Fine-grained Object Categorization for Service Robots

Songsong Xiong
Department of Artificial Intelligence
University of Groningen
Groningen, Netherlands
s.xiong@rug.nl

Hamidreza Kasaei
Department of Artificial Intelligence
University of Groningen
Groningen, Netherlands
hamidreza.kasaei@rug.nl

***Abstract*—A robot working in a human-centered environment is frequently confronted with fine-grained objects that must be distinguished from one another. Fine-grained visual classification (FGVC) still remains a challenging problem due to large intra-category dissimilarity and small inter-category dissimilarity. Furthermore, flaws such as the influence of illumination and information inadequacy persist in fine-grained RGB datasets. We propose a novel deep mixed multi-modality approach based on Vision Transformer (ViT) and Convolutional Neural Network (CNN) to improve the performance of FGVC. Furthermore, we generate two synthetic fine-grained RGB-D datasets consisting of 13 car objects with 720 views and 120 shoes with 7200 sample views. Finally, to assess the performance of the proposed approach, we conducted several experiments using fine-grained RGB-D datasets. Experimental results show that our method outperformed other baselines in terms of recognition accuracy, and achieved 93.40 % and 91.67 % recognition accuracy on shoe and car dataset respectively. We made the fine-grained RGB-D datasets publicly available for the benefit of research communities. The video is available at https://youtu.be/c8Tqy6uLV08**

***Index Terms*—Fine-grained visual classification, RGB-D dataset, Vision transformer, Deep convolutional neural network**

I. Introduction

Nowadays, fine-grained visual categorization received much attention due to its widespread application in various fields, such as intelligent retail [1], [2], automatic biodiversity monitoring [3], [4] and etc. The FGVC aims to differentiate category instances from numerous subcategories subordinating the basic-level categories, for example, classifications of flowers and fruits, the species of birds and dogs, and the different models of cars. The FGVC, however, still is confronted with crucial challenges for two reasons. The first one is to discriminate fine-grained objects due to the large intra-category dissimilarity and the small inter-category variance [5](see Fig. 1). Another reason is the lack of a large number of fine-grained RGB-D objects dataset that can be used to train a deep learning based model.

In traditional FGVC, many studies leveraged the RGB fine-grained datasets, such as CUB-200-2011 [6], Oxford Flowers [7], Aircraft [8], and Pets [9]. Because these data are gathered from various sources on the internet, they have some inherent flaws. The external information of the object is, therefore, essential to obtain a more detailed representation for object recognition.

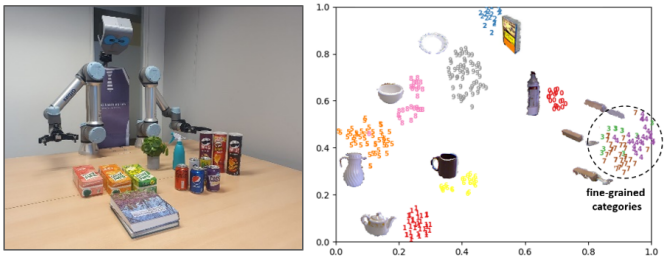

Fig. 1. An example of fine-grained object visual classification (FGVC) scenario: (*left*) the dual arm robot perceives the environment through an RGB-D camera, and performs the FGVC tasks; (*right*) The t-SNE plot displays the distribution of various object categories across the feature space. This plot, in particular, demonstrates that distinguishing knife, fork, and spoon from each other (fine-grained-level recognition) is more difficult than distinguishing mug from bottle (basic-level recognition).

As RGB-D sensor technology advances, the cost of RGB-D devices has gradually decreased in recent years. As a result, many studies are conducted using both RGB and depth sensors for computer vision, for example, object classification [10]–[12], action recognition [13], and object detection [14], [15].

Such studies revealed that when we consider both RGB and depth information, we can learn a better representation and achieve a higher recognition accuracy. To the best of our knowledge, there is no FGVC RGB-D object datasets, except for GUN-71 [16] for fine-grained hand-grasp classification and non-public FGBD-FG [17] for vegetables and fruits.

