IMPROVING GENERALIZATION OF DEEP NETWORKS FOR ESTIMATING PHYSICAL PROPERTIES OF CONTAINERS AND FILLINGS

Hengyi Wang[1], Chaoran Zhu[2*], Ziyin Ma[3], Changjae Oh[2]*

[1] School of Computer Science, University College London
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London
[3] School of Computer Science, University of Birmingham

ABSTRACT

We present methods to estimate the physical properties of household containers and their fillings manipulated by humans. We use a lightweight, pre-trained convolutional neural network with coordinate attention as a backbone model of the pipelines to accurately locate the object of interest and estimate the physical properties in the CORSMAL Containers Manipulation (CCM) dataset. We address the filling type classification with audio data and then combine this information from audio with video modalities to address the filling level classification. For the container capacity, dimension, and mass estimation, we present a data augmentation and consistency measurement to alleviate the over-fitting issue in the CCM dataset caused by the limited number of containers. We augment the training data using an object-of-interest-based re-scaling that increases the variety of physical values of the containers. We then perform the consistency measurement to choose a model with low prediction variance in the same containers under different scenes, which ensures the generalization ability of the model. Our method improves the generalization ability of the models to estimate the property of the containers that were not previously seen in the training.

Index Terms— Multi-modal perception, generalization, data augmentation

1. INTRODUCTION

Human-robot collaboration is an important concept in the modern industry, where robots are required to interact and cooperate with humans [1, 2, 3, 4, 5, 6]. In this work, we aim at addressing tasks that are crucial in human-robot object handover: estimating the physical properties of target containers and their fillings. A good estimation model should generalize to containers with various materials, shapes, textures, and transparency. However, estimating the physical properties of the container is still challenging due to the invisible content, occlusion, and variety of household containers.

Existing methods [7, 8, 9] address the classification of filling level and type mostly with audio data, whereas the estimation of capacity, dimension, and mass of the container are remaining problems. A dataset available and commonly used to estimate these properties is the CORSMAL Containers Manipulation (CCM) dataset [10]. Training a neural network for regressing the capacity, dimension and mass value on this dataset, which with only 9 objects appears repeatedly in the total 684 videos, can be easily over-fitted and eventually make the regression similar to classification. Therefore, neural networks trained directly on the dataset, such as [9], usually lead to poor generalization. Other works [7, 8] that apply
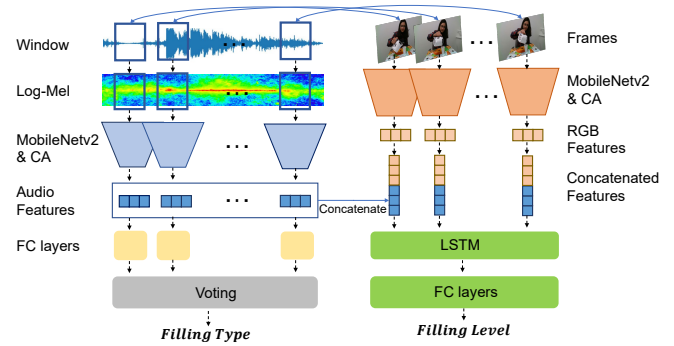

Fig. 1. Overview of our approach for the filling level and type classification of the container. Key– CA: Coordinate Attention, FC: Fully Connected, LSTM: Long Short Term Memory, Log-Mel: Log-Mel spectrogram.

non-deep-learning approaches for estimating the container capacity show certain generalization ability but have limited performance.

