MEC-Enabled Hierarchical Emotion Recognition and Perturbation-Aware Defense in Smart Cities

Yi Zhao, *Graduate Student Member, IEEE*, Ke Xu, *Senior Member, IEEE*, Haiyang Wang, *Member, IEEE*, Bo Li, *Member, IEEE*, Meina Qiao, and Haobin Shi

***Abstract*—With the explosive growth of Internet of Things (IoT) devices and various emerging network technologies, IoT-enabled smart cities are further refined into health smart cities. For example, IoT devices can automatically recognize emotional states through collected facial expressions, which can further serve mental health assessment, human–computer interaction, etc. On the other hand, existing facial expression recognition algorithms emphasize the application of deep neural networks (DNNs), and it is difficult for resource-constrained IoT devices to provide sufficient computing resources to optimize parameters for DNN-based structures. To solve the challenge of resource constraints, we propose the hierarchical emotion recognition system enabled by mobile edge computing (MEC). Specifically, MEC nodes provide IoT devices with short-delay and high-performance computing services, satisfying the requirements of training DNN-based algorithms. Moreover, our proposed emotion recognition system leverages a pretrained feature extraction module on the remote cloud to accelerate optimization and provides a localization module for specific tasks of IoT devices. In addition to evaluating the accuracy and efficiency, we also clarify that the DNN-based emotion recognition system exposes obvious vulnerability to perturbation. Due to the uncertainty of the environment, it is common for facial expressions collected by IoT devices to be accompanied by perturbation. To address this issue, we propose the proactive perturbation-aware defense mechanism. It has been demonstrated that the newly proposed defense mechanism can maintain state-of-the-art performance on the publicly available LIRIS-CSE dataset while defending against known and unknown perturbation. This can promote the deployment of our proposed MEC-enabled hierarchical emotion recognition system and defense mechanism in real-world scenarios.**

***Index Terms*—Emotion recognition, facial expression, Internet of Things (IoT), mobile edge computing (MEC), proactive perturbation-aware defense, robustness.**

Manuscript received August 31, 2020; revised April 4, 2021; accepted April 28, 2021. Date of publication May 11, 2021; date of current version November 19, 2021. This work was supported in part by the China National Funds for Distinguished Young Scientists under Grant 61825204; in part by NSFC Project under Grant 61932016; in part by Beijing Outstanding Young Scientist Program under Grant BJJWZYJH01201910003011; in part by the National Key R&D Program of China under Grant 2018YFB0803405; in part by Beijing National Research Center for Information Science and Technology (BNRist) under Grant BNR2019RC01011; and in part by PCL Future Greater-Bay Area Network Facilities for Largescale Experiments and Applications under Grant LZC0019. *(Corresponding author: Ke Xu.)*

Yi Zhao is with the Department of Computer Science and Technology and the Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China (e-mail: zhaoyi16@mails.tsinghua.edu.cn).

Ke Xu is with the Department of Computer Science and Technology and the Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China, and also with Peng Cheng Laboratory (PCL), Shenzhen 518066, China (e-mail: xuke@tsinghua.edu.cn).

Haiyang Wang is with the Department of Computer Science, University of Minnesota at Duluth, Duluth, MN 55812 USA (e-mail: haiyang@d.umn.edu).

Bo Li is with the Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: lbo@illinois.edu).

Meina Qiao is with the Department of Computer Vision Technology, Baidu Inc., Beijing 100871, China (e-mail: qiaomeina@baidu.com).

Haobin Shi is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: shihaobin@nwpu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2021.3079304

I. Introduction

DUE to the explosive growth of Internet of Things (IoT) devices (e.g., wearable sensors), the rapid development of various network technologies (e.g., 5G [1], edge computing [2], and cloud computing [3]) and the widespread application of artificial intelligence (e.g., deep learning [4], federated learning [2], and transfer learning [5]), IoT-enabled smart cities have been integrated into all aspects of people's lives [6], [7]. For example, wearable robotics can have cognitive abilities [8], thereby enhancing the perception of autism. As a nonnegligible module of health smart cities, automatic emotion recognition [9]–[11] can enable IoT devices to provide better healthcare services. For example, via perceiving the emotional states, if the driver is detected to be nervous or tired, an alarm can be issued to assist the driver in cultivating healthy and safe driving behaviors.

Regarding automatic emotion recognition, the existing methods are mainly based on visual [12], audio [13], or text [14] information. Compared with audio and text collection equipments, various visual sensors can be widely deployed indoors and outdoors. While the coverage of these devices is wide, the visual information collected is also a more effective clue for emotion recognition. In other words, visual-based emotion recognition is more suitable for application in IoT scenarios. Therefore, in this article, we focus on IoT-enabled emotion recognition system via facial expressions.

To decode the emotional state from facial expressions, the existing methods [15], [16] with excellent performance emphasize the application of deep neural networks (DNNs). For instance, Li *et al.* [17] integrated the convolutional neural network (CNN) with the attention mechanism to achieve an end-to-end learning framework. Due to the powerful characterization ability of huge parameters and the ability to perceive the occlusion regions, it can not only accurately decode facial expressions but also effectively solve the challenge caused by partially occluded faces. In various IoT

scenarios, however, resource-constrained devices account for most of the proportion. Although these resource-constrained IoT devices can run well-trained DNN-based models, they lack the required computing resources, as well as sufficient data support in model training and parameter optimization. This brings a challenging issue for the deployment of IoT-enabled emotion recognition system in real-world scenarios.

Another challenge for the IoT-enabled emotion recognition system is that visual information will frequently be slightly perturbed. First, even for visual information-oriented collection devices in the same scenario, the facial expressions they collect will have differences in image quality due to various hardware differences. Moreover, the environment of IoT devices is complex and changeable, so slight perturbations are inevitable. Due to the lack of interpretability in the current DNN-based architectures, it has been demonstrated in many fields [18], [19] that DNN-based models are highly sensitive to input noise (i.e., perturbed input), exposing obvious vulnerabilities. The perturbation encountered by IoT devices can also greatly reduce the performance of the remarkably well-trained emotion recognition system, which can be found in Section III-C. This requires that the automatic emotion recognition system deployed in IoT scenarios should emphasize robustness at the beginning of its design.