In this paper, to improve the recognition accuracy for fine-grained objects, we develop a novel deep mixed multi-modality approach based on CNN-ViT networks. An overview of our approach is shown in Fig. 2. To train and evaluate the model, we generate synthetic RGB-D FGVC datasets with shoes and cars, inspired by [18]. To assess the performance of the proposed approach, we performed extensive sets of experiments. Experimental results show that our mixed multi-modality approach surpassed the selected state-of-the-art approaches regarding recognition accuracy. In summary, our key contributions are twofold:

- We propose a deep mixed multi-modality method based on CNN-ViT networks to perform FGVC.
- To the best of our knowledge, we are the first group

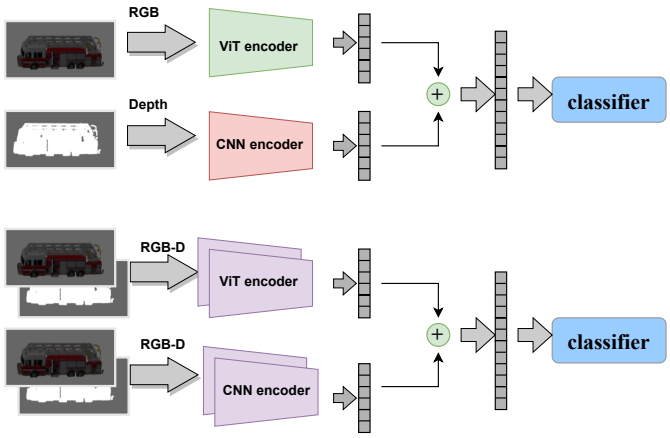

Fig. 2. Multi-modal deep learning framework for RGB-D fine-grained classification: (*top-row*) Pipeline of mixed Multi-model sturcture with RGB or Depth; (*lower-row*) Pipeline of mixed Multi-modal framework with RGB-D.

to build the publicly available RGB-D dataset for fine-grained object instance classification. The datasets are publically available online at: *https://github.com/github-songsong/Fine-grained-Pointcloud-Object-Dataset*

II. RELATED WORK

A. Fine-grained object recognition

Many studies on fine-grained object recognition tasks have been conducted for decades, mainly grouped into three categories [5]: localization methods, feature encoding methods, and transformer methods.

Localization FGVC methods: These approaches aim to acquire the discriminative partition areas via training a detection model and then classifying using this trained model. For instance, Branson et al. [19], and Wei et al. [20] proposed to superintend the learning procedure of the localization process via part annotations. However, due to the high costs and lack of availability of these annotations of the above approaches, weakly supervised learning using sole image labels has gradually received much more attention. Yang et al. [21] introduced a re-ranking method to rerank the global categorization via the region representations enhancement. However, each method requires a specially designed model to identify potential areas. Moreover, these selected sections must go through the backbone for final classification.

Feature-encoding Methods: These approaches, as one of the FGVC measures, are designed to enrich the object representation to gain better classification performance. Yu et al. enhanced the representation performance for categorization by utilizing the hierarchical bilinear pooling framework, which combines the multiple cross-layer bilinear features [22]. Zheng et al. [23] categorized the input channels into several semantic meanings and ensembled the intra-group bilinear descriptions for FGVC. However, these approaches are usually inexplicable, and the performance of these models with a single encoding attribute (*convolutional processing*) is also limited [24].

Transformer methods: In recent years, transformers have made great progress in Natural language processing [25], [26]. In the meanwhile, increasing studies started applying transformers to computer vision tasks, such as object detection [27], [28], segementation [29], [30], and object tracking [31]. In particular, nowadays sole ViT model for FGVC has become increasingly popular. For example, Dosovitskiy et al. [32] proposed the ViT model with superior performance in the image classification field. Subsequently, Swin [33], DeiT [34], and MAE [35] are introduced respectively for computer vision tasks. Based on that, He et al. [24]. extend the ViT-only model to FGVC based on the traditional FGVC RGB-only dataset and evaluate the ViT framework in the FGVC community.

Many researchers have recently utilized CNN-only or ViT-only to fulfill FGVC with the RGB-only datasets. Ullrich et al. leveraged a multi-CNN network to extract RGB and Depth images for 3D object recognition [36]. Then, in the FGVC field, the RGB and Depth image representations from CNN-only models separately are also used for the single-view FGVC [17]. For ViT and CNN, their essence is to acquire the object representation. Their performance spectrum, however, is confined under their fixed architectures. With the training dataset increasing, their accuracy can be improved but only approach the maximum of their fixed architectures. To improve the performance of the single-view FGV, we proposed the mixed multi-modality approach with CNN and ViT for fine-grained RGB-D object classification.