In this paper, we present methods to estimate the physical properties of household containers, such as filling type and level classification, container capacity, mass, and dimension estimation using the CCM dataset [10]. We use MobileNetv2 [11] with the coordinate attention (CA) [12] as a backbone model, with minor changes of input modalities for each task. Our backbone model incorporates positional information, which shows good generalization ability with both audio and RGB-Depth (RGBD) inputs. The audio data is converted to clips of log-Mel spectrogram [13] which are used as input for the filling type classification. To address the filling level classification, we jointly combine the intermediate audio feature from the filling type classification with the intermediate visual feature from the corresponding RGB frames. For the estimation of container capacity, dimensions and mass, we use RGBD images cropped by YOLO-v5 [14] to filter out the irrelevant information in the target frame. To address the generalization problem, we augment the annotations for container capacity, dimension, and mass using the geometric relationship between 2D images and 3D objects by randomly resizing the detected object of interest and changing the value of the label based on the geometric relationship. We further present a variance-based evaluation that measures the consistency of the model in each container to evaluate the generalization ability of trained models.
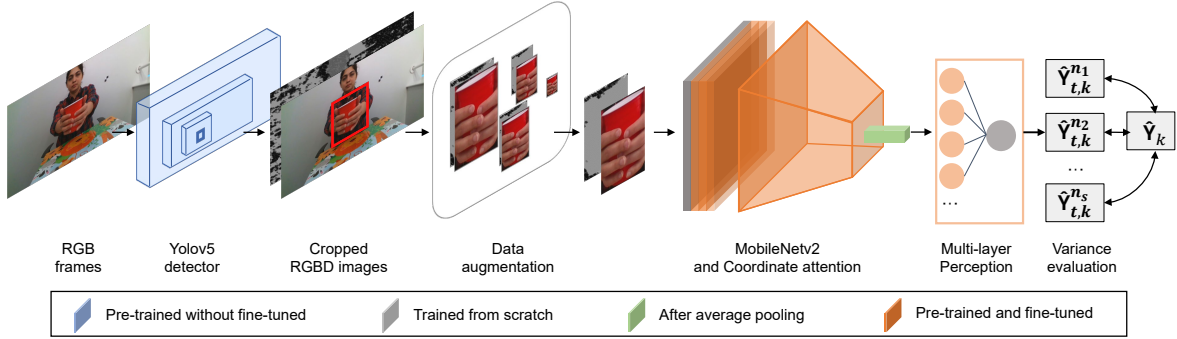
---

*Equally contributed.

Fig. 2. Overview of our framework to estimate container capacity, mass, and dimension. We use the same network architecture for these tasks and employ the data augmentation strategy to augment the target objects detected from YOLO-v5 [14]. At training time, we perform training with RGBD input and evaluate the variance of the videos with the same container k.

2. PROPOSED METHOD

We present our methods to perform filling level and type classification (Section 2.1) and container capacity, mass, and dimension estimation (Section 2.2). We use the CCM dataset [10] that provides audio-visual recordings where people manipulate containers including cups, drinking glass, and food boxes by pouring the filling into the container, shaking the food box, or performing no action. The container filling is labeled with one of four classes, $\{empty, pasta, rice, water\}$, with three different filling levels, $\{0\%, 50\%, 90\%\}$. Among the video data from four views, we use the RGB and depth images captured from a camera facing the human operator.

2.1. Filling level and type classification

Figure 1 shows the pipeline of estimating the filling type and level of the container. Inspired by [8], we also rely on the clipped audios and voting strategy for the filling type estimation and we adopt the LSTM [15] network passed by the intermediate feature from the filling type estimation to predict the filling level. However, instead of using the VGG backbone [16] on the audio features, we use a more efficient backbone, MobileNetv2 [11] with CA [12]. For the filling level classification, we employ a multi-modal approach that integrates information from audio and RGB video. We combine the intermediate audio feature from the filling type classification with the RGB feature since combining different modalities can utilize more relevant object features and improve the generalization of the model in a small training set.

Backbone model. To emphasize the information for the target tasks, we include the CA [12] to MobileNetv2 [11]. Unlike the Squeeze-and-Excitation attention [17] which neglects the positional information, the CA generates the channel attention with the spatial information, which captures the useful information from the audio feature and RGB frames by encoding the long-range positional information.

Feature extraction. An audio sequence usually contains various events, which may not be relevant to our objective. Training a model on such audio sequences may introduce noise caused by other events. Using additional existing annotation containing the starting and ending time of events in the audio signals [8], we divide the signals into a series of audio clips overlapping in time. The clips which do not contain the target events will be assigned to an empty label. These audio clips will be used for extracting the log-Mel spectrogram [13] and the training will be performed on these clips instead of the whole

audio sequence. The corresponding RGB video frames of each audio clip are also extracted to jointly estimate the filling level in a time-series manner.

