GarbageNet: A Unified Learning Framework for Robust Garbage Classification

Jianfei Yang, Zhaoyang Zeng, Kai Wang, Han Zou, and Lihua Xie, *Fellow, IEEE*

***Abstract*—The recyclability of domestic waste plays a crucial role in the modern society, which helps reduce multiple types of pollution and brings economic effect. To achieve this goal, garbage classification is one of the most important steps during the recycling process. Prevailing deep learning techniques empower high-performance visual recognition models and can benefit the automation of garbage classification task. Nevertheless, there exist three challenges when directly leveraging deep recognition models for this task, referring to lack of sufficient data, high cost of category increment, and noisy data quality. In this article, we present a novel incremental learning framework, GarbageNet, to address the aforementioned challenges. First, weakly-supervised transfer learning guarantees the capacity of feature extractor. Second, for new categories of garbages, GarbageNet embeds them as anchors for reference and classifies the test samples by finding their nearest neighbors in the latent space. Third, an attentive mixup of training data is utilized for suppressing the negative effect of mislabeled data. We evaluate our method on real-world datasets, and the empirical results demonstrate that GarbageNet achieves the state-of-the-art performance with regard to accuracy, robustness, and extendability. The proposed method won the first place in the HUAWEI Cloud Garbage Classification Challenge in 2019.**

***Impact Statement*—This article contributes to a vision-based garbage recognition system for the recyclability of domestic waste. To the best of our knowledge, it is the first article that explores a realistic vision-based solution for the garbage recognition with a real-world dataset and a comprehensive benchmark on current state-of-the-art visual recognition models. The proposed method addresses several challenges and achieves state-of-the-art performance. We believe that this interdisciplinary research contributes to the AI for the environment, which helps promote environmental ethics, bring rotation economy, and relieve the pressure of consumption doctrine in smart city.**

***Index Terms*—Convolutional neural network, garbage classification, incremental learning, recyclability, transfer learning.**

Manuscript received January 18, 2021; revised April 1, 2021; accepted May 12, 2021. Date of publication May 18, 2021; date of current version October 13, 2021. This work was supported in part by the Nanyang Technological University under the Presidential Postdoctoral Fellowship Program. This paper was recommended for publication by Associate Editor G. Yen upon evaluation of the reviewers' comments. *(Jianfei Yang and Zhaoyang Zeng are co-first authors.) (Corresponding author: Jianfei Yang.)*

Jianfei Yang and Lihua Xie are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yang0478@e.ntu.edu.sg; elhxie@ntu.edu.sg).

Zhaoyang Zeng is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: zengzhy5@mail2.sysu.edu.cn).

Kai Wang is with the Institutes of Data Science and School of Computing, National University of Singapore, Singapore 119077 (e-mail: kai.wang960112@gmail.com).

Han Zou is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA (e-mail: enthalpyzou@gmail.com)

Color versions of one or more figures in this paper are available online http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAI.2021.3081055

I. Introduction

THE rapid development of human society comes with the cost of severe environmental pollution. One of the most common pollutants is the domestic waste (garbage). It is anticipated that the annual solid waste will reach 2.2 billion tonnes by 2025 globally, and the corresponding cost for waste management would be $375.5 billion [1]. The Environmental Protection Agency (EPA) proposes municipal solid waste recycling as an effective strategy [2]. In fact, many cities have started to explore such strategy such as Berlin and Singapore [3], which can promote environmental ethics, bring rotation economy, and relieve the pressure of consumption doctrine. To achieve the recycling purpose, garbage classification is the essential work, which categorizes the wastes to recyclable waste, kitchen waste, harmful waste, and other waste.

The current garbage classification mainly relies on bin-level sorting and later manual sorting, which is hazardous, cumbersome, and inefficient. Therefore, a garbage classification algorithm is highly demanded to guide the people to categorize the household waste. The recent progress in deep learning has empowered many useful applications in computer vision, such as object recognition and detection [4]. Driven by large-scale labeled data and powerful computer vision algorithms, garbage objects may be accurately classified for recycling by visual recognition models. Nevertheless, it is still a nontrivial task due to several difficulties, as illustrated in Fig. 1. For example, collecting and annotating sufficient data for training can be quite highly costly. The garbage images usually contain many noises, which makes it difficult for human labeling and degrades the performance of standard visual recognition model. Moreover, the number of garbage categories will become larger and larger, as new products with different appearances come into existence continuously. Extending the well-trained neural networks to new categories usually leads to tremendous computational overhead. These challenges hinder the performance of prevailing object recognition models for the garbage classification problem in the real world.

In this article, motivated by the realistic observations of garbage classification problem, we focus on three perspectives to deal with these challenges: 1) transfer learning, 2) incremental learning, and 3) noise-robust learning. First, since public garbage datasets have not been found by now, we can only rely on image search via Internet and, thus, limited data are available. For a state-of-the-art visual recognition model, as long as the feature extractor, i.e., convolutional neural network (CNN), is powerful enough, fine-tuning the model can achieve satisfactory