To solve the above challenges, we, for the first time, propose and implement the hierarchical emotion recognition system enabled by mobile edge computing (MEC), supporting resource-constrained IoT devices to achieve efficient emotion recognition capabilities, as well as the proactive perturbation-aware defense mechanism. More specifically, the edge network is close to the IoT devices. Coupled with the support of advanced wireless network technologies, the transmission delay between the IoT devices and the edge network is much lower than the transmission delay with the remote cloud. Moreover, the computing resources possessed by edge nodes are far superior to those of IoT devices. Since MEC [5], [20] can provide short-delay and high-performance computing services for IoT devices, we deploy model training and optimization tasks on the edge network, while IoT devices only perform real-time emotion recognition. In addition, edge nodes exploit various pretrained models on the remote cloud as the feature extraction module. In conjunction with the scenario-oriented (i.e., task-oriented) localization module, only a small amount of facial expressions are required to complete the task-oriented model fine-tuning. In fact, in addition to the combination of the mobile edge network and remote cloud, distributed deep learning [21] and collaborative learning [2] can also solve the constraints of computing, data, and other resources. Both of them are technical extensions of DNNs, and the framework proposed in this article is compatible with them. While the accuracy and efficiency are significantly improved, it also reveals the vulnerability of robustness to perturbation. To cope with possible perturbations, we further propose the proactive perturbation-aware defense mechanism. Experimental evaluation demonstrates that our proposed defense mechanism can significantly improve the performance of the IoT-enabled emotion recognition system, defending against the negative effects of both known and unknown perturbations. This can facilitate the deployment of the automatic emotion recognition system in real-world scenarios.

In summary, the key contributions we have made in this article are as follows.

1) To enable resource-constrained IoT devices with emotion recognition capabilities, we propose the MEC-enabled hierarchical emotion recognition system, achieving state-of-the-art performance in resource-constrained scenarios.
2) Since the facial expressions collected by IoT devices are inevitably accompanied by slight perturbations, we clarify the vulnerability of the IoT-enabled emotion recognition system to possible perturbations.
3) To enhance the robustness against various perturbations, we, for the first time, propose the proactive perturbation-aware defense mechanism, which can defend against both known and unknown perturbations.
4) Extensive experimental results demonstrate that our proposed hierarchical framework, in conjunction with the proactive defense mechanism, is compatible with feature extraction modules of different architectures, while achieving superior performance.

The remainder of this article is organized as follows. Section II reviews the related studies. The newly proposed MEC-enabled hierarchical emotion recognition system for resource-constrained IoT devices will be introduced in Section III. To cope with the robustness issues, we propose the proactive perturbation-aware defense mechanism for the emotion recognition system in Section IV. Section V describes the performance evaluation and analysis on the real-world dataset. Finally, Section VI concludes this article.

II. Related Work

With the rapid development of various IoT-related technologies, health smart cities [11], [22], [23] enable people to live healthier and happier. Since emotion recognition is one of the basic modules that assists people in a healthy life, many studies [10], [24] have focused on how to construct an effective automatic emotion recognition system. For example, Tariq *et al.* [25] proposed the CNN-based speech emotion detection system for elder adults in nursing homes, which can accurately decode the voice collected by audio IoT devices. In addition to audio information, visual information is also a valid clue. Since facial expressions may not match the predefined discrete emotions, Vemulapalli and Agarwala [26] utilized a compact expression embedding to describe facial expressions, formalizing them into continuous rather than discrete fashion.

There are also approaches [27]–[29] to integrate multiple types of information. Ma *et al.* [27] implemented a deep belief network to fuse audio and visual features for emotion recognition. To optimize the facial expression analysis in the industrial IoT scenario, Xi *et al.* [30] constructed a parallel neural network to integrate text features with original images. These fused features can improve the model's robustness in situations where the extracted facial features are inaccurate.

The above automatic emotion recognition methods are all based on DNNs. Although they can achieve better performance, the required computing resources and data

volume are challenging issues for resource-constrained IoT devices. To solve the challenge caused by the lack of sufficient labels, Chen and Hao [9] innovatively proposed the hybrid automatic labeling strategy. It can automatically supplement labeled data without human intervention.

While the IoT devices are subject to resource constraints, the input data accompanied by noise also pose challenges for DNN-based models. Many studies [18], [19] have demonstrated that DNN-based models are vulnerable to perturbed inputs. To facilitate the analysis of the vulnerability of DNN-based models to perturbations in security-sensitive domains, Ling *et al.* [31] for the first time implemented a uniform platform to evaluate the impact of various attack and defense mechanisms. Regarding improving the robustness of decoding partially occluded facial expression, multiple weighed facial regions of interest with different representations are adopted by Li *et al.* [17], forcing the DNN-based model to focus on non-occluded regions instead of occlusions.

In addition to the perturbation caused by natural factors such as the environment, the edge-driven intelligent architecture also puts forward higher requirements for the robustness of the MEC-enabled hierarchical emotion recognition system [2], [32]. Regarding edge-driven IoT systems, Dai *et al.* [32] discussed possible malicious attacks. For example, malicious nodes can provide false data [32] or backdoored model [2] to other edge nodes, which can further force DNN-based models lose its original performance, and even bury potential safety hazards. In response to this problem, Dai *et al.* [32] innovatively proposed an edge-driven security framework to improve robustness. Zhao *et al.* [2] proposed a stability-based defense mechanism, which can not only defend against potential backdoor attacks but also enhance the robustness of edge-driven intelligent services.

Based on the review of the referred studies, it can be found that resource constraints and robustness challenges need to be solved in the IoT scenario. Different from the existing methods, in this article, we propose the MEC-enabled hierarchical emotion recognition system for resource-constrained IoT devices. It can solve the issues of high-performance computing and data volume by deploying different modules of emotion recognition hierarchically in suitable locations. Moreover, the actively integrated perturbation-aware defense mechanism allows the IoT-enabled emotion recognition system to deal with the negative effects of various perturbations.

III. MEC-Enabled Hierarchical Emotion Recognition System for IoT Devices

In this section, we first introduce the proposed hierarchical deep learning framework, which enables more resource-constrained IoT devices to achieve efficient emotion recognition. Subsequently, we formally define the emotion recognition system and the implementation details of our proposed network architecture. Finally, we conduct performance analysis of the proposed framework and clarify the robustness issues that may be encountered in the deployment of the emotion recognition system in real-world scenarios.

TABLE I
MAJOR NOTATION LIST

Symbol	Description
x_k, x'_k	Benign and perturbed facial expression with index k
y_k, $\hat{y}_k$	Real and predicted labels associated with x_k
θ	All parameters of the deep neural network
η	Perturbation added to the benign facial expression
ϵ	Coefficient factor describing the level of perturbation
β_0	Coefficient factor achieving trade-off between benign and perturbed facial expressions
$\mathcal{S}$	Collection of selected perturbation generation methods
s	A specific perturbation generation method

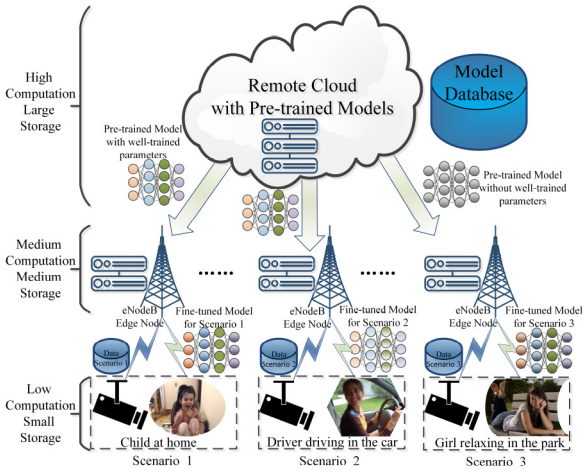


Fig. 1. MEC-enabled hierarchical deep learning framework for emotion recognition with IoT devices.