Filling type classification. Let us denote an n-th audio sequence with multiple clips, where the t-th audio feature map, i.e. the log-Mel spectrogram, X_t^n, is extracted using the audio clips with the length of 0.025s and the stride of 0.01s, with an assigned label, Y_t^n. In this task, we aim to find a model f_{θ_T}, with learned parameters θ_T on the audio features, to perform the classification on the four classes of filling types. The objective function is:

$$\theta_T = \arg\min_\theta \sum_n \sum_t \mathcal{L}\left(f_\theta\left(X_t^n\right), Y_t^n\right), \quad (1)$$

where $\mathcal{L}$ is the cross-entropy loss. At the inference time, we perform classification on each X_t^n in audio. As each X_t^n generates an independent prediction, we adopt a voting strategy to decide the final predictions. If the number of non-empty predictions does not exceed the threshold, which we empirically set to 5, the sequence is labelled to *empty*. Otherwise, the label is the class with the most predictions.

Filling level classification. We adopt a multi-modal fusion strategy, which integrates information from multiple modalities. As each modality captures specific characteristics of the underlying information, combining multiple complementary modalities can provide a highly comprehensive set of data. As the filling level of the containers varies over time, we perform this task in a time-series manner. We first align the frames with the center of the generated audio clips. These audio and RGB frames are then used as input to the pre-trained audio model (f_{θ_T}) from the filling type estimation and the pre-trained vision model from ImageNet classification [18], respectively, since the CCM training data is not sufficient to train the whole pipeline from scratch. Note that the parameters for the pre-trained audio and vision models are fixed during training to avoid the model over-fit to the limited training samples. We retrieve the embeddings before the fully connected (FC) layer from audio and vision models. The intermediate features from these two models are then concatenated and fed into a LSTM network [15] with FC layers to classify the filling level. We aim to find the model g_{θ_L} with the learned parameters θ_L as follows:

$$\theta_L = \arg\min_\theta \sum_n \mathcal{L}(g_\theta(f_{\theta_T}\left(X_{1\ldots t}^n\right), Z_{1\ldots t}^n), Y^n) \quad (2)$$

where $X_{1\ldots t}^n, Z_{1\ldots t}^n$ are the audio and RGB data, respectively.

2.2. Container capacity, mass, and dimension estimation

Figure 2 shows the pipeline of estimating the capacity, mass, and dimensions. We utilize the same network structures with previous tasks, i.e. the MobileNetv2 with CA, as the backbone model. We first crop the object of interest using bounding boxes containing the target object. We then perform the data augmentation and use the augmented data as input to train the model for better generalization. The transfer learning strategy, which uses the trained parameters on the dimension estimation, is also adopted to improve the generalization ability of our model on capacity and mass estimation. After generating the predictions, we apply the proposed consistency evaluation to find the model with less variance, which may have a better generalization on those unseen containers.

Backbone model. As an RGBD image requires 4 channel inputs, we initialize the parameters for the first 3 input channels with MobileNetv2 pre-trained on ImageNet classification [18] and randomly initialize the parameters for the 4-th channel of our model. We also include the CA [12] for better generalization of the model.

Container detection. Prior to estimating the capacity, mass, and dimension, we first locate the container's position in the images extracted from videos since the background does not contribute to the predictions directly but occupies a large portion in the image. We apply a YOLO-v5 [14] object detector to localize the object in the RGB frame and crop the object of interest in both the RGB frame and corresponding depth image (already spatially aligned). As there is no annotation of bounding boxes in the CCM dataset, we use the detector pre-trained on COCO dataset [19].

Data augmentation. Even if the training set of CCM contains 684 audio-visual recordings with different manipulations and conditions, the labels for container capacity, mass and dimensions are limited to the number of containers used in the set, i.e. 9. Considering that the model estimates the continuous values, training the model to address such a regression problem on the CCM dataset may cause an over-fitting issue. Assuming that, given a fixed depth of the target container, the capacity, dimension, and mass can be proportional to the number of pixels that belong to the target object, we make use of this relationship to perform the data augmentation. Specifically, we randomly resize the target container image by a scale factor s and change their corresponding labels with such a scale factor. Considering that the depth value has not been changed and by neglecting the correction for the perspective projection of the object, we multiply s^2 to the capacity and mass and s to the dimension values according to scaling the target container image by s. Note that this augmentation is an approximation that might not be physically correct. To compensate for the noise introduced by the proposed approach, we sample s according to a normal distribution with the mean $\mu = 1$ and the standard deviation $\sigma \in [0.5, 1.0]$. The distribution is truncated to make sure $s \in [0.5, 1.5]$. The final images are padded to $[640, 640]$ and resized to $[320, 320]$ for reducing the computation cost.