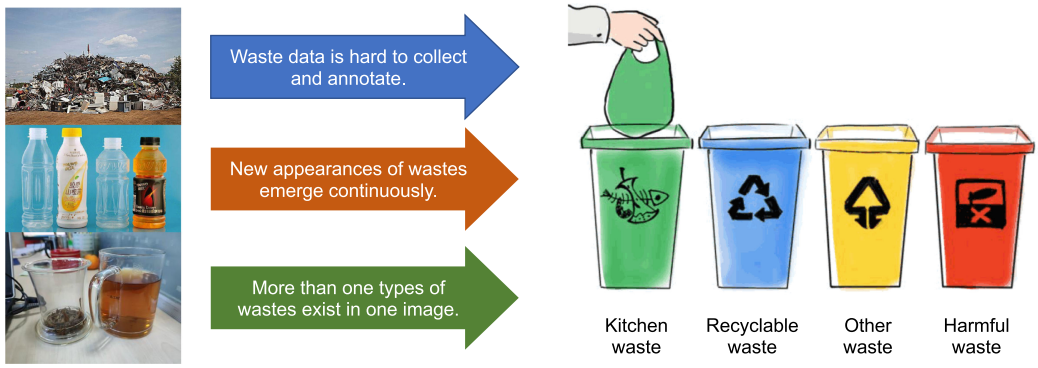

Fig. 1. Challenges of vision-based garbage classification for waste recycling.

performance [5]. Therefore, it is significant to apply transfer learning and pretrain the CNN model on large-scale dataset, such as ImageNet [6], which helps transfer knowledge from a well-annotated dataset to a realistic lack-of-data application. Second, the model needs to be flexible enough in order that new categories of garbages can be recognized by the model without much efforts. Enlarging the recognizable classes for existing models requires retraining the model with sufficient data. In our design, we borrow the idea from metric-based face recognition [7] and propose a unified learning framework for garbage recognition. Our design makes it possible to introduce new categories by only embedding new training samples and conduct the classification by measuring the embedding distances between the test sample and new training samples. Third, the annotation of garbage dataset is quite noisy because the garbage images are hard to recognize and sometimes more than one type of garbages may exist in one image. With noisy labels, deep learning models tend to have decreasing performance without particular treatment [8]. To mitigate such negative effect, we utilize a mixup method that suppresses the mislabeled data and simultaneously alleviates overfitting by inheriting the vicinal risk minimization from mixup [9].

The contributions of this article are as follows.

1) We analyze and summarize the challenges of the garbage classification problem with respect to the prevailing visual recognition methods and deep learning techniques.
2) To address these issues, we propose a novel unified learning framework, GarbageNet, which enables data-efficient learning by transferring knowledge from a rich visual domain, flexibility of introducing new categories to the model by metric learning, and noise-robust learning by data mixup for garbage classification.
3) To the best of our knowledge, this is the first paper that releases a garbage classification dataset and offers a comprehensive benchmark for a deep-model-based garbage recognition problem.
4) Extensive experiments are conducted to demonstrate the superiority of GarbageNet. Moreover, leveraging limited data for training, we show that new categories can be well recognized by GarbageNet, which avoids the cumbersome retraining process of deep neural networks for category increment.

The proposed method won the first place among 200 teams in the HUAWEI Cloud Garbage Classification Challenge in 2019, achieving 94.6% accuracy for a real-world dataset with 43 categories of common wastes. The code and dataset will be released after this article is published.

II. Related Work

A. Visual Recognition and Garbage Classification

Deep learning empowers various visual recognition tasks, of which a fundamental one is the object recognition. Since Lecun *et al.* [10] proposed to leverage CNNs for handwritten digit recognition, CNNs have been a universal backbone for visual recognition tasks. As the development of GPU boosts the computational capability, deeper and deeper CNNs begin to come into existence. AlexNet [5] is the first deep CNNs that improve the performance significantly on the challenging dataset, ImageNet [6]. Then, GoogleNet (Inception) increases the depth and width of the network while keeping the computational budget constant [11], and VGG utilizes small convolutional kernels with the same receptive field [12]. To address the gradient vanishing problem, He *et al.* [13] propose a residual connection in ResNet, which enables the network to go deeper. These deep neural networks have been further revamped using different convolution layers [14] or for more applications, such as facial expression recognition [15], gesture recognition [16], and multimodal activity recognition [17]. In this article, these models can be utilized for garbage classification by fine-tuning the model using waste images [18]. Considering the complexity, the MobileNet can achieve satisfactory performance with less computational overhead [19]. There are also other research works on detecting garbages on apps [20] and grasping garbages by robots [21]. However, the three aforementioned challenges for garbage classification have not been investigated and addressed yet, which constitutes the main contributions of this article.

B. Transfer Learning

Deep neural networks require massive labeled data that are usually not available for some specific applications, such as garbage classification. The common solution is to apply transfer

learning that trains the CNN model on public datasets and, then, fine-tunes the classifier using specific data [22]. The key step of transfer learning is the pretraining process that enables the model to learn a robust feature extractor, which can be achieved by standard supervised learning [22], weakly-supervised learning [23], and unsupervised learning [24]. Note that the pretraining should be conducted on a broader domain for the target task, in case that "negative transfer" would not happen. If there exists a domain shift during the transfer process, domain adaptation method helps align the marginal distribution among domains [25], which deals with the problem of insufficient data for visual recognition [26], [27] and ubiquitous computing [28].

C. Incremental Learning

In garbage classification, new classes of wastes will come into existence, so the model should be updated using incremental learning techniques. Given an existing well-trained model, it is desirable to learn such new capabilities without forgetting the existing knowledge [29]. Incremental learning, also called lifelong learning, drives the model to learn new capacities by adding new component models [30] or memory-based framework [31] while overcoming the catastrophic forgetting [32]. In this article, we not only focus on how to boost the model capability, but also manage to avoid retraining for saving costs. Motivated by metric learning [33], [34], the existing and new samples are projected to a feature space via a robust CNN feature extractor, and then, the classification can be easily conducted by finding the nearest neighbor of a test sample.