B. Fine-grained object datasets

Recently, fine-grained object categorization tasks have attracted substantial attention with the advancement of deep learning techniques. In recent studies on FGVC, Nilsback [7] contributed a fine-grained flower dataset with 17 different species for FGVC, followed by the fine-grained Birds dataset containing 11788 images from 200 bird species [6]. Since then, FGVC has gradually gained more attention. For example, the Standford Dogs [9] and Cars [37] datasets for FGVC were published, respectively. Fine-grained VegFru [38] consisting of vegetables and fruits, and Kuzushiji-MNIST [39], have also been introduced recently.

Considering the limitation of the RGB-only data, RGB-D images rapidly emerged in computer vision tasks due to providing additional rich information. For example, Andreas et al. [40] released the object segmentation dataset, which comprises 111 RGB-D images of stacked and occluding objects on the table. Latter, the UR Fall detection dataset, contributing to tracking human skeletal research, was created by Bogdan et al [41]. For object recognition, it designs to differ diverse objects via the object representation descriptors. It is undeniable that RGB-D images enhance the descriptiveness of objects compared to RGB-only images. Several datasets for basic-level object recognition tasks also released, such as the Wahington RGB-D object dataset with 300 instances of

household objects [42], NUY Depth V2 [43], and Restaurant object dataset [44].

However, almost all the publicly available FGVC datasets comprise RGB-only images or grayscale-only images, except for the hand-grasp dataset [16] and the RGB-D dataset with vegetables and fruits [45] which is not publicly available. To make up the gap of RGB-D data for FGVC, and inspired by google research [46], we created FGVC RGB-D datasets, which consist of 120 categories instances of shoes, including 7200 frames, and 13 car classes with 780 instance views.

III. METHOD

As shown in Fig. 2, our approach uses RGB-D data as inputs. The deep networks then process the inputs and generate the corresponding fine-grained representation that will be used to learn and recognize objects. In this section, we first introduce the process of synthetic FGVC RGB-D dataset generation, and then discuss our deep mixed multi-modality approach, which is built based on CNN-ViT networks.

A. RGB-D datasets generation for fine-grained visual categorization

One of the primary goals of this research is to generate RGB-D datasets for FGVC tasks. Towards this goal, we develop a simulation environment in Gazebo to record data, which is consisting of a table and a Kinect camera mounted on a tripod (see Fig. 3). We import the model of several cars and shoes from Google scanned objects [18] for our FGVC datasets. Google scanned objects used an object scanning system [18], equipped with a camera for detecting object shapes, an HDR camera for color extraction, and a computer projector for recognizing patterns. Besides, the scanning devices also leverage a structured light to obtain the 3D shape of the object. Finally, the scanned objects of the natural world are downscaled and packaged to the Gazebo models, which contain thumbnail pictures, obj mesh, and SDF files. In our fine-grained object RGB-D datasets, there are 7200 shoe views categorized into 120 objects and 780 car views divided into 13 targets for cars.

As shown in Fig. 3, our datasets provide extremely difficult challenge for FGVC tasks because the included objects have nearly identical properties such as geometry, textures, and colors, which is even hard for human to identify the differences. We spawn the object on top of the table, and move it in front of the camera along a rose trajectory. We then record 60 partial views of each instance.

B. Structure of multi-modality representations with ViT and CNN

In recent research on deep learning approaches for FGVC, the ViT-only framework or CNN-only structure has received a lot of attention. Considering the fixed CNN or ViT structures, they have their sole pros and cons [47]. When the training data is plentiful, deep network perform well. Several experiments revealed that the CNNs often outperformed ViT on small-scale data, while ViT can gradually outperform the CNN as

Fig. 3. Fine-grained point cloud object datasets: (*top-row*) car objects in Gazebo environment; (*lower-row*) shoe objects in Gazebo environment.

data size increases. In particular, due to differences in feature extraction module of ViT and CNN, they learn local and global representations of objects, respectively. As a result, we propose a deep mixed multi-modality approach for FGVC that takes into account ViT and CNN to capture both local and global features of the object. An overview of our model is shown in Fig. 2.

