITION AND ORIENTATION.

		Average Mean Square Error (MSE)		
 ᶜz {Camera Frame} ᶜy ᶜx yaw roll {Object Frame} pitch	Position Error (cm)	ΔX_0	ΔY_0	ΔZ_0
		0.21	0.20	0.33
	Orientation Error (degree)	ΔRoll	ΔPitch	ΔYaw
		16.82	15.15	21.04

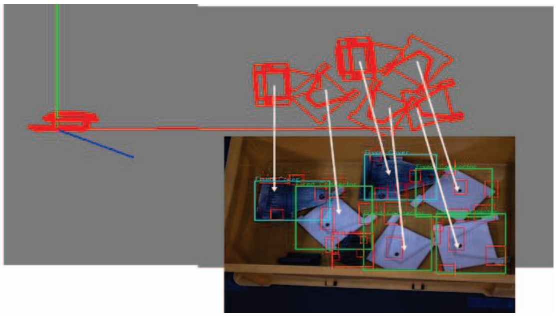


Fig. 14. The CAD representation of parts with the estimated part poses are overlapped with the 3D point cloud representation of a bin picking scene.

C-3. Deep 2D-to-3D Transformation

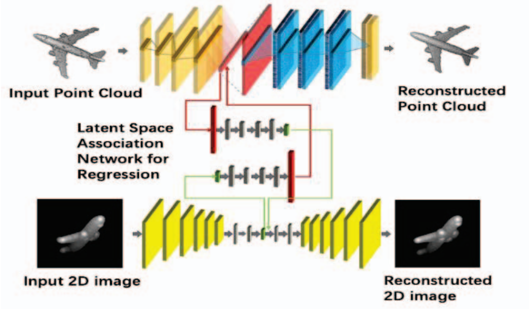

Input Point Cloud
Reconstructed Point Cloud
Latent Space Association Network for Regression
Input 2D image
Reconstructed 2D image

Fig. 15. Dual Auto-Encoder Architecture for generating the 2D-to-3D and 3D-to-2D transformations. The top-side network is a standard 2D auto-encoder structure which can extract the latent vector of each 2D image. The bottom-side network is our proposed Point AE

The 2D images with their backgrounds removed by the above background removal process can be input to the deep learning based 2D-to-3D transformation network, as shown in Fig. 15, For the reconstruction of 2D-to-3D, we first input the projected 2D image to the 2D encoder to obtain its feature vector, and then connect to the 2D-to-3D association network to predict the global feature of the corresponding 3D point cloud. Finally, use 3D decoder to reconstruct the point cloud with the predicted global feature as input.

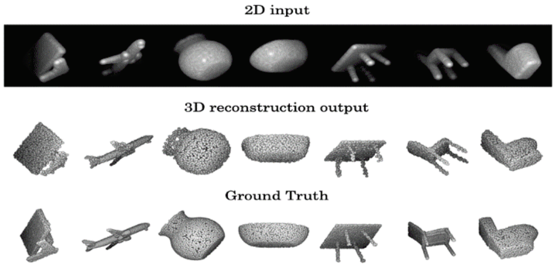


Fig. 16. 2D-to-3D transformation. The input is the 2D image that captured by a virtual camera in 3D space from a specific 3D testing object point cloud. The output is the transformation output of corresponding 2D projection image.

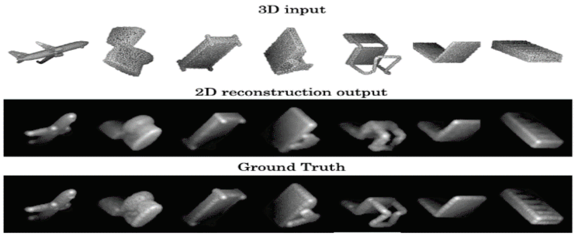


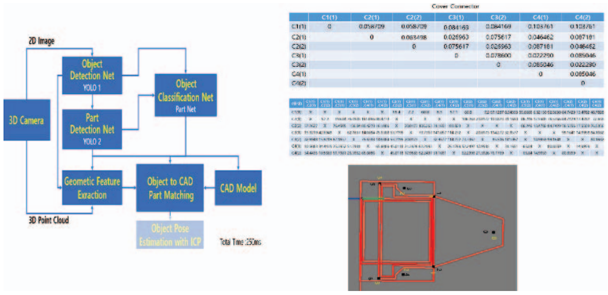
Fig. 17. 3D-to-2D transformation. The input is the raw 3D point cloud data from testing dataset. The output is the transformation output in 2D space.