D. Label Noise in Deep Learning

The high performance of deep learning models is enabled by high-quality annotations, which are usually laborious and expensive in the real world. Thus, learning with label noise becomes popular especially for realistic applications [35], which consists of three categories: 1) noise-cleaning methods, 2) semisupervised methods, and 3) noise-robust methods. Noise-cleaning methods learn to identify the noisy samples by various filters [36]. Semisupervised methods rely on a manually verified clean set and assume there exists a label mapping from noisy labels to clean labels [37]. The noise-robust learning methods directly learn from noisy data, which mostly conforms with the practical scenarios. Many methods are developed to suppress the weights of noisy samples, such as curriculum learning [38] and attentive feature mixup (AFM) [9]. In this article, we leverage AFM for learning noisy garbage images.

III. Methodology

A. Problem Formulation

The garbage classification problem based on images is formulated as a K-way classification problem. In the waste standard [3], four types of wastes are defined including recyclable waste, kitchen waste, harmful waste, and other waste. However, each type of waste may include many kinds of garbages. Therefore, the normal solution is to classify them to a specific category such as "bottle" and "battery," which can be easily classified to "recyclable waste" and "harmful waste," respectively. In this article, we collect 43 classes of common garbages that belong to the four types of wastes.

Denote x as the input image and y as the class label. As illustrated in Fig. 2, the proposed GarbageNet is composed of a feature extractor $G_f(x;\theta_f)$, a normal classifier $G_y(z;\theta_y)$, a memory pool of existing categories, and a metric-based classifier. The feature extractor $G_f(x;\theta_f)$, which is a CNN parameterized by θ_f in visual recognition, maps the input image to a latent space $z \in \mathbb{R}^Z$. The normal classifier $G_y(z;\theta_y)$ is a combination of fully connected (FC) layers parameterized by θ_y. The memory pool restores the features of the training samples. In garbage classification, we consider three data streams: 1) the original labeled data $X_o = \{x_o^i, y_o^i\}_{i=1}^{N_o}$, 2) the supplementary labeled data $X_s = \{x_s^i, y_s^i\}_{i=1}^{N_s}$, and 3) the test data $X_t = \{x_t^i\}_{i=1}^{N_t}$, where N_o, N_s, N_t denote their numbers of samples.

An overview of the proposed GarbageNet is shown in Fig. 2. Using the original labeled data, we train the feature extractor G_f and the normal classifier G_y by cross-entropy loss. To suppress the noisy samples, the attentive feature mixup is utilized [9]. After the first training process, the features of the selected samples (i.e., prototypes) are saved to the memory pool $\mathcal{M} \in \mathbb{R}^{(K\times P)\times Z}$ where P denotes the number of prototypes for each category. The testing phase is conducted via a metric-based classifier that outputs the predicted category by K-nearest neighbors (KNN) in the memory pool. To reduce the computational complexity, a novel metric-based classifier is proposed with a revised cosine similarity used, so that a dot-product operation can achieve the distance measurement. In particular, when the supplementary-labeled data of new categories are ready, the GarbageNet can be simply incremented by enlarging the memory pool using the new features generated by $G_f(X_s)$. The details of these components are introduced in the following sections.

B. Transfer Learning for Feature Extractor

Transfer learning deals with the problem when insufficient training data are available [22], which enables deep learning models to learn robust features by pretraining. For the feature extractor G_f in the GarbageNet, we also apply this technique because the garbage data are hard to collect and annotate. Different from the normal pretraining strategy based on ImageNet [6], we choose to utilize the state-of-the-art pretraining parameters based on semisupervised learning and ResNeXt-101 [23]. Around 1 billion images are leveraged for weak supervision by self-training and teacher–student distillation. In Section IV, we demonstrate that such semisupervised pretraining leads to significant improvement on realistic garbage classification scenarios, even compared to deeper and wider models [39]. This highlights the importance of transfer learning parameters. Since the garbage images collected in the real world consist of various categories and usually have intrusive backgrounds, a robust feature extractor trained by billions of images helps a lot. Furthermore, the ResNeXt-101 is composed of repeated building blocks that aggregate kinds of transformations, which helps learn rich features for garbage recognition.