To encode the object, we first extract the RGB and Depth images from the point cloud of the object using orthographic projection method as discussed in [48]. To construct a compact deep representation for the given object, the obtained RGB and Depth images are then fed into ViT and CNN, both pre-trained on ImageNet, respectively. As shown in Fig. 2(*top-row*), the obtained representations from RGB and depth images are then concatenated to form a single deep representation for the given object. The obtained representation is then imported into a classifier for classification purposes. We compare the performance of various classifiers in the content of fine-grained classification. We discovered that the K-nearest neighbor classifier outperformed other classifiers, especially when limited training data was available, which is consistent with

our previous findings [49] (see Section IV-A). Therefore, we select KNN classifier and use Motyka distance function as to measure the similarity of objects, and set the K value to 1.

Furthermore, to further refine the performance of FGVC, we consider both RGB-D images as the input for each of the network, as shown in Fig. 2(*lower-row*). The RBG-D images of the object are fed into both *CNN or ViT* networks, and then, the obtained representation are then fused using a function such as average (AVG), maximum (MAX), and appending (APP). Eventually, the obtained representation is used for categorization purposes.

IV. Result and Discussion

We conducted several experiments to evaluate the performance of the proposed method. To select the best classifier, we first evaluate the performance of various classifiers using the car dataset. We then performed an extensive set of experiments using 10-fold cross-validation algorithm [50]. In particular, we randomly divided the dataset into ten folds, with one fold serving as test data and the remaining nine serving as training data in each iteration. This experiment is repeated ten times, so each fold is used as test data once. To measure the performance of object recognition we used instance accuracy ($\frac{\#\text{ true predictions}}{\#\text{ predictions}}$).

A. multi-classifiers

To better classify the fine-grained objects, we assess the performance of various classifiers on the fine-grained car data by performing a group of 10-fold cross-validation experiments.

As shown in Fig. 4, k-Nearest Neighbors (kNN) [51], Multi-layer Perceptron (MLP) [52], Support Vector Machine (SVM) [53], Decision Tree (DT) [54], Gaussian Process (GP) [55], Random Forest (RF) [56], and Gaussian Naive Bayes (GNB) [57] are considered in this study. Compared to other classifiers, kNN classifier outperformed others in terms of accuracy. In the subsequent experiments, we use the k-NN as the base classifier for fine-grained object recognition tasks.

B. Analysis of t-SNE

To better analyze the effect of multi-modality representations on fine-grained instance accuracy, we first conduct a t-SNE analysis with mix-network representations based on DenseNet [44] and ViTMAE [45], followed by recognition results on the fine-grained car and shoe data.

As shown in Fig. 2 (top-row), we extract the RGB, depth views of the object, and then fed the RGB view into ViTMAE and the depth view into DenseNet, respectively. Afterward, we performed feature dimension reduction for the t-SNE analysis. As shown in Fig. 5(*top-row*), it is visible that the object representations from single modality (either RGB or depth) are not descriptive enough to distinguish fine-grained objects from each other, for example, check object seven and object zero in t-SNE plots. Instead, we can easily separate the objects using the representations obtained from multi-modal (Fig. 2 *lower-row*). According to the t-SNE visualization, multi-modal representations can more effectively distinguish finer-grained car objects.

C. Offline Experiments

We evaluated the performance of the proposed approach using fine-grained cars and shoes datasets. Following the above procedure, we first constructed car object embeddings based on ViT(MAE), DenseNet, and mixed multi-model, and used k-NN classifier. For k-NN, distance function and K value play essential roles in the classification process. According to our experiments discussed in Section IV-A, the distance function and k value are set as Motyka and 1, respectively. To obtain more accurate and fair results, we use 10-fold cross-validation protocol.

In Fig. 5(*lower-row*), our approach with mixed multi-model method achieved 91.67% classification accuracy, which outperformed the other approaches. By comparing the confusion matrices, we can see that ambulance and bus were often incorrectly classified when we used either ViT or CNN, while when we considered both ViT and CNN, the performance of the agent remarkably improved. In the same vein, Fig. 5(*top-row*)shows that, in general, our approach with both CNN and ViT obtained better classification accuracy than ViT(MAE)-only and DenseNet-only models. To further verify the superiority of the mixed multi-modality approach with ViT(MAE) and DenseNet for FGVC, we performed a new round of experiments based on the fine-grained shoe object dataset.