For the sake of clarity, Table I lists the major notions in our proposed emotion recognition framework. It is worth noting that the specific meaning of the symbol also depends on its superscript and subscript.

A. MEC-Enabled Hierarchical Deep Learning

To enable more resource-constrained IoT devices (e.g., computation-constrained cameras, energy-constrained GoPro, and storage-constrained robots) to achieve emotion recognition via facial expression as clue, we propose the MEC-enabled hierarchical deep learning framework for emotion recognition with IoT devices, illustrated in Fig. 1.

In our proposed hierarchical architecture, it is mainly composed of three modules: 1) IoT devices; 2) MEC networks; and 3) remote cloud networks. They cooperate with each other to support efficient emotion recognition in different scenarios. Specifically, with the explosive development of smart cities and smart homes, various IoT devices that can collect facial expressions emerge in endlessly and are deployed in many indoor or outdoor scenarios. Thus, IoT devices distributed in different scenarios are the first real-time performers of emotion recognition. Many studies have demonstrated that convolutional networks can effectively extract features of images (i.e., facial expression data). Regarding the convolutional network architecture with superior performance, its massive parameters require large-scale data and high-performance computing resources to learn the optimal configuration. However, most

of IoT devices are resource constrained and cannot satisfy the training requirements of effective convolutional networks. To cope with this issue, in this article, we deploy the parameter learning task for emotion recognition in the MEC networks (i.e., the 2nd layer), and the IoT device (i.e., the 1st layer) only utilizes the well-trained DNN-based models to perform real-time emotion recognition. This is because emerging mobile and wireless technologies have greatly reduced the transmission delay between IoT devices and MEC networks. Moreover, unlike the remote cloud, the edge network is adjacent to IoT devices, and there is almost no time delay for uploading the required data and downloading the well-trained model. Another important factor is that compared to IoT devices, edge networks have better computing and storage resources, and can efficiently complete model optimization training. In addition, in order to further improve performance, we exploit the model library on the remote cloud (i.e., the 3rd layer). Regarding the remote cloud, it has a variety of large-scale data and high-performance computing capabilities and, thus, stores a large number of pretrained models. MEC networks can download the pretrained model and its pretrained parameters from the remote cloud and expand it locally according to the specific scenario, which greatly reduces the computational overhead.

To formally describe the proposed MEC-enabled hierarchical learning framework, we assume that there is one remote cloud, denoted by $\mathcal{C}$, and the model library associated with $\mathcal{C}$ is denoted by $\mathcal{M}_C = \{M_C^1, M_C^2, \ldots, M_C^j, \ldots\}$. These pretrained models can be utilized by multiple different edge nodes. Note that each individual edge node can be associated with multiple IoT devices or scenarios. In this article, we focus on the performance of emotion recognition. For the sake of simplicity, an edge node associated with multiple different IoT devices can be regarded as multiple virtual edge nodes. In other words, each individual IoT device for emotion recognition is associated with an edge node. Therefore, we utilize $\mathcal{E} = \{E_1, E_2, \ldots, E_N\}$ and $\mathcal{U} = \{U_1, U_2, \ldots, U_N\}$ to represent the set of N edge nodes and the set of N users (i.e., IoT devices), respectively. Overall, according to the logic from top to bottom (i.e., 3rd to 2nd, and then 1st), the model is gradually refined to specific scenarios.

Considering that IoT devices and edge nodes are performers and learners of the same model, we use M_i to represent the model associated with U_i and E_i. As mentioned above, the model on the edge node will utilize a pretrained model on the remote cloud to extract features. Thus, $M_{i,j}$ means that the model on edge node E_i is expanded based on M_C^j on the remote cloud, denoted by

$$M_{i,j}(\theta_{i,j}) = f\left(i, M_C^j\right) \tag{1}$$

where $\theta_{i,j}$ refers to all the parameters of the DNN-based model after localization on edge node E_i. f represents the localization for a specific scenario and the details will be discussed in Section III-B.

B. Emotion Recognition System

Regarding emotion recognition, we focus on visual information (i.e., facial expressions). Specifically, these facial expressions can come from videos or separate images. Given a facial expression example, our emotion recognition system can judge discrete emotion categories, such as happiness, disgust, surprise, etc. Regarding our proposed emotion recognition system, we define the image containing facial expression as x_k, and the real label and predicted label associated with x_k are defined as y_k and $\hat{y}_k$, respectively.

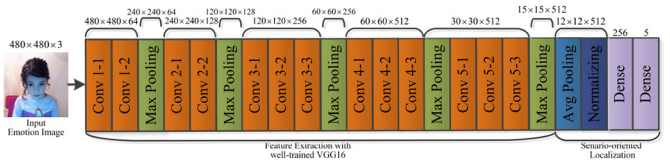


Fig. 2. Architectural details of our proposed MEC-enabled emotion recognition system with *VGG16*-based feature extraction module, as well as the output dimensions of each layer with $(480 \times 480 \times 3)$ facial expression as the initial input.

To accurately perform emotion recognition, we utilize the DNN-based model to map input x_k to output $\hat{y}_k$. The referred model is $M_{i,j}$ in Section III-A, which requires to learn its own optimal parameters $\theta_{i,j}$ through labeled facial expressions. For the convenience of description, we simplify $M_{i,j}$ to M. Regarding image data, convolutional networks can effectively perform feature extraction [33], [34]. In addition, many high-quality convolutional network architectures are stored on the remote cloud, such as *VGG16* [35], *ResNet18* [36], *AlexNet* [37], etc. As illustrated in Fig. 2, in addition to utilizing the feature extraction module (i.e., the convolutional module) of the pretrained *VGG16*, we perform specific dimensionality reduction operations (i.e., scenario-oriented localization) for the specific emotion recognition. Note that the *Conv* in Fig. 2 refers to the convolutional layer with a (3×3) filter and the *rectified linear unit* (ReLU) [38] activation function, and *Normalizing* refers to a batch normalization layer. More detailed description of the architecture and the purpose of each component can be found in Section V-B1.