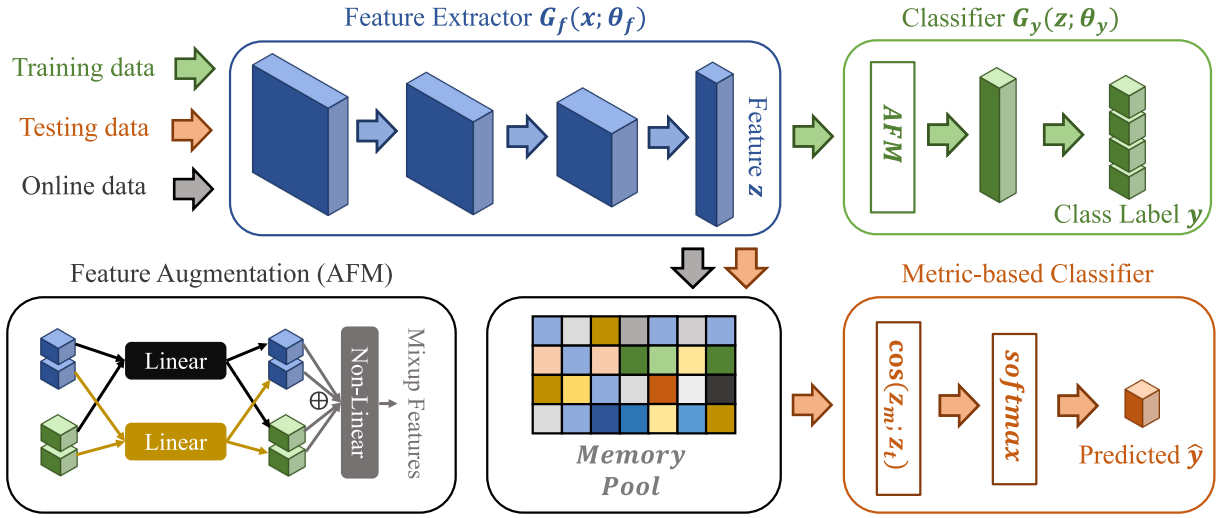

Fig. 2. Illustration of the GarbageNet. The original training data are learned by the feature extractor and a normal classifier using cross-entropy loss. Then, these features are saved in a memory pool where the incremental categories of online data are also saved after mapping the data to the features via G_f. In the testing phase, the test samples are matched to the categorical memories via cosine distance, which outputs the most similar category as the predicted result.

Using the pretrained parameters, we further fine-tune the feature extractor G_f with a normal classifier G_y via the cross-entropy loss. This process can be formulated as

$$\min_{G_f,G_y} \mathcal{L}_{\text{CE}} = -\mathbb{E}_{x\sim X_o} \sum_{n=1}^{N_o} [\mathbb{I}_{[l=y_o]} \log G_y(G_f(x_o))]. \quad (1)$$

Fine-tuning feature extractor enables the model to learn garbage-specific features and cultivates a good manifold, which motivates us to design a metric-based classifier for our method.

C. Noise-Robust Training via Attentive Mixup

Though transfer learning empowers robust feature extractor with limited garbage images, the challenges still exist in the quality of labeled data X_o. We find that more than one type of wastes may appear at one training image, which can be regarded as a type of label noise and is rather detrimental to supervised learning only based on a cross-entropy loss. To suppress the negative effect of these images, we expect to incorporate the noise-robust training into the normal training phase of the GarbageNet. To this end, we attempt to apply an attentive feature mixup [9] that is a plug-and-play module for noise-robust learning. It originally aims to deal with the incorrect annotations, but we find it effective to mitigate the negative effect of mixed categories in the garbage classification problem, which benefits the GarbageNet without involving any complexity into the model inference phase.

Let $\mathcal{B} = \{(x^1, y^1), (x^2, y^2), \ldots, (x^b, y^b)\} \in X_o$ be the mini-batch set of the original labeled data X_o with the batch size of b and the noisy one-hot label vector y^i. Applying AFM to GarbageNet consists of three steps. First, depending on the batch size and GPU memory, we use randomly composite numbers of groups with the group size K. Second, the ordered samples of each group are projected into feature embeddings by applying K FC layers. In the GarbageNet, we utilize $K = 2$ and set two FC layers F_a and F_b, which generates the features $\hat{x}^i = F_a(x^i)$ and $\hat{x}^j = F_b(x^j)$. The third step is the most important one that enforces $\hat{x}^i$ and $\hat{x}^j$ to interact, and thus, the groupwise weights are learned. We leverage a light-weight attention network A_t for group attention weights estimation. The attention network is a simple FC-FC-ReLU-Sigmoid structure [9]. Regarding the sum interaction $\oplus$, the attention weights are obtained by

$$[\alpha_k^i, \alpha_k^j] = A_t(F_a(x^i) \oplus F_b(x^j)) \quad (2)$$

where k denotes the kth group. It is the feature interaction that enables the attention weight learning by comparing the clean and noisy samples in an efficient manner.

Having obtained the attention weights, the mixup between the group members $\{x^i, x^j\}$ is formulated as

$$\tilde{x} = \frac{\alpha^i x^i + \alpha^j x^j}{\alpha^i + \alpha^j} \quad (3)$$

$$\tilde{y} = \frac{\alpha^i y^i + \alpha^j y^j}{\alpha^i + \alpha^j} \quad (4)$$

where $\tilde{x}$ and $\tilde{y}$ are the mixup feature and soft label, respectively. To enhance the robustness of the feature extractor, we externally use an auxiliary classifier G_a that helps obtain the AFM loss as follows:

$$\min_{G_f,G_a} \mathcal{L}_{afm} = -\mathbb{E}_{x\sim\tilde{x}} \sum_{n=1}^{N_m} \left[\mathbb{I}_{[l=\tilde{y}]} \log G_a(G_f(\tilde{x}))\right] \quad (5)$$

where N_m denotes the number of mixup samples.

It is noteworthy that introducing AFM to the GarbageNet does not bring external computations like other noise-robust learning methods [40]. The AFM module helps overcome the negative effect of noisy labels and the annotation of uncertain images that may consist of multiple types of garbages.

D. Memory Pool and Metric-Based Classifier

Apart from the noisy-robust module in the training procedure, the extendability of the GarbageNet is a nontrivial issue. The extendability of the neural network refers to the cost of increasing the recognizable categories in terms of time and space. Normally, it requires retraining the whole framework using both the original data and new data. For garbage recognition, there are many kinds of urban rubbish, of which the number is becoming larger and larger as new products continuously come into existence. Therefore, it is highly demanded to augment the recognition model without much effort.

To this end, the design of the GarbageNet first draws lessons from the few-shot learning framework [41]. Having obtained a robust feature extractor, we construct a memory pool as a support set by mapping the prototypical images to feature embeddings. When a new kind of waste is collected, the model can be simply augmented by enlarging the support set, which avoids the cumbersome retraining. In the testing phase, we employ cosine distance to search for the neighbors and, then, transform them to the prediction probabilities by *K*-nearest neighbors.