Transfer learning. The container capacity and mass can be obtained if the dimensions and the materials (i.e density) are known. We thus consider that the container capacity and mass estimations are higher-level tasks than the dimension estimation as the model needs to not only detect the object shape but also learn semantic context, such as material and volume from the image. We thus employ a transfer learning that uses knowledge learned from the container dimension estimation for the container capacity and mass estimation. We first train our model on the dimension estimation and then use the pre-trained weights as initial weights to train models for the capacity and mass estimation. Figure 4 visualizes the intra-class variance in our validation dataset. We can observe that the model trained with the
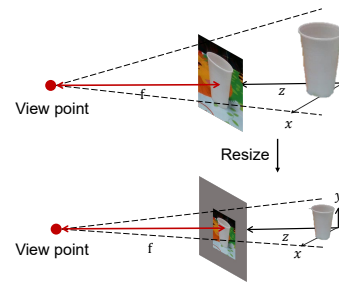

Fig. 3. Illustration of our data augmentation. The object is resized while keeping the camera matrix and distance fixed. The image is resized to a fixed aspect ratio with zero padding, the gray area surrounding the container. f is the focal length from the camera matrix, x,y,z are the world coordinates.
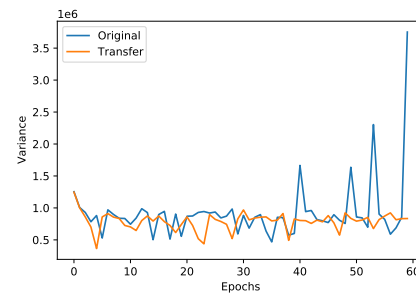

Fig. 4. Effect of the transfer learning on capacity estimation.

weights transferred from the dimension data can maintain a more stable variance curve compared to the model trained from scratch. The model can encode the representations of the target container after learning to predict the width and height, which improves the generalization ability of our model.

Consistency evaluation. Although we perform the augmentation on the training set, the predictions of the same containers should be consistent at the test time, which indicates a good generalization. Given the predictions of n-th video frame t with container k, $\hat{Y}_{t,k}^n$, and its real-value label, Y_k, we measure the total variance, Var, as:

$$Var = \sum_k \sum_n \sum_t (\hat{Y}_{t,k}^n - Y_k)^2. \tag{3}$$

Using the estimated variance of the models on our manually chosen validation set, we select the model with high accuracy with low variance to ensure the generalization ability.

3. EXPERIMENTS

3.1. Setup

We evaluate our models on the CCM dataset, which is divided into the public train (9 containers, 684 configurations), public test (3 unseen containers, 228 configurations), and private test (3 unseen containers, 228 configurations). We then split the training set into two parts: 7 objects for training and 2 objects for self-validation. For object detection, we use YOLO-v5 pre-trained on COCO dataset [19]. As classes in COCO dataset are different from those in the CCM dataset, similar to [9], we detect objects according to the categories: *cups*, *wine*, *glasses*, *vases* (for *wine glasses*), *books*, and *laptops* (for *boxes*). If object detection is failed, we use the average prediction from the training dataset. For training, we use Adam [20] optimizer with a learning rate of 0.001 for 100 epochs for each task except

Table 1. Results of the filling type and level classification and container capacity estimation. The best results are highlighted in bold. Key– A: audio, C: color image, D: depth image.

Method	Filling Type		Filling Level		Capacity	
	Input	Score	Input	Score	Input	Score
KEIO-ICS [21]	A	80.72	A	65.73	C	**72.26**
BIT [9]	A	94.26	A,C	79.65	C	60.57
HVRL [8]	A	96.95	A	78.65	C,D	54.79
Concatenation[7]	A,C	41.83	A,C	42.70	C	62.57
ACC [22]	A	94.50	A	**80.84**	-	-
NTNU [23]	A	86.89	-	-	D	67.30
Ours	A	**99.13**	A,C	77.40	C,D	59.51

Table 2. Results of the container mass and dimension, width (top and bottom) and height estimation. The best results are in bold. Key– A: audio, C: Color image, D: depth image.