Although the feature extraction module is pretrained, its combination with the scenario-oriented localization module still requires fine-tuning via local facial expressions. We divide the available facial expressions into the training dataset ($\mathcal{X}_{\text{train}}$ and $\mathcal{Y}_{\text{train}}$), validation dataset ($\mathcal{X}_{\text{validation}}$ and $\mathcal{Y}_{\text{validation}}$), and testing dataset ($\mathcal{X}_{\text{test}}$ and $\mathcal{Y}_{\text{test}}$). In the training phase, we can find the optimal parameter θ for the model M by minimizing the loss function $\mathcal{L}$ over the training dataset, which can be written as follows:

$$\theta = \arg\min_{\theta} \mathcal{L}(\mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}}; \theta) \tag{2}$$

where $\mathcal{L}$ refers to the *cross-entropy* loss function.

C. Performance Analysis of Emotion Recognition

To analyze the performance of the proposed emotion recognition system and clarify possible issues, we have analyzed the accuracy, efficiency, and robustness through experiments on the publicly available LIRIS-CSE dataset [39]. Specifically, we choose two different convolutional networks (i.e., $\mathcal{M}_C = \{M_C^1 = VGG16, M_C^2 = AlexNet\}$) in the model library of the remote cloud and then supplement

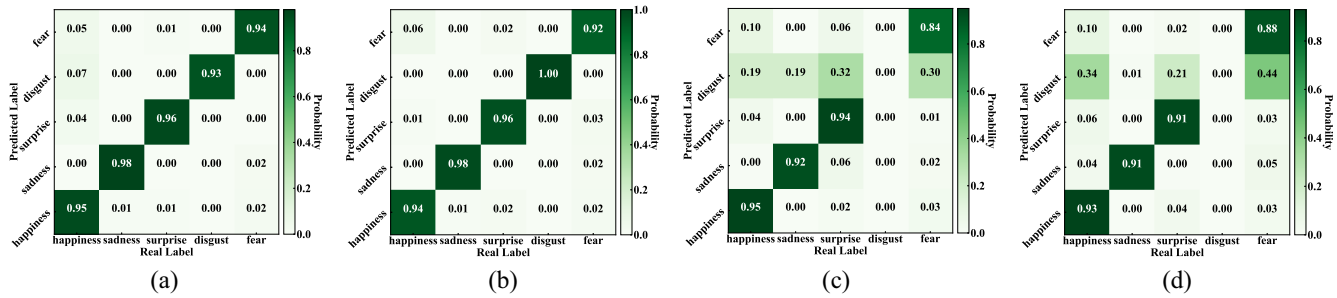


Fig. 3. Confusion matrix. (a) *VGG16*-based feature extraction in conjunction with localization module. (b) *AlexNet*-based feature extraction in conjunction with localization module. (c) *VGG16*-based feature extraction without localization module. (d) *AlexNet*-based feature extraction without localization module.

them with localization module for children's spontaneous facial expressions, respectively. Note that choosing two different feature extraction modules instead of a single one is to verify the generalization performance and the ubiquity of relevant issues. A more detailed introduction to the dataset and experimental configuration can be found in Section V.

1) Accuracy: First, we utilize the confusion matrix to analyze the accuracy of two different convolutional architectures under our proposed framework. In Fig. 3, the x-axis and y-axis, respectively, represent the real and predicted emotion categories. The values of the diagonal elements indicate the probability of each category being accurately classified, while the values of the nondiagonal elements indicate the probability of misclassifying the facial expression.

As illustrated in Fig. 3(a), our proposed MEC-enabled hierarchical emotion recognition system with *VGG16*-based feature extraction achieves an average accuracy exceeding 0.93 on all categories of facial expressions. It achieves 95.67% accuracy on the entire dataset. When it does not apply our proposed localization operation, as illustrated in Fig. 3(c), the overall accuracy of emotion recognition is reduced to 88.32%. This is because the localization module we proposed optimizes the network structure for specific scenarios, making the network structure more accurate and efficient. Moreover, our localization module is compatible with convolutional networks of different architectures.

As illustrated in Fig. 3(b), the MEC-enabled hierarchical emotion recognition system with *AlexNet*-based feature extraction can achieve 95.09% accuracy on the entire dataset. However, when there is no localization module, i.e., Fig. 3(d), its overall accuracy is only 87.29%. In other words, the application of our localization module can increase its performance by 8.94%. In the following content, our framework refers to the MEC-enabled hierarchical emotion recognition system that applies the localization module. Overall, whether we leverage *VGG16* or *AlexNet* as the basis for constructing the MEC-enabled hierarchical emotion recognition system, our proposed framework achieves the state-of-the-art performance on the publicly available LIRIS-CSE dataset [39].

2) Efficiency: In addition to improving the accuracy of emotion recognition for IoT devices, our proposed framework also emphasizes the cooperation of a remote cloud and a mobile edge network (i.e., the utilization of pretrained feature extraction modules) to achieve the improvement of model convergence efficiency in the training phase.

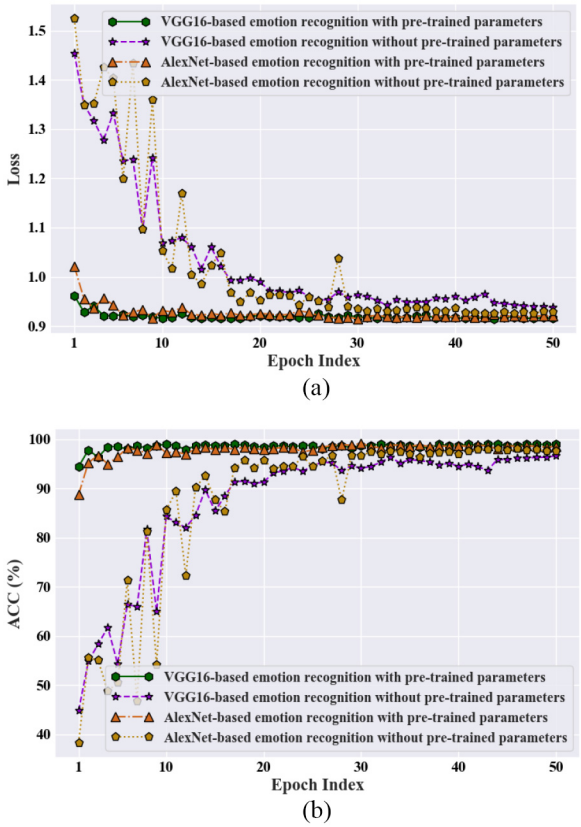


Fig. 4. Comparison of model convergence efficiency in the training phase. (a) Trend of loss. (b) Trend of accuracy.