After the feature learning in Section III-B and III-C, we first map the training data X_o to the memory pool $\mathcal{M}$ as the support set. Specifically, we randomly choose P samples for each category and use them as the prototypes in $\mathcal{M} \in \mathbb{R}^{K\times P\times Z}$. This not only prevents the computation overhead when one category has too many training samples, but also benefits the classification by constructing a class-balanced memory pool. The memory pool can be simply enlarged by mapping new training samples X_s via the feature extractor G_f without retraining, which is convenient to use.

With the support set, a test sample $x_t \in X_t$ is predicted by estimating its similarity in $\mathcal{M}$. It can be achieved by an image retrieval technique that retrieves the most similar sample in $\mathcal{M}$ or makes the prediction by KNN. In the GarbageNet, we utilize a KNN-like method for classification as each category has multiple samples in our support set. To this end, we design the metric-based classifier by revamping the traditional KNN.To simplify the computations, we propose to use a simplified cosine distance in our method. For the feature embedding $z_t \in \mathbb{R}^{Z\times 1}$ of a test sample x_t, the cosine similarity is calculated as

$$d_{\cos}(z_t, z_m) = \cos\langle z_t, z_m\rangle = \frac{z_t \cdot z_m}{\|z_t\|_2\|z_m\|_2} \tag{6}$$

where z_m is a feature embedding of a support sample in $\mathcal{M}$ and $\|\cdot\|_2$ denotes the L_2-norm. Here, we need to calculate the cosine distances between a test sample and all the support vectors in $\mathcal{M}$. To simplify such process, we ignore the normalized part and only calculate the inner product $z_t \cdot z_m$. In this fashion, the similarity vector $D_{\text{sim}} \in \mathbb{R}^{P\times K}$ can be calculated by

$$D_{\text{sim}} = M \cdot z_t \tag{7}$$

where $M \in \mathbb{R}^{(P\times K)\times Z}$ is the support matrix with its row being a support vector. Then, we sum up the P cosine scores for K categories in D_{sim} and obtain the category-level similarity vector $S \in \mathbb{R}^K$. The predicted class is inferred by

$$\hat{y}_t = \arg\max \sigma(S) \tag{8}$$

Algorithm 1: GarbageNet Algorithm.

Data: X_o, X_s, X_t
1. Training Phase;
while $current\_epoch < total\_epoch$ **do**
Retrieve a batch of data $X_o^b \in X_o$;
$\min_{G_f,G_y,G_a} \mathcal{L}_{CE} + \mathcal{L}_{afm}$
end
Result: obtain the support matrix $M \in \mathbb{R}^{(P\times K)\times Z}$.
2. Incremental Phase;
Calculate the incremental feature set $G_f(X_s)$;
Enlarge the memory pool to $M \in \mathbb{R}^{(P\times(K+K'))\times Z}$;
Result:the incremental support matrix M
3. Testing phase;
Calculate the query feature z_t;
Obtain the similarity vector D_{sim};
Obtain the class-wise probability by $\hat{y}_t = \arg\max \sigma(S)$;
Result: the predicted label $\hat{y}_t$

where σ is the *softmax* function that transforms the cosine similarity to probability for each category. When K' new categories from X_s are imported to the support set, only the support matrix is enlarged to $M \in \mathbb{R}^{(P\times(K+K'))\times Z}$, which increases the space complexity by a little margin. For the time complexity, the proposed method does not lead to more training time since its training strategy is the same as other normal learning frameworks [13]. The testing phase calculates the similarity between the query feature and memory pool, which is achieved by an efficient dot-product calculation and, thus, avoids the FC layer as classifier in existing networks [42].

E. Overview of Garbage Recognition

In summary, the GarbageNet consists of the training phase, incremental phase and testing phase. Our method offers a solution that enables offline training, lifelong incremental learning, and online testing for the real-world scenario. We summarize the three phases in Algorithm 1. In the training phase, the feature extractor G_f is learned using the original labeled data X_o based on a noise-robust strategy with FC-based classifiers G_y and G_a. Then, the memory pool $\mathcal{M}$ is established by mapping P prototypes from X_o for K categories. Then, in the incremental phase, the supplementary labeled data X_s can be imported to the memory pool and, thus, enhance the model capability. Such incremental operations can be conducted whenever needed in practice. In the testing phase, the test sample is predicted by a novel metric-based classifier that utilizes a modified cosine similarity and a normalized classwise probability estimation, of which the advantages are more computational-efficient when compared to the traditional KNN algorithm based on L_2-distance.

The parameters of the GarbageNet are not increased by these revamps. First, the transfer learning in Section III-B only provides the pretrained weights and does not change the architecture. Second, the noise-robust module only works during the training of the network and, hence, does not increase the computation complexity of model inference. Third, the proposed metric-based classifier is achieved by a dot-product whose

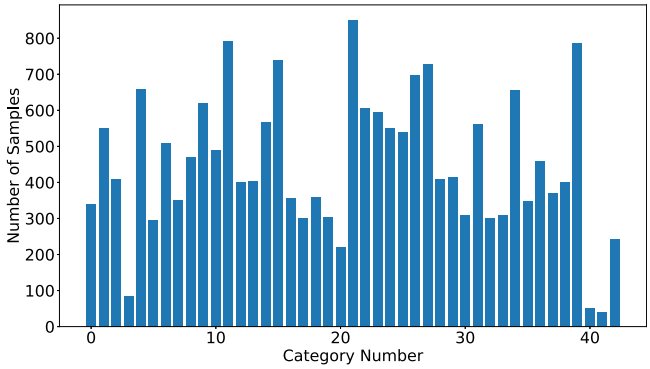

Fig. 3. Number of samples for each category in the dataset.