	Input	Mass	Width (top)	Width (bottom)	Height
KEIO-ICS [21]	C	40.19	69.09	59.74	70.07
Visual [24]	C,D	49.64	-	-	-
Ours	C,D	**58.78**	**80.01**	**76.09**	**74.33**

the filling type classification trained for 200 epochs. We use cross-entropy loss for the filling level and type classification and $\ell 1$ loss the container capacity, mass, dimension estimation.

3.2. Implementation Details

For the filling level and type classification, we extract the log-Mel spectrogram of 64 dimensions from 8-channel audios. The fully-connected layers in the filling type classification contain two layers of size 512×128 and $128 \times N$, where N is the number of classes, and in the filling level classification the size becomes $1280 \times N$ with 0.2 dropout rate. The LSTM network contains 3 layers of 256 units each with 0.15 dropout rate. For the MobileNetv2 backbone, we use the original structure [11], except the last output layer that is modified to fit the desired output. The feature of the audio and color (RGB) image extracted by MobileNet is (1280,1) after average pooling. For the container capacity, mass and dimensions estimation, we use the depth information provided by the CCM dataset, i.e. the distance from the object to the camera. To handle the inaccurate depth information, e.g. the flattened depth of a cylindrical surface, we concatenate the depth map with the RGB channels and process them in the network in the same way.

3.3. Results

Table 1 and 2 show the results of the existing methods and ours for the filling type and level classification and container capacity, mass, and dimension estimation. The benchmark scores[1], averages of the public and private test dataset, are based on 10 performance scores that quantify the accuracy of the estimations and consider the impact of the estimations on the quality of human-to-robot handover and delivery of the container by the robot in simulator [6].

As shown in Table 1, our method outperforms other methods for the filling type classification. ACC [22] shows the best result in the filling level estimation, which uses multiple classifiers to first detect the action by the human and then the filling type. The audio data can generalize well on different containers for the filling type classification. However, in the filling level estimation, the model using audio-only may be distracted by some background noise. We examine our multi-modality strategy for the filling level estimation. Table 3 show the results of the ablation study. Combination with the vision-based

[1] http://corsmal.eecs.qmul.ac.uk/challenge.html

Table 3. Effect of input modalities. Key– A: audio, C: color image.

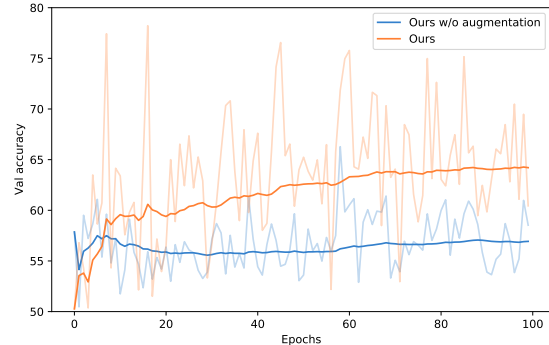| Method	Input	Filling Level
Ours	A	75.53
	A,C	77.40

Fig. 5. Effect of the proposed data augmentation on the capacity estimation. The train and validation set are manually selected with different containers using the public training set. The original curves (light color) are smoothed (dark color) for better visualization.

encoder improves the generalization of our model, which makes the total score higher than the audio-based model. Compared with KEIO-ICS [21], we obtain a higher score using the audio as input only, showing the effectiveness of our backbone model.

For the container capacity estimation, KEIO-ICS [21] shows the best result. While HVRL [8] and KEIO-ICS use a model specific to solve each task, we use the same backbone for the five estimation tasks, which can be more generalizable. In Table 2, our method outperforms others for estimating the container mass and dimensions, i.e., the width at the top and bottom, and height of the container.

3.4. Analysis

We perform ablation studies to analyze the effect of the modality fusion and the proposed data augmentation.

We examine the effect of the modality-fusion strategy in the filling level estimation in Table 3. The results show that fusing the audio and visual features improves the performance of the filling level estimation as the vision cues can provide additional information, e.g., the filling level of the transparent containers.