Fig. 4 shows the trend of loss and the trend of accuracy as the training epoch increases. As introduced in Section III-A, we use the pretrained convolutional architecture (e.g., *VGG16* or *AlexNet*) on the remote cloud as the feature extraction module. For the same architecture, we can also choose to start training from scratch, i.e., without pretrained parameters for the feature extraction module. Regarding the trend of loss in Fig. 4(a), whether it is *VGG16* or *AlexNet* as the feature extraction module, if we utilize the pretrained parameters on the remote cloud, the loss of the model can be reduced to less than 1 after 2 epochs of training. Conversely, if the model is trained from scratch, it will take at least 17 epochs to get the loss below 1. As illustrated in Fig. 4(b), similar phenomena are

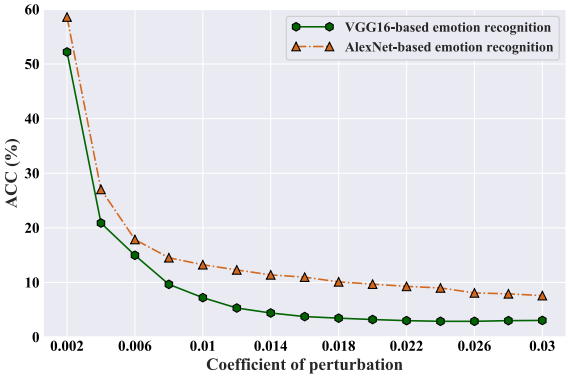


Fig. 5. Impact of different perturbation levels (i.e., ϵ) on model robustness.

also reflected in the trend of accuracy. With pretrained parameters for the feature extraction module, the accuracy of the model exceeds 95% after 2 epochs of training. While training from scratch requires at least 20 epochs to make the accuracy exceed 95%. Overall, using the pretrained parameters on the remote cloud can enable the model to converge within 5 epochs, and the time cost is much lower than that of training from scratch. This makes sense for resource-constrained scenarios. Although edge nodes are more resourceful than IoT devices, each edge node is associated with multiple IoT devices. Accelerating convergence can help more IoT devices use limited computing resources to perform more tasks within a limited time. In the following content, our framework refers to the MEC-enabled hierarchical emotion recognition system that leverages the pretrained parameters on the remote cloud for the feature extraction module.

3) Robustness: In addition to accuracy and efficiency, a major challenge for emotion recognition with IoT devices is robustness to perturbation. Because the environment of IoT devices is complex and changeable, the collected facial expression frequently receives various uncertain perturbation, resulting in differences from the perturbation-free training data. These subtle differences cannot affect the human brain's perception of emotions. However, for the DNN-based emotion recognition system, it may produce a wide range of errors and even cause security problems.

To evaluate the robustness of the emotion recognition system, we actively add different levels of perturbation to facial expressions in the testing dataset. Specifically, we utilize the fast gradient sign method (FGSM) [40] to generate perturbation. Compared with random noise in the environment, the crafted perturbation generated by FGSM is more effective in weakening the performance of DNN-based models [40]. The coefficient of perturbation is represented by ϵ, i.e., the value of x-axis in Fig. 5. More details of perturbation generation will be introduced in Section IV-A.

As illustrated in Fig. 5, as the perturbation coefficient increases, the accuracy of the emotion recognition system decreases significantly. In fact, even if the perturbation coefficient reaches 0.03, there is almost no visual difference between perturbed facial expressions and benign facial expressions. However, when the perturbation coefficient exceeds 0.006, the accuracy of the emotion recognition system is lower than that of random guessing (i.e., 20% = 1/5). This is because the accuracy of our emotion recognition system on benign facial expressions is close to 100%, and perturbation makes most of the correctly classified examples misclassified. These results demonstrate that the MEC-enabled hierarchical emotion recognition system is obviously vulnerable to perturbation. Therefore, our framework requires an effective defense mechanism to improve robustness in real-world scenarios.

IV. Proactive Perturbation-Aware Defense Mechanism for Emotion Recognition System

Our proposed framework for emotion recognition has been demonstrated to achieve state-of-the-art performance on the publicly available LIRIS-CSE dataset, while also revealing the lack of robustness in the perturbation-rich environment. In this section, we further propose data augmentation to actively defend against known and unknown perturbations. Meanwhile, we also optimize the training details of the proposed DNN-based model to achieve stable and efficient emotion recognition.

A. Immune-Driven Data Augmentation Enabled Proactive Defense

In Section III-C3, the DNN-based emotion recognition system has been demonstrated to be vulnerable to crafted facial expressions. For these crafted facial expressions, there is almost no visual difference with the original facial expressions. However, the slight perturbation can indeed cause a non-negligible degradation in accuracy. For IoT-enabled smart emotion recognition, the environment of IoT devices is complex, and the hardware configuration of IoT devices are uneven, so similar perturbations are prone to occur. Especially, for some security-related emotion recognition scenarios, it will cause immeasurable serious consequences.

To avoid the negative impact of perturbation on the DNN-based emotion recognition system, motivated by immune thoughts, we propose the data augmentation for proactive defense and strengthen the robustness of IoT-enabled smart emotion recognition. The referred data augmentation is similar to proactive immunity. In biological immunology, humans can inject vaccines to defend the body from possible pathogens in the future. Specifically, we actively add some small perturbation to benign facial expression (i.e., x_k) to construct perturbed facial expression (i.e., x'_k) to achieve data augmentation. The relationship between x'_k and x_k can be expressed as follows:

$$x'_k = x_k + \eta_k \tag{3}$$

where η_k refers to the corresponding perturbation.

Regarding generating perturbations to reduce the performance of DNN-based models, gradient-based perturbation methods [40]–[42] can effectively and accurately fool the well-trained DNN-based models. It can cause significant performance degradation with minimal perturbations. Fig. 6 provides some examples of perturbed facial expressions, and it can be found that there is almost no difference between perturbed images and the original image. The implementation

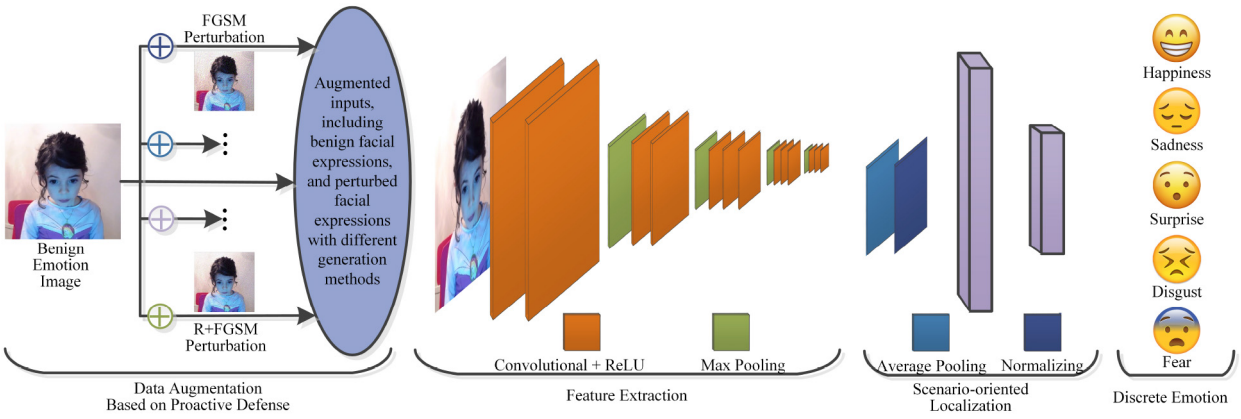


Fig. 6. Our proposed MEC-enabled hierarchical emotion recognition system in conjunction with immune-driven data augmentation. Note that the *VGG16*-based feature extraction module can be replaced with other architectures (e.g., *AlexNet*), and the output of the localization module can also be adjusted for specific scenarios. In addition to FGSM and R+FGSM, data augmentation can also utilize other perturbation generation methods [e.g., projected gradient descent (PGD)] to achieve proactive defense.

details of possible perturbation generation methods are as follows.