Fig. 4. Three challenges of the real-world garbage dataset.

computation complexity is equal to an FC-based classifier. In summary, our method overcomes the aforementioned issues without the cost of increasing model complexity for model inference.

IV. Experiment

A. Datasets

To evaluate the proposed method, we utilize the garbage classification dataset in the HUAWEI Cloud Garbage Classification Challenge. The dataset is composed of 43 categories of common wastes including: 1) disposable snack box, 2) stained plastic, 3) cigarette butt, 4) toothpicks, 5) broken pots and bowls, 6) bamboo chopsticks, 7) leftovers, 8) big bones, 9) fruit peel, 10) rotten pulp, 11) tea residue, 12) vegetable leaves, 13) egg shell, 14) fish bone, 15) power bank, 16) bag, 17) cosmetic bottle, 18) plastic toy, 19) plastic bowl, 20) plastic hanger, 21) express paper box, 22) plug, 23) old clothes, 24) can, 25) pillow, 26) plush toy, 27) shampoo bottle, 28) broken glass, 29) leather shoes, 30) cutting board, 31) Cardboard box, 32) seasoning bottle, 33) wine bottle, 34) metal food can, 35) pot, 36) edible oil drum, 37) drink bottle, 38) dry battery, 39) ointment, 40) expired medicine, 41) towel, 42) drink box, and 43) paper bag, which belong to four kinds of wastes according to the recycle standard. There are 19 459 images in the training set and similar number of images in the testing set.

The label distribution of the dataset is shown in Fig. 3. The label space is highly imbalanced because some categories of wastes are rare and hard to collect, such as toothpicks. We show some representative images in Fig. 4. In addition to the lack of training data, one kind of waste, e.g., *disposable snack box*, can have many appearances, the number of which keeps increasing as new products come into existence. This requires our model to have the ability of lifelong learning that enriches the memory pool conveniently. In the third row of Fig. 4, the garbage images may be composed of multiple classes of wastes, such as the *disposable snack box* and *leftovers* that usually exist in one image. As such, these real-world situations bring more difficulties to the garbage classification problem.

B. Implementation Details

For the implementation details of our method, we employ ResNeXt 32×16d and ResNeXt 32×8d with billions of images pretrained as the CNN backbones [23]. The training phase requires 30 epochs with the batch size of 128. The initial learning rate is 1×10^{-3} and it decays at 15th and 25th epoch by ten times. The network is optimized by stochastic gradient descent with the momentum of 0.9. The data augmentation is applied to all experiments including the random cropping and random horizontal flipping. The number of prototypes P is set to be dynamic, which is the number of samples for each category. The last epoch of the model is evaluated in the test set and the result is reported. The GarbageNet is compared to the prevailing visual recognition models: ResNet-50 [13], Dual Path Network (DPN-107) [39], ResNeXt-101 [42], Squeeze-and-Excitation Networks (SENet-154) [44], Progressive Neural Architecture Search (PNasNet-5-large) [45], AmoebaNet [46], and EfficientNet (EfficientNet-b5 and EfficientNet-b7) [47]. These models achieve the state-of-the-art performance on ImageNet for object recognition and detection [6]. The criterion is the overall accuracy among all categories. Moreover, the inference time of the model is also provided using a cloud server with one NVIDIA Tesla V100, so that the complexity of the model can be fairly compared.

C. Overall Performance

In Table I, we evaluate the proposed method and other prevailing image classification methods with respect to accuracy, parameter number, and inference time. The number of parameters of other methods is directly obtained from their original papers. First, with limited training data, we can see that the data augmentation improves the ResNet-50 baseline by 12%. Second, deeper and wider neural networks further improve the accuracy with the cost of the parameter number and inference complexity. However, for the same category of network architecture, the performance is not significantly improved when only increasing the number of layers, such as ResNet-50 (90.3%) versus ResNet-101 (90.5%). Third, the proposed GarbageNet achieves state-of-the-art performance and outperforms all other

TABLE I
OVERALL PERFORMANCE OF OUR METHOD AND OTHER PREVAILING APPROACHES. THE BOLD VALUE INDICATES THE STATE-OF-THE-ART ACCURACY.

Model	Accuracy (%)	# Params	Inference Time (ms)	Extendability
ResNet-50* [13]	78.30	25M	2.4	×
ResNet-50 [13]	90.30	25M	2.4	×
ResNet-101 [13]	90.50	44.5M	4.2	×
X-DenseNet [43]	91.25	-	4.4	×
DPN-107 [39]	93.60	83M	28.7	×
ResNeXt-101 32x8d [42]	93.10	88M	19.1	×
ResNeXt-101 32x16d [42]	93.59	193M	23.6	×
SENet-154 [44]	94.38	114M	50.4	×
PNASNet [45]	94.38	86.1M	-	×
AmoebaNet-A [46]	94.42	87M	-	×
EfficientNet-b5 [47]	94.20	30M	82.1	×
EfficientNet-b7 [47]	95.20	66M	27.8	×
GarbageNet-101 32x8d	96.48	88M	19.1	√
GarbageNet-101 32x16d	**96.96**	193M	23.6	√

Data augmentation is applied to all methods except ResNet-50*. G_f in the GarbageNet is derived from ResNeXt-101 32×8d and 32×16d.