The reconstruction of 3D-to-2D is the reverse of the above process. Firstly, the original test set point cloud is input to the 3D encoder to obtain a global feature, and then input to the 3D-to-2D association network to predict the corresponding 2D image feature vector. Finally, the 2D decoder is used for image reconstruction.

TABLE III. AVERAGE 2D-TO-3D RECONSTRUCTION LOSS APPLIED IN WHOLE SHAPENET16 TESTING SET WITH CHAMFER DISTANCE METRIC.

Average 2D-to-3D loss	0.014

V. 3D POSE ESTIMATION WITH ICP



	C1(1)	C2(1)	C2(2)	C3(1)	C3(2)	C4(1)	C4(2)
C1(1)	0	0.058709	0.058709	0.084169	0.084169	0.103761	0.103761
C2(1)		0	0.063498	0.026963	0.075617	0.046462	0.087181
C2(2)			0	0.075617	0.026963	0.087181	0.046462
C3(1)				0	0.078600	0.022290	0.085046
C3(2)					0	0.085046	0.022290
C4(1)						0	0.085046
C4(2)							0

Fig. 18. The fine-tuning of 3D object/part pose based on ICP.

The 3D poses estimated based on part feature matching are refined by ICP as the last step for the precision pose estimation. The results are shown in Figs. 18 and 19.

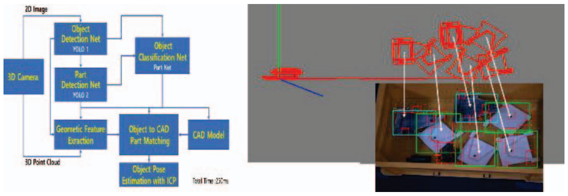


Fig. 19. Overlay of CAD model based on the 3D object/part poses refined by ICP.

VI. EXPERIMENT

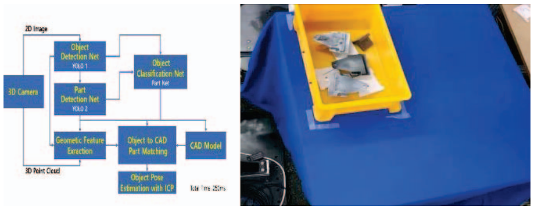


Fig. 20. Experimental setup of bin picking operation with 3 obejct/part.

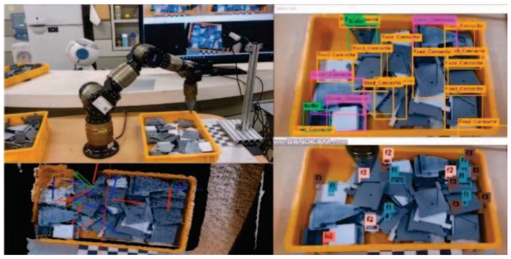


Fig. 21. Actual robotic bin picking operation for experiment.

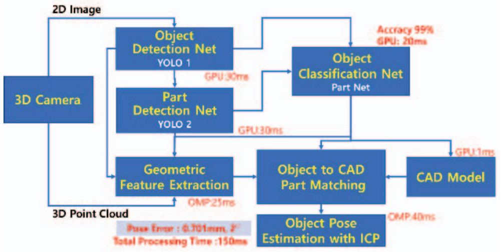


Fig. 22. Computational time of individual components for real-time bin picking operation.

We have conducted a real experiment with 3 industrial parts of arbitrary geometric shapes, as shown in Fig. 20 and 21. Fig. 22 shows that the proposed bin picking system can be operated in real-time as indicated by the computational time of individual components.

Fig. 23 shows the laboratory set-up for the experimental demonstration of the proposed CM-GAN based real-time robotic bin picking operations. Experiments show that the accuracy as listed in Table II has been sufficient for a successful bin picking operation when a vacuum gripper is used for the pick and place operation, where less than 200 ms is consumed to estimate the 3D part poses. However, in the case where we need a higher precision than that reported in Table II for picking, we can always go one more step further by applying ICP to the estimated 3D pose for the final refinement.