1) Fast Gradient Sign Method [40]: It is an attack for an L_∞-bounded adversary. As illustrated in

$$\eta_k^{\text{FGSM}} = \epsilon \cdot \text{sign}\big(\nabla_{x_k} \mathcal{L}(x_k, y_k; \theta)\big) \tag{4}$$

in the white-box scenario, FGSM calculates the gradient $\nabla_{x_k} \mathcal{L}(x_k, y_k; \theta)$ of the model with respect to the input, and then utilizes the *sign* function to get its specific gradient direction, and then multiplies it by a coefficient factor to get the required perturbation. It can achieve efficient attacks by maximizing the inner part of the saddle point formulation.

Regarding (4), ϵ refers to the sufficiently small coefficient factor. It can ensure that η_k in (3) is small enough to be discarded. Note that in the implementation, we apply a batch as the unit to calculate the required gradient through backpropagation.

2) Projected Gradient Descent [43]: It can be regarded as a multistep variant of FGSM. Different from FGSM, PGD can achieve more effective perturbation via multiple iterations of FGSM. In our model implementation process, the application of non-linear functions (e.g., *ReLU* activation function) makes our model also non-linear. For a non-linear model, if the gradient is calculated only once to generate the perturbation, the direction of gradient may not be expected. In the process of multiple iterations, PGD can continuously adjust to find the right direction of gradient.

3) R+FGSM [41]: Similar to FGSM, both R+FGSM and FGSM are single-step attacks. It has been found that the real direction of the steepest gradient may be obscured by sharp curvature artifacts located near the data points. To avoid the non-smooth vicinity of the data point before linearizing $\mathcal{L}$, Tramèr *et al.* [41] proposed to provide a small random step before the single-step FGSM attack, denoted by (5). Since it still only requires to calculate the gradient once, it achieves similar performance to PGD while ensuring the FGSM-like computational efficiency. Note that the similar performance to PGD refers to finding the correct gradient direction. For R+FGSM and PGD, the perturbation generated by each of them has different visual effects on the original facial expression image

$$\eta_k^{R+\text{FGSM}} = \tilde{x}_k - x_k + (\epsilon - \alpha) \cdot \text{sign}\big(\nabla_{\tilde{x}_k} \mathcal{L}(\tilde{x}_k, y_k; \theta)\big). \tag{5}$$

Regarding (5), $\tilde{x}_k = x_k + \alpha \cdot \text{sign}(\mathcal{N}(\mathbf{0}^d, \mathbf{I}^d))$ and $\epsilon > \alpha$.

As illustrated in Fig. 6, the perturbed facial expressions based on different generation methods will be mixed with benign facial expressions as input data. The corresponding perturbed facial expressions are defined as $\mathcal{X}'_{\text{train}}$, $\mathcal{X}'_{\text{validation}}$, and $\mathcal{X}'_{\text{test}}$. Therefore, all datasets after data augmentation are $\mathcal{X}_{\text{train}} \cup \mathcal{X}'_{\text{train}}$, $\mathcal{X}_{\text{validation}} \cup \mathcal{X}'_{\text{validation}}$, and $\mathcal{X}_{\text{test}} \cup \mathcal{X}'_{\text{test}}$. Note that we are not generating all the perturbed facial expressions at once and then training the model multiple times. In the training phase, each batch will calculate the gradient according to the current latest model to generate perturbed facial expressions with perturbation. Since the model parameters are updated in each batch, even for the same batch of data, the perturbed facial expressions generated in different epochs are different. In other words, the data augmentation we proposed allows the DNN-based emotion recognition system to encounter different perturbed facial expressions in each epoch of training and validation. This can not only produce effective perturbations but also be more in line with the dynamic characteristics of the perturbation received by IoT devices in real-world scenarios.

B. Multiscenario Perturbation-Aware Differential Training

In addition to the proposed data augmentation, we have further restructured the training architecture to more effectively defend against different perturbations in multiple scenarios.

Regarding the DNN-based emotion recognition system, the optimization of relevant models is mainly based on the gradient and backpropagation. As for the learning architecture in Section III-B, our loss function [i.e., (2)] is aimed at benign facial expressions (i.e., perturbation-free facial expressions). To learn the optimal parameters that can cope with the perturbation, we have to redefine the loss function. From a technical point of view, the newly proposed loss function should be associated with both perturbed and benign facial expressions. In addition, due to different scenarios and differences in

hardware performance, the perturbation encountered by IoT devices is complex and diverse. Moreover, the fault tolerance of the emotion recognition system in different scenarios is also different. Considering the requirements of technology and scenarios, we redefine a differential loss function for multiscenario perturbation, denoted in

$$\theta = \arg\min_{\theta} \left(\beta_0 \cdot \mathcal{L}(\mathcal{X}_{\text{train}}, \mathcal{Y}_{\text{train}}; \theta) + (1-\beta_0) \sum_{s \in \mathcal{S}} \beta_s \cdot \mathcal{L}\left(\mathcal{X}^s_{\text{train}}, \mathcal{Y}_{\text{train}}; \theta\right) \right) \quad (6)$$

Regarding (6), β_0 is the coefficient to achieve the tradeoff between perturbed and benign facial expressions. β_0 should not be less than 0.5. $\mathcal{S}$ refers to the collection of selected perturbation generation methods. $\mathcal{X}^s_{\text{train}} \subset \mathcal{X}'_{\text{train}}$, and s represents a specific perturbation generation method, such as FGSM, and β_s is the corresponding coefficient factor. In the specific implementation, we can assign different coefficients to multiple perturbation generation methods according to the requirements of specific scenarios, and $\sum_{s \in S} \beta_s = 1$.

V. Experiments

In this section, we will provide details of the selected datasets, which are collected from various real-world scenarios. Then, we will introduce the experimental setup, including the details of the network structure implementation and the configuration of various parameters. Finally, we will present the evaluation results and give the corresponding analysis.