TABLE II
ABLATION STUDY OF THE GARBAGENET ON THE VALIDATION SET

Solution	Accuracy (%)
ImageNet Pre-train	92.1
+Billion-level Pre-train	93.6
+Noise-robust AFM	94.1
+Metric-based Classifier	95.0

methods, achieving the best accuracy of 96.96% and a satisfactory inference time 23.6 ms. Even the simpler version, i.e., GarbageNet-101 32×8d, can surpass the other approaches in terms of classification accuracy.

D. Ablation Study

It is first seen that the metric-based classifier and AFM improve the original ResNeXt-101 32×8d and 32×16d by 3%–4%, as shown in Table I. To further demonstrate the effectiveness of each novel module in the GarbageNet, we build a balanced validation dataset that consists of 2000 images and conduct the ablation study. As shown in Table II, the contribution of each module differs. Using ResNeXt-101 32×8d as the baseline, the ImageNet pretrained model yields 92.1% accuracy while the weakly-supervised billion-level pretrained model improves it by 1.5% without involving any computation complexity. Then, the plug-and-play AFM helps overcome the noisy annotations and achieves 94.1% accuracy. This shows that such noises exist but are not severe in the validation data split. More importantly, the metric-based classifier leads to the most improvement, achieving an accuracy of 95.0%. The memory design not only simplifies the complexity of the incremental phase for our method, but also improves the classification performance.

E. Evaluation of Incremental Learning

The GarbageNet is trained on the original data X_o, and then, it can be extended to be able to classify more categories by supplementary data X_s. In this experiment, we use one part of the classes as X_o and the other as X_s, whereas the test data X_t keeps the same. As shown in Fig. 5, the category number of the training data (i.e., X-axis) indicates X_o. Considering that augmenting garbage classification should cost least computations, we compare our method with two classic but efficient incremental learning methods: 1) LwF [32] and 2) iCaRL [48], though they require retraining. The best performance is obtained by the setting of fine-tuning the model using complete X_o, achieving 96.96%. For the GarbageNet, the accuracy increases as the training data X_o becomes more. Nevertheless, we can observe that the model can achieve 91.7% accuracy using 20 categories as training and extra 23 categories as support set, which is still satisfactory. Furthermore, when we employ more categories for training, the performances of our method are 94.21% and 96.02% with the 30 and 40 categories of training data, respectively. In comparison, LwF and iCaRL can achieve more than 80% accuracy for all settings, but they require retraining using the new training data, which still suffers from catastrophic forgetting [32]. This demonstrates that the GarbageNet still yields good performance in the incremental learning scenario, which makes it more convenient to use in the real-world scenario.

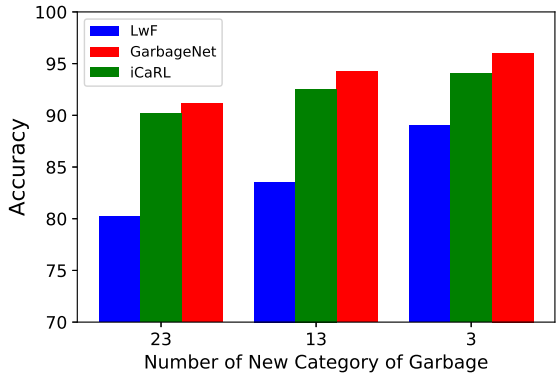

Fig. 5. Performance of the GarbageNet with different category number of images as training dataset X_o.

V. Conclusion

In this article, we study the garbage classification problem by analyzing the challenges, collecting the data, and designing a novel method, GarbageNet. The challenges include the lack of data, noisy annotations, and the requirement of incremental learning scenario. To address these problems, the GarbageNet leans on the state-of-the-art transfer learning techniques, and learns noise-robust features by a feature mixup module. Furthermore, a memory pool and a metric-based classifier are developed to enhance the capability of the model without retraining it. The GarbageNet is evaluated using real-world garbage data. It achieves the best performance of 96.96% with acceptable inference speed, surpassing all other prevailing methods. Moreover, it is observed that the GarbageNet trained on partial categories of data can still yield satisfactory performance in the incremental setting. The future work will be focused on the object detection so that various types of wastes in one image can be recognized simultaneously, which benefits the automation of waste sorting and recycle process.

References

[1] D. Hoornweg and P. Bhada-Tata, *What a Waste: A Global Review of Solid Waste Management*. Washington, D.C., USA: World Bank, 2012.

[2] R. Sanderson, "Environmental protection agency office of federal activities' guidance on incorporating EPA's pollution prevention strategy into the environmental review process," *Environ. Protection Agency*, Washington, D.C., USA, 1993.

[3] D. Zhang, T. S. Keat, and R. M. Gersberg, "A comparison of municipal solid waste management in Berlin and Singapore," *Waste Manage.*, vol. 30, no. 5, pp. 921–933, 2010.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015, Art. no. 436.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[8] Y. Xu, P. Cao, Y. Kong, and Y. Wang, "l_{DMI}: A novel information-theoretic loss function for training deep nets robust to label noise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6225–6236.

[9] X. Peng, K. Wang, Z. Zeng, Q. Li, J. Yang, and Y. Qiao, "Suppressing mislabeled data via grouping and self-attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 786–802.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. Learn. Repres.*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[14] C. Wang, J. Yang, L. Xie, and J. Yuan, "Kervolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 31–40.

[15] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, Jan. 2020.

[16] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Learning gestures from WiFi: A Siamese recurrent convolutional architecture," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10763–10772, Dec. 2019.

[17] H. Zou, J. Yang, H. P. Das, H. Liu, Y. Zhou, and C. J. Spanos, "WiFi and vision multimodal learning for accurate and robust device-free human activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 426–433.