We also ran experiments by using two-CNNs (MnasNet and DenseNet), and two-ViTs (MAE and MAE-L), mixed CNN-ViT approach (ViTMAE and DenseNet). Based on Table I, It is clear that the accuracies of the multi-model classification

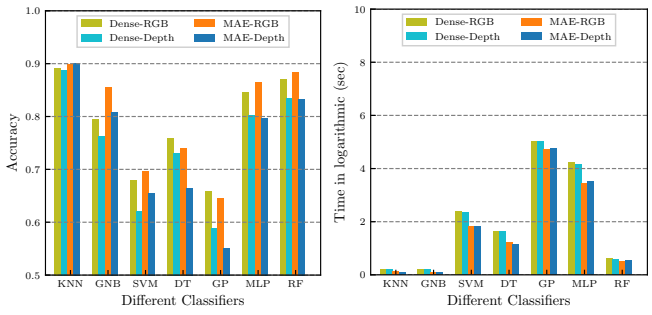

Fig. 4. Fine-grained car objects recognition performance of various classifiers: (*left*) average class accuracy; (*right*) computation time for recognition.

TABLE I
Summary of multi-modalities of single-view with RGB and Depth.

Networks \ Modalities		RGB	Depth	RGB+Depth
CNN	Mnas	0.8226	0.8099	0.8899
	Dens	0.8564	0.8365	0.9031
Dens(RGB)+Mnas(Depth)		-		**0.9042**
ViT	MAE	0.9144	0.8069	0.9242
	MAE-L	0.9121	0.8175	0.9172
VAE-L(RGB)+VAE(Depth)		-		**0.9249**
VAE(RGB)+Dens(Depth)		-		**0.9313**