A. Datasets

To evaluate our proposed emotion recognition system and perturbation-aware defense mechanism, we have conducted relevant experiments on the publicly available LIRIS-CSE dataset [39]. Compared to adults, children have insufficient social experience and their health and emotions urgently need the assistance of artificial intelligence. LIRIS-CSE collected spontaneous facial expression videos of 12 ethnically diverse children. Moreover, these videos record a variety of different scenarios, which are more suitable for evaluating the intelligent emotion recognition system. Since children's behaviors are recorded in a constraint-free environment, the collected facial expressions are more natural and real, which is consistent with the natural facial expressions collected by IoT devices. Specifically, these spontaneous facial expressions in LIRIS-CSE can be divided into six categories, including *happiness*, *sadness*, *anger*, *surprise*, *disgust*, and *fear*. More details (e.g., variations in recording scenarios and recording parameters) can be found in the work of Khan *et al.* [39].

It has been found that young children use expressions of *disgust* and *anger* interchangeably [44], and Khan *et al.* [39] only provided one video labeled *anger*. Thus, we have removed the only one video from LIRIS-CSE. In this article, we focus on the accuracy and robustness of single-label classification tasks, so we have removed 18 videos with multiple labels. Therefore, we finally selected $208 - 1 - 18 = 189$ videos from the 208 videos in LIRIS-CSE to evaluate our framework.

TABLE II
Statistical Properties of Original Images

	happiness	*sadness*	*surprise*	*disgust*	*fear*
$\lvert\mathcal{X}_{train}\rvert$	6,231	4,097	4,309	770	3,427
$\lvert\mathcal{X}_{validation}\rvert$	731	490	499	88	409
$\lvert\mathcal{X}_{test}\rvert$	805	522	558	100	438
Total	7,767	5,109	5,366	958	4,274

In terms of dataset division, we ensure that each video in different scenarios can be evaluated. To this end, we divide all frames of each video into the training dataset, validation dataset, and testing dataset. Among them, the training dataset can optimize the model parameters, the validation dataset can select the optimal model, and the testing dataset is used to evaluate the performance. The statistic properties of various facial expressions used in our experiment can be found in Table II. Note that Table II is for the original images. For images with perturbation (i.e., $\mathcal{X}^s_{\text{train}}$, $\mathcal{X}^s_{\text{validation}}$, and $\mathcal{X}^s_{\text{test}}$), the statistical properties can be calculated according to $\mathcal{S}$ and Table II. Since LIRIS-CSE is a public dataset, the amount of images for each emotional state is fixed, the imbalance between them is inevitable. The experimental results in Sections III-C and V-C show that our proposed framework can still achieve superior performance under this uneven condition.

B. Experimental Configuration

1) Detailed Descriptions of Architectures: For our MEC-enabled hierarchical emotion recognition system, the pretrained model on the remote cloud is utilized as our feature extraction module. In the experiments, we choose two feature extraction modules of different architectures (i.e., *VGG16* [35] and *AlexNet* [37]) to evaluate our proposed emotion recognition system and perturbation-aware defense mechanisms. Through the evaluation of different feature extraction modules, it can be demonstrated that our proposed hierarchical framework in conjunction with the proactive defense mechanism achieves broad applicability and superior compatibility. Specifically, for the pretrained *VGG16* model, we utilize its 13 convolutional layers with very small (3×3) convolutional filter to extract image features. For $(480 \times 480 \times 3)$ input images of facial expressions, the output dimensions of each layer have been illustrated in Fig. 2. Subsequently, we carry out localization operations oriented to emotion recognition. First, to reduce the features of the last convolutional layer from the pretrained *VGG16* model, we apply a global average pooling layer with a (4×4) filter, which can maintain a larger receptive field with global context information. Then, a batch normalization layer is applied to avoid internal covariant shift and speed up the training. Finally, we apply two dense layers as the tail of the entire architecture, where the first dense layer can reduce the dimensionality of the features to 256 and the second dense layer can capture independent representations for each discrete emotional state.

For the *AlexNet*-based emotion recognition architecture, we utilize five convolutional layers of pretrained *AlexNet* to extract image features. The convolutional filters of these convolutional layers are (11×11), (5×5), (3×3), (3×3), and (3×3)

Algorithm 1: Decay Strategy for Our Learning Rate

Input: $\xi_{\text{initial}} = 0.01$, $decayrate = 0.9$, $epochs = 50$, $decaystride = 5$, $k = 0$

```
decaystart = epochs * 0.5;
for k < epochs do
    if k > decaystart then
        frac = (k − decaystart)//decaystride;
        decayfrac = decayrate * *frac;
        ξ_k = ξ_{k−1} * decayfrac;
    else
        ξ_k = ξ_initial;
    k++;
```

in turn. Its localization operations are consistent with those of *VGG16*-based emotion recognition architecture, including a global average pooling layer, a batch normalization layer, and two dense layers.

2) Machine Configuration and Repeatability: Regarding the training and evaluation of our proposed framework, all relevant experiments are conducted on a single Linux machine with NVIDIA Tesla V100 GPU. The methods are programmed in Python 3.7.4 and PyTorch 1.4.0.

3) Parameter Settings: Here, we provide the detailed parameter settings in the experimental training and evaluation. In terms of hyperparameter settings, appropriate *batch size* (denoted by *BS*) and *learning rate* (denoted by ξ) can enhance the training efficiency and improve generalization. In the training phase, we set *batch size* to $BS = 128$. We utilize a decay strategy to set the learning rate. Specifically, the initial learning rate is $\xi_{\text{initial}} = 0.01$. The number of epochs for training is $epochs = 50$. For the k-th epoch training, the decaying learning rate can be calculated as Algorithm 1.

Regarding the *immune-driven data augmentation* in the training phase, we have $\mathcal{S} = \{\text{FGSM}, R + \text{FGSM}\}$, so that we can comprehensively evaluate the newly proposed defense mechanism in different perturbation scenarios, including known perturbation generation methods (i.e., FGSM and R+FGSM) and unknown perturbation generation methods (e.g., PGD). In addition, we have $\epsilon = 0.01$ in (4) and (5) for generating perturbed expressions during the training phase. As for α in (5), we set it to half of ϵ, i.e., $\alpha = 0.5 \cdot \epsilon$.

C. Experimental Results and Analysis

To fully evaluate our proposed perturbation-aware defense mechanism for the MEC-enabled hierarchical emotion recognition system, we conduct massive experiments with different pretrained feature extraction modules of different architectures, as well as multiple perturbation generation methods.