[18] U. Ozkaya and L. Seyfi, "Fine-tuning models comparisons on garbage classification for recyclability," 2019, *arXiv:1908.04393*.

[19] S. L. Rabano, M. K. Cabatuan, E. Sybingco, E. P. Dadios, and E. J. Calilung, "Common garbage classification using MobileNet," in *Proc. IEEE 10th Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ., Manage.*, 2018, pp. 1–4.

[20] G. Mittal, K. B. Yagnik, M. Garg, and N. C. Krishnan, "SpotGarbage: Smartphone app to detect garbage using deep learning," in *Proc. 2016 ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 940–945.

[21] C. Zhihong, Z. Hebin, W. Yanbo, L. Binyan, and L. Yu, "A vision-based robotic grasping system using deep learning for garbage sorting," in *Proc. 36th IEEE Chin. Control Conf.*, 2017, pp. 11223–11226.

[22] S. J. Pan *et al.*, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[23] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019, *arXiv:1905.00546*.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[25] H. Zou, Y. Zhou, J. Yang, H. Liu, H. P. Das, and C. J. Spanos, "Consensus adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5997–6004.

[26] J. Yang, H. Zou, Y. Zhou, Z. Zeng, and L. Xie, "Mind the discriminability: Asymmetric adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, Cham, 2020, pp. 589–606.

[27] J. Yang, H. Zou, Y. Zhou, and L. Xie, "Towards stable and comprehensive domain alignment: Max-margin domain-adversarial training," 2020, *arXiv:2003.13249*.

[28] J. Yang, H. Zou, S. Cao, Z. Chen, and L. Xie, "MobileDA: Towards edge domain adaptation," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6909–6918, Aug. 2020.

[29] A. Rosenfeld and J. K. Tsotsos, "Incremental learning through deep adaptation," in *IEEE Trans. pattern anal. mach. intell.*, vol. 42, no. 3, pp. 651–663, 2018.

[30] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 177–186.

[31] E. Belouadah and A. Popescu, "IL2M: Class incremental learning with dual memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 583–592.

[32] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Nov. 2017.

[33] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[34] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 281–292, Jan. 2018.

[35] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[36] A. L. Miranda, L. P. F. Garcia, A. C. Carvalho, and A. C. Lorena, "Use of classification algorithms in noise detection and elimination," in *Proc. Int. Conf. Hybrid Artif. Intell. Syst.*, 2009, pp. 417–424.

[37] K.-H. Lee, X. He, L. Zhang, and L. Yang, "CleanNet: Transfer learning for scalable image classifier training with label noise," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5447–5456.

[38] S. Guo *et al.*, "CurriculumNet: Weakly supervised learning from large-scale web images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–150.

[39] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[40] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5051–5059.

[41] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[42] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[43] S. Meng, N. Zhang, and Y. Ren, "X-DenseNet: Deep learning for garbage classification based on visual images," *J. Phys.: Conf. Series*, vol. 1575, no. 1, 2020, Art. no. 012139.
[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[45] C. Liu *et al.*, "Progressive neural architecture search," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–34.
[46] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4780–4789.
[47] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 6105–6114.
[48] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2001–2010.

Jianfei Yang received the B.Eng. degree in software engineering from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2016, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2021.

His research interests include deep transfer learning with applications in Internet of Things and computer vision.

Dr. Yang has won many AI and data challenges in the visual and interdisciplinary fields, such as ACM ICMI EmotiW-18, and IEEE CVPR-19 UG2+ Challenge. He is currently a Presidential Postdoctoral Research Fellow and an independent PI with NTU.

Zhaoyang Zeng received the B.Eng. degree in software engineering in 2016 from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, where he is currently working toward the Ph.D. degree in computer science with the School of Computer Science and Engineering.

His research interests include object detection and vision-language cross-modality understanding.

Mr. Zheng has won many AI challenges in the computer vision fields, such as VQA Challenge and ActivityNet Challenge.

Kai Wang received the M.Eng. degree in computer science from the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently working toward the Ph.D. degree in computer science with the National University of Singapore, Singapore.

He did research internships in HUAWEI Cloud and Alibaba DAMO Academy from 2020 to 2021. His research interests include deep learning with applications in face analysis, affective computing, and resource-efficient AI.

Mr. Wang is the champion of ACM ICMI Group Emotion Recognition Challenge 2017 and ACM ICMI Engagement Regression Challenge 2018. He is a Reviewer for the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Han Zou received the B.Eng. (first class hons.) and Ph.D. degrees in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012 and 2016, respectively.

He is currently a Postdoctoral Scholar with the Department of Electrical Engineering and Computer Sciences, The University of California, Berkeley, Berkeley, CA, USA. His research interests include ubiquitous computing, statistical learning, signal processing, and data analytics with applications in occupancy sensing, indoor localization, smart buildings, and Internet of Things.

Lihua Xie (Fellow, IEEE) received the B.E. and M.E. degrees in electrical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from The University of Newcastle, Callaghan, NSW, Australia, in 1992.

Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a Professor and was the Head of the Division of Control and Instrumentation from July 2011 to June 2014. He held teaching appointments with the Department of Automatic Control, Nanjing University of Science and Technology, from 1986 to 1989, and Changjiang Visiting Professorship with the South China University of Technology, from 2006 to 2011. His research interests include robust control and estimation, networked control systems, multiagent control, and unmanned systems.

Dr. Xie was an Editor of IET Book Series in Control and an Associate Editor of a number of journals including the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, the IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-II. He is a fellow of Academy of Engineering Singapore, IFAC, and Chinese Automation Association.