We analyze the effect of the proposed data augmentation in our validation set (split from the training set). As shown in Figure 5, in the container capacity estimation, the validation accuracy of the model with the data augmentation is higher as the epochs increase compared to the model without the augmentation.

4. CONCLUSION

We presented methods to estimate the physical properties of the household containers for human-robot collaboration. The key idea of our approach is to improve the generalization ability of the models using multi-modal information with attention and geometric relationship between 2D images and 3D containers. Future works include the investigation of a physically-correct data augmentation method and transfer learning on each task.

5. REFERENCES

[1] R. Sanchez-Matilla, K. Chatzilygeroudis, A. Modas, N.F. Duarte, A. Xompero, P. Frossard, A. Billard, and A. Cavallaro, "Benchmark for human-to-robot handovers of unseen containers with unknown filling," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1642–1649, 2020.

[2] J.R. Medina, F. Duvallet, M. Karnam, and A. Billard, "A human-inspired controller for fluid human-robot handovers," in *IEEE-RAS Int. Conf. Humanoid Robot.*, 2016.

[3] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan, V. Ortenzi, P. Corke, and M. Grafinger, "Object-independent human-to-robot handovers using real time robotic vision," *IEEE Robot. Autom. Lett.*, vol. 6, no. 1, pp. 17–23, 2020.

[4] V. Ortenzi, A. Cosgun, T. Pardi, W.P. Chan, E. Croft, and D. Kulić, "Object handovers: a review for robotics," *IEEE Trans. Robot.*, 2021.

[5] W. Yang, C. Paxton, A. Mousavian, Y. Chao, M. Cakmak, and D. Fox, "Reactive human-to-robot handovers of arbitrary objects," in *IEEE Int. Conf. Robot. Autom.*, 2021.

[6] Y.L. Pang, A. Xompero, C. Oh, and A. Cavallaro, "Towards safe human-to-robot handovers of unknown containers," in *IEEE Int. Conf. Robot Hum. Interact. Commun.*, 2021.

[7] Q. Liu, F. Feng, C. Lan, and R.H. Chan, "Va2mass: Towards the fluid filling mass estimation via integration of vision and audio learning," in *Int. Conf. Pattern Recognit.*, 2021.

[8] R. Ishikawa, Y. Nagao, R. Hachiuma, and H. Saito, "Audio-visual hybrid approach for filling mass estimation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021.

[9] V. Iashin, F. Palermo, G. Solak, and C. Coppola, "Top-1 corsmal challenge 2020 submission: Filling mass estimation using multi-modal observations of human-robot handovers," in *Int. Conf. Pattern Recognit. Workshop*, 2021.

[10] A. Xompero, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Corsmal containers manipulation (1.0) [dataset]," *DOI: https://doi.org/10.17636/101CORSMAL1*, 2020.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[12] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021.

[13] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Int. Symp. Music Inf. Retr.*, 2000.

[14] G. Jocher, K. Nishimura, T. Mineeva, and R. Vilariño, "Yolov5," *Code repository https://github.com/ultralytics/yolov5*, 2020.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018.

[18] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009.

[19] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C.L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014.

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

[21] T. Matsubara, S. Otsuki, Y. Wada, H. Matsuo, T. Komatsu, Y. Iioka, K. Sugiura, and H. Saito, "Shared transformer encoder with mask-based 3D model estimation for container mass estimation," in *IEEE Int. Conf. Acoustic, Speech, Sig. Proc. Grand Challenges: Audio-Visual Object Classification For Human-Robot Collaboration*, 2022.

[22] S. Donaher, A. Xompero, and A. Cavallaro, "Audio classification of the content of food containers and drinking glasses," in *Europ. Signal Proc. Conf.*, 2021.

[23] G. Christmann and J. Song, "2020 CORSMAL Challenge - Team NTNU-ERCReport," 2020, `https://corsmal.eecs.qmul.ac.uk/resources/challenge/2020.11.30_CORSMAL_NTNU-ERC_Report.pdf`.

[24] T. Apicella, G. Slavic, R. Ragusa, P. Gastaldo, and L. Marcenaro, "Container localisation and mass estimation from an RGB-D camera," in *IEEE Int. Conf. Acoustic, Speech, Sig. Proc. Grand Challenges: Audio-Visual Object Classification For Human-Robot Collaboration*, 2022.