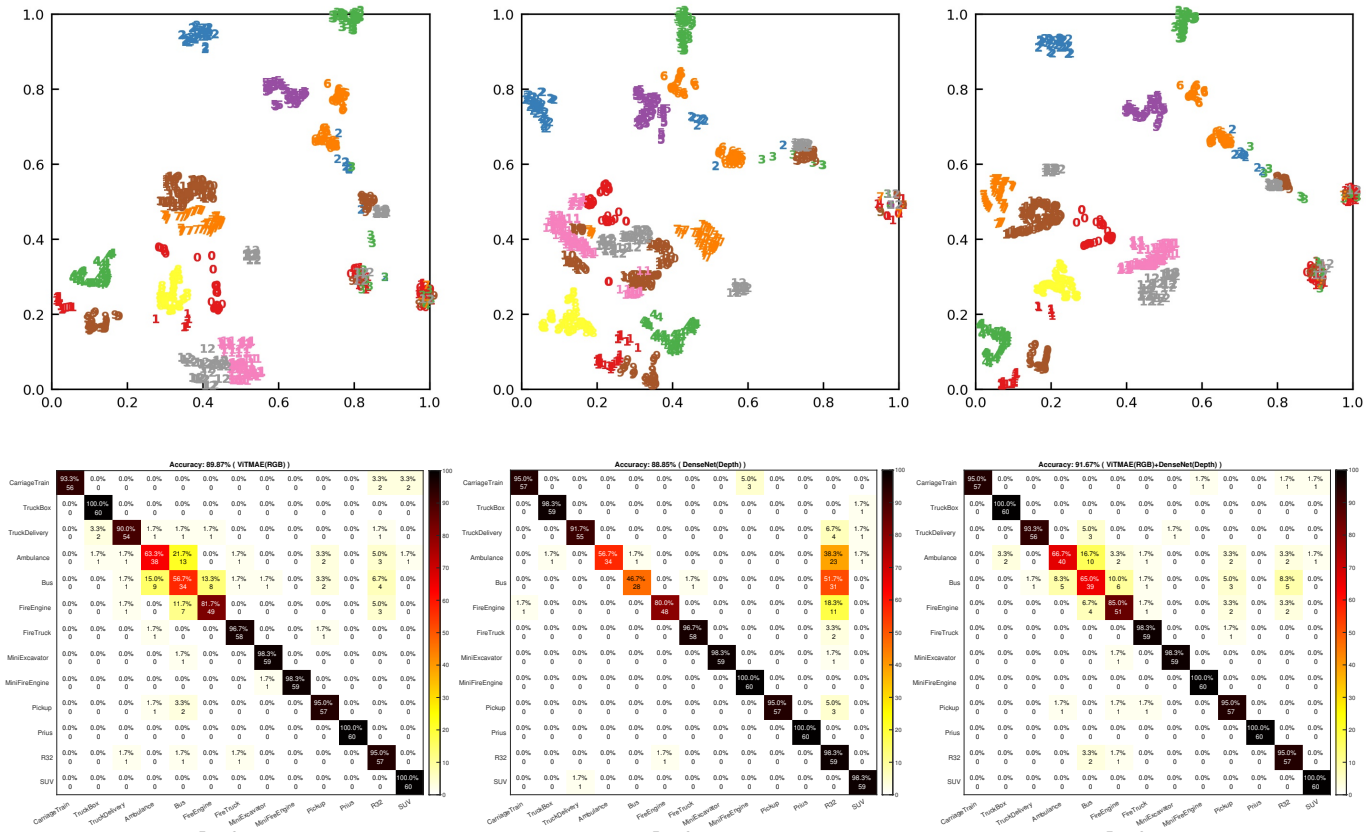

Fig. 5. (top-row): T-SNE analysis for fine-grained car objects (*left*) ViTMAE with RGB image; (*center*) DenseNet with Depth image; (*right*) mix-model with ViTMAE (RGB) and DenseNet (Depth). (lower-row) Accuracy for fine-grained car objects recognition (*left*) Accuracy of the ViTMAE with RGB image; (*center*) Accuracy of the DenseNet with Depth image; (*right*) Accuracy of mix-modal with ViTMAE(RGB) and DenseNet(Depth).

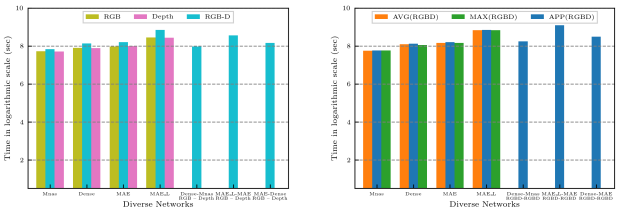

Fig. 6. Computation time of diverse networks (*left*) Computation time under RGB and Depth images; (*right*) Computation time under the AVG, MAX, and APP functions (RGB-D).

TABLE II
SUMMARY OF MULTI-MODALITIES OF SINGLE-VIEW WITH RGB-D.

Networks \ Modalities		AVG	MAX	APP
CNN	Mnas (RGB-D)	0.8757	0.8847	0.8899
	Dens (RGB-D)	0.8911	0.899	0.9031
Dens(RGB-D)+Mnas(RGB-D)		-		**0.9201**
ViT	MAE (RGB-D)	0.91	0.8186	0.9242
	MAE-L (RGB-D)	0.9042	0.6656	0.9172
VAE(RGB-D)+VAE-L(RGB-D)		-		**0.9337**
VAE(RGB-D)+Dens(RGB-D)		-		**0.9340**

outperform those of the single models. In particular, the accuracy of the mixed multi-modality with ViT(MAE) and DenseNet reached 93.13%. Besides, the mixed multi-models have a mild increase compared to the logarithm computation time of individual models, as shown in Fig. 6(*left*).

We also evaluated the performance of the multi-modal input for each of the base network as shown in Fig. 2(*low-row*). Unlike the above experiments, in this round of the experiment, all representations from single network contained the RGB-D information. For example, RGB and Depth images, from the same angle view of each object, are input into the same CNN, and ViT networks. The obtained results are summarized in Table II and Fig. 6(*right*).When the results are compared, it is clear that the multi-model approach outperforms the single model method with RGB-D images. The accuracy of the mixed multi-modality with DeseNet (RGB-D) and MAE (RGB-D) was 93.40%, which outperformed all single models, multi-CNNs, and multi-ViTs.