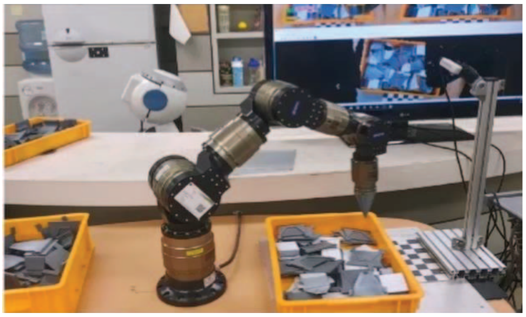


Fig. 23. A robotic bin picking scene that is set up to experimentally verify the application of the proposed CM-GAN to the 3D pose estimation of industrial parts in a bin

Experiments demonstrate that the CM-GAN application to robotic bin picking is able to perform real-time bin picking of multiple industrial parts of arbitrary 3D geometries with a high precision.

VII. Conclusion

This paper demonstrate the real-time pick and place of 3D industrial parts randomly filed in a bin based on a hybrid approach of deep learning and engineering as well as on a sole deep approach. We aim at making a breakthrough to overcome the current limitations of industrial bin picking by enabling the real-time pick and place of multiple 3D parts of arbitrary geometries. The experimental results demonstrate that the proposed approaches are able to perform a real-time multiple part bin picking operation for multiple 3D parts of arbitrary geometries with a high precision.

Acknowledgment

This work was supported in part by the 3D Recognition Project of Korea Evaluation Institute of Industrial Technology (KEIT) under Grant 10060160, in part by the "Deep Learning Based Cross-Sensory Transfer for Visually Impaired. " project of National Research Foundation (NRF) of Korea (2020R1A2C200956811) sponsored by Korea Ministry of Science and ICT, and in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government(MSIT) by AI Graduate School Support Program under Grant 2019-0-00421.

References

[1] Li, Deping, et al. "3D object recognition and pose estimation for random bin-picking using Partition Viewpoint Feature Histograms." Pattern Recognition Letters 128 (2019): 148-154.

[2] Mitash, Chaitanya, Abdeslam Boularias, and Kostas Bekris. "Robust 6D object pose estimation with stochastic congruent sets." arXiv preprint arXiv:1805.06324 (2018).

[3] Lee, J., Kang, S. and Park, S-Y, 3D Pose Estimation of Bin Picking Object using Deep Learning and 3D Matching, In Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics(ICINCO 2018) - Volume 2, pages 318-324

[4] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[5] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[6] Fan H, Su H, Guibas L J. A point set generation network for 3D object reconstruction from a single image. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 605-613.

[7] Choy C B, Xu D, Gwak J Y, et al. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. European conference on computer vision. Springer, Cham, 2016: 628-644.

[8] Wu Z, Song S, Khosla A, et al. 3D shapenets: A deep representation for volumetric shapes. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1912-1920.

[9] Smith E, Meger D. Improved adversarial systems for 3D object generation and reconstruction. arXiv preprint arXiv:1707.09557, 2017.

[10] Zamorski M, Zięba M, Nowak R, et al. Adversarial Autoencoders for Generating 3D Point Clouds. arXiv preprint arXiv:1811.07605, 2018.

[11] Achlioptas P, Diamanti O, Mitliagkas I, et al. Learning representations and generative models for 3D point clouds. arXiv preprint arXiv:1707.02392, 2017.

[12] Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3D shape descriptors. Symposium on geometry processing. 2003, 6: 156-164.

[13] Wu J, Zhang C, Xue T, et al. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. Advances in neural information processing systems. 2016: 82-90.

[14] Achlioptas P, Diamanti O, Mitliagkas I, et al. Representation learning and adversarial generation of 3D point clouds. arXiv preprint arXiv:1707.02392, 2017, 2(3): 4.

[15] Islam N U, Lee S. Learning Typical 3D Representation from a Single 2D Correspondence Using 2D-3D Transformation Network. International Conference on Ubiquitous Information Management and Communication. Springer, Cham, 2019: 440-455.

[16] LEE, Sukhan; ISLAM, Naeem Ul; LEE, Soojin. Robust image completion and masking with application to robotic bin picking. Robotics and Autonomous Systems, 2020, 103563.

[17] Cheng Wencan, and Sukhan Lee. "Point Auto-Encoder and Its Application to 2D-3D Transformation." International Symposium on Visual Computing. Springer, Cham, 2019.