D. Robotic demonstrations

We performed a real-robot experiment to show the real-time performance of the proposed mixed multi-modality approach. We evaluated the proposed approach in the context of "*robot-assitant-packaging*" based on fine-grained objects. In this round of experiment, we used VAE (RGB-D) + Dens (RGB-

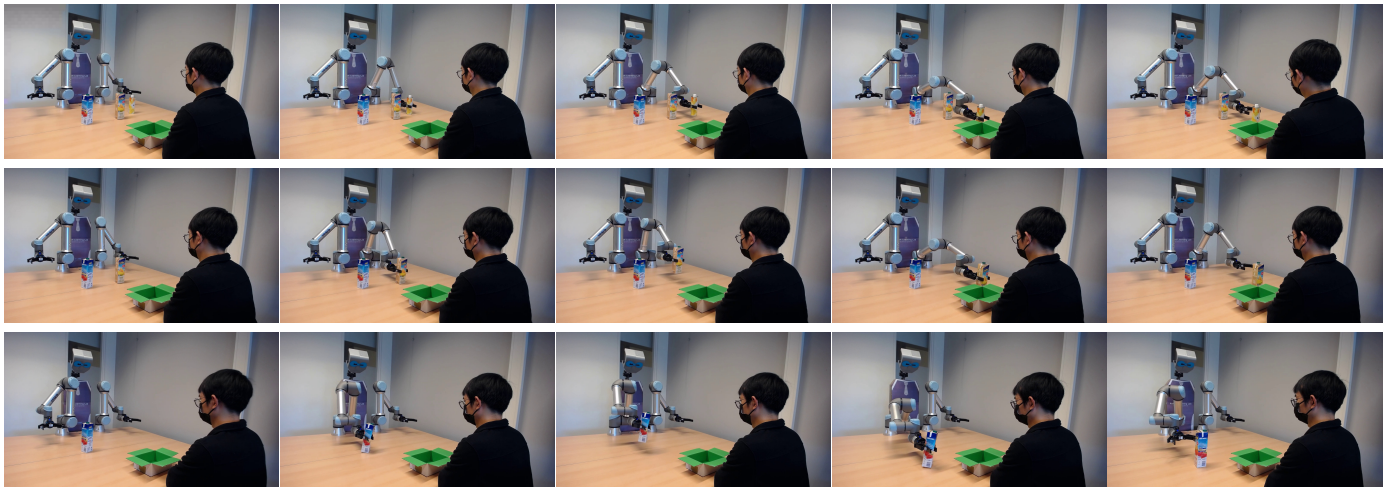

Fig. 7. A series of snapshots demonstrating the performance of our dual-arm robot in a robot-assisted packaging scenario: We used fine-grained objects (*i.e. mango_box and strawberry_box*). The objects are not reachable by human user, and therefore, the robot should handover the requested object to the user.

D) to represent objects. We randomly placed three objects including two fine-grained juice boxes, and one bottle object as shown in Fig. 7. In this setting, the robot first should recognize all objects correctly. As soon as the user asks for an specific object, the robot should grasp the object and deliver it to the user. A sequence of snapshots showing the performance of the robot during this experiment is shown in Fig. 7. In this experiment, we observed that our dual-arm robot could recognize the fine-grained objects (i.e., *strawberry_box* and *mango_box*) precisely and deliver them to the user. A video of this experiment has been attached to the paper as supplementary material.

V. Conclusion

In this paper, we presented a deep mixed multi-modality approach based on ViT-CNN networks to handle fine-grained object classification. In particular, we encode the local information of the object using ViT network and encode the global representation of the object using a CNN network through both RGB and depth views of the object. Since there is no other fine-grained RGB-D household objects datasets, we generated two synthetic fine-grained RGB-D datasets to train and evaluate our approach. Furthermore, we made the datasets publicly available to the benefits of research communities. We performed several sets of experiments to evaluate the proposed approach. To show the usefulness of the proposed approach in real-world applications, we integrated our approach in a real-robot scenario. Experimental results showed that our approach could recognize fine-grained objects from each other with the accuracy of 93.40%. In the continuation of this work, we would like to investigate the possibility of improving the performance by considering different views of the object from various perspectives.

Acknowledgment

We thank the center for Information Technology of the University of Groningen for their support and for providing access to the peregrine high performance computing cluster. Songsong Xiong is funded by the China Scholarship Council.

References

[1] L. Karlinsky, J. Shtok, Y. Tzur, and A. Tzadok, "Fine-grained recognition of thousands of object categories with single-example training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4113–4122.

[2] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "Rpc: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2019.

[3] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[4] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha, "Benchmarking representation learning for natural world image collections," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 884–12 893.

[5] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[7] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1447–1454.

[8] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[9] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, no. 1. Citeseer, 2011.

[10] L. Shao, Z. Cai, L. Liu, and K. Lu, "Performance evaluation of deep feature learning for rgb-d image/video classification," *Information Sciences*, vol. 385, pp. 266–283, 2017.

[11] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1817–1824.

[12] S. H. Kasaei, M. Oliveira, G. H. Lim, L. Seabra Lopes, and A. M. Tomé, "Interactive open-ended learning for 3d object recognition: An approach and experiments," *Journal of Intelligent & Robotic Systems*, vol. 80, no. 3-4, pp. 537–553, 2015.

[13] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for rgb-d action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1651–1664, 2015.

[14] M. M. Rahman, Y. Tan, J. Xue, and K. Lu, "Recent advances in 3d object detection in the era of deep neural networks: A survey," *IEEE Transactions on Image Processing*, vol. PP, no. 99, pp. 1–1, 2019.

[15] M. M. Rahman, Y. Tan, J. Xue, L. Shao, and K. Lu, "3d object detection: Learning 3d bounding boxes from scaled down 2d bounding boxes in rgb-d images," *Information Sciences*, vol. 476, pp. 147–158, 2019.

[16] G. Rogez, J. S. Supancic, and D. Ramanan, "Understanding everyday hands in action from rgb-d images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3889–3897.

[17] Y. Tan, K. Lu, M. M. Rahman, and J. Xue, "Rgbd-fg: A large-scale rgb-d dataset for fine-grained categorization," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[18] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," *arXiv preprint arXiv:2204.11918*, 2022.

[19] M. Klompas, R. Branson, E. C. Eichenwald, L. R. Greene, M. D. Howell, G. Lee, S. S. Magill, L. L. Maragakis, G. P. Priebe, K. Speck *et al.*, "Strategies to prevent ventilator-associated pneumonia in acute care hospitals: 2014 update," *Infection Control & Hospital Epidemiology*, vol. 35, no. S2, pp. S133–S154, 2014.

[20] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, vol. 76, pp. 704–714, 2018.

[21] S. Yang, S. Liu, C. Yang, and C. Wang, "Re-rank coarse classification with local region enhanced features for fine-grained image recognition," *arXiv preprint arXiv:2102.09875*, 2021.

[22] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 574–589.

[23] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 852–860.

[25] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[28] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and D. D. Dai J F, "Deformable transformers for end-to-end object detection," in *Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria: OpenReview. net*, 2021.

[29] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Trans2seg: Transparent object segmentation with transformer," 2021.

[30] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5463–5474.

[31] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple object tracking with transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[35] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[36] M. Ullrich, H. Ali, M. Durner, Z.-C. Márton, and R. Triebel, "Selecting cnn features for online learning of 3d objects," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5086–5091.

[37] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[38] S. Hou, Y. Feng, and Z. Wang, "Vegfru: A domain-specific dataset for fine-grained visual categorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 541–549.

[39] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[40] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 4791–4796.

[41] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[42] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1817–1824.

[43] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[44] S. H. Kasaei, A. M. Tomé, L. S. Lopes, and M. Oliveira, "Good: A global orthographic object descriptor for 3d object recognition and manipulation," *Pattern Recognition Letters*, vol. 83, pp. 312–320, 2016.

[45] Y. Tan, K. Lu, M. M. Rahman, and J. Xue, "Rgbd-fg: A large-scale rgb-d dataset for fine-grained categorization," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[46] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," *arXiv preprint arXiv:2204.11918*, 2022.

[47] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[48] S. H. Kasaei, "Orthographicnet: A deep transfer learning approach for 3-d object recognition in open-ended domains," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 2910–2921, 2020.

[49] H. Kasaei and S. Xiong, "Lifelong ensemble learning based on multiple representations for few-shot object recognition," *arXiv preprint arXiv:2205.01982*, 2022.

[50] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross validation, encyclopedia of database systems (edbs)," *Arizona State University, Springer*, vol. 6, 2009.

[51] O. Kramer, "K-nearest neighbors," in *Dimensionality reduction with unsupervised nearest neighbors*. Springer, 2013, pp. 13–23.

[52] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[53] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[54] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[55] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[56] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[57] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Updating formulae and a pairwise algorithm for computing sample variances," in *COMPSTAT 1982 5th Symposium held at Toulouse 1982*. Springer, 1982, pp. 30–41.