Specifically, we separately evaluate the MEC-enabled hierarchical emotion recognition system with *VGG16*-based feature extraction and *AlexNet*-based feature extraction, which can demonstrate the broad applicability of our defense mechanism. Simultaneously, for benign facial expressions and two types of perturbed (i.e., FGSM and R+FGSM) facial expressions, our defense mechanism enables the newly proposed hierarchical emotion recognition system to still maintain state-of-the-art performance. Fig. 7 shows the accuracy comparison of the referred two architectures on benign and perturbed facial expressions. For our proposed MEC-enabled emotion recognition system with *VGG16*-based feature extraction, after leveraging the defense mechanism (i.e., perturbation-aware training) to improve the robustness to perturbation, its performance can still reach 94.965% on all benign facial expressions. Compared to the MEC-enabled emotion recognition system without the defense mechanism (i.e., perturbation-free training), the performance is reduced by 0.734%. This is because our defense mechanism emphasizes defense against different perturbations by improving the generalization ability of the model. While improving the generalization ability, it will inevitably affect the performance on benign facial expressions. We can utilize β_0 in (6) to coordinate the tradeoff, ensuring that the model is suitable for different scenarios.

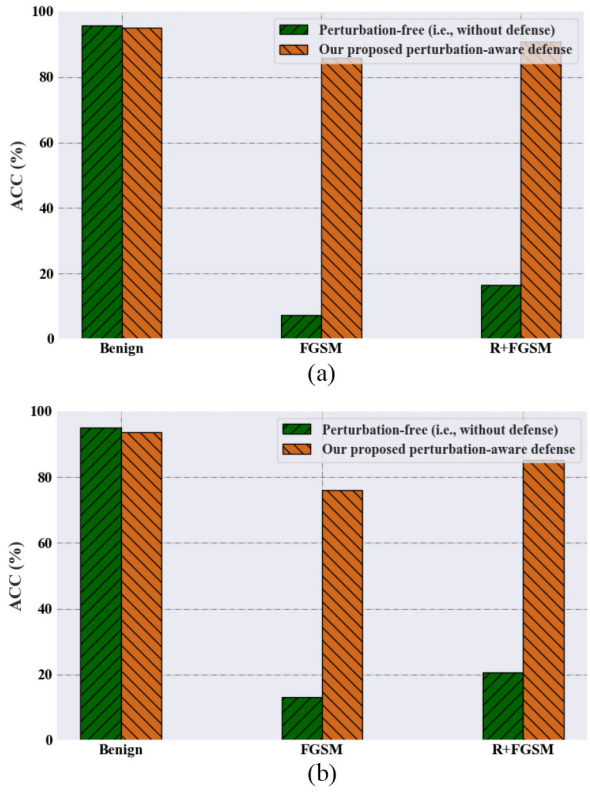


Fig. 7. Comparison of accuracy on benign facial expressions, perturbed facial expressions based on FGSM, and perturbed facial expressions based on R+FGSM. (a) *VGG16*-based feature extraction. (b) *AlexNet*-based feature extraction.

In addition to the negligible performance degradation on benign facial expressions, via the newly proposed perturbation-aware defense mechanism, our proposed MEC-enabled emotion recognition system achieve a significant performance improvement on perturbed facial expressions. As illustrated in Fig. 7(a), compared to the training without defense mechanism, the performance of emotion recognition in conjunction with the perturbation-aware defense mechanism is improved by 1087.497% on FGSM-based perturbed facial expressions and 453.014% on R+FGSM-based perturbed facial expressions, respectively. Note that the reason why the

TABLE III
ACCURACY DETAILS AND CORRESPONDING PERFORMANCE IMPROVEMENTS WITH *VGG16*-BASED FEATURE EXTRACTION MODULE

	Accuracy with benign facial expressions: 94.965% (↓ 0.734%)					
	$\epsilon_{test} = 0.005$		$\epsilon_{test} = 0.01$		$\epsilon_{test} = 0.015$	
FGSM	90.838%	(↑ 411.850%)	85.761%	(↑ 1,087.497%)	82.749%	(↑ 1,967.175%)
R+FGSM	93.314%	(↑ 159.884%)	90.838%	(↑ 453.014%)	87.990%	(↑ 864.697%)
Average	93.039%	(↑ 86.927%)	90.521%	(↑ 127.600%)	88.568%	(↑ 144.231%)

performance improvement is so high (e.g., $1,087.497\%$) is that the performance of emotion recognition system without defense mechanism is too low on perturbed facial expressions. For example, $1087.497\% = (85.761\% - 7.222\%)/7.222\%$. This demonstrates that even in perturbed scenarios, our defense mechanism ensures that the MEC-enabled emotion recognition system can achieve state-of-the-art performance. More details about accuracy of the MEC-enabled hierarchical emotion recognition system can be found in Table III.

Similar to Fig. 7(a), it can be found in Fig. 7(b) that the perturbation-aware defense mechanism is also applicable to the MEC-enabled emotion recognition system with *AlexNet*-based feature extraction. More specifically, it allows the emotion recognition system to basically maintain its original performance on benign facial expressions. Meanwhile, compared to the emotion recognition system without the defense mechanism, the performance of the emotion recognition system with our perturbation-aware defense mechanism is improved by 473.211% on FGSM-based perturbed facial expressions and 418.087% on R+FGSM-based perturbed facial expressions, respectively.

As formerly notified, we have $\epsilon = 0.01$ in (4) and (5) for generating perturbed expressions in the training phase. Regarding Fig. 7(a) and Fig. 7(b), in the testing phase, we also have $\epsilon = 0.01$ to generate perturbations and then add perturbations to all benign facial expressions, producing perturbed facial expressions for testing. To evaluate the generalization ability of the defense mechanism more comprehensively, we also generated the facial expression with different levels of perturbations (i.e., $\epsilon = 0.005$ and $\epsilon = 0.015$) for testing, which can be found in Tables III and IV. Note that for all emotion recognition systems used in the testing, the value of ϵ is 0.01 in the training phase. As illustrated in Table III, for more minor perturbations (i.e., $0.005 < 0.01$), our defense mechanism with $\epsilon = 0.01$ is more effective, achieving an accuracy of 90.838% and 93.314% on facial expressions with FGSM-based perturbation and facial expressions with R+FGSM-based perturbation, respectively. For more prominent perturbations (i.e., $0.015 > 0.01$), our perturbation-aware defense mechanism is still effective, and the performance improvement is more obvious, up to $1,967.175\%$ on facial expressions with FGSM-based perturbation. This is because, for $\epsilon = 0.015$ perturbations, the performance of the emotion recognition system without the defense mechanism is worse; thus, the relative improvement will be more prominent. For the MEC-enabled hierarchical emotion recognition system with *AlexNet*-based feature extraction, the above analysis is still valid, which can be found in Table IV. Therefore, we utilize a single level of $\epsilon = 0.01$ for data augmentation, and then it can deal with different levels (e.g., $\epsilon = 0.005$, $\epsilon = 0.01$, and $\epsilon = 0.015$) of perturbation. These results demonstrate that our perturbation-aware defense mechanism has a strong generalization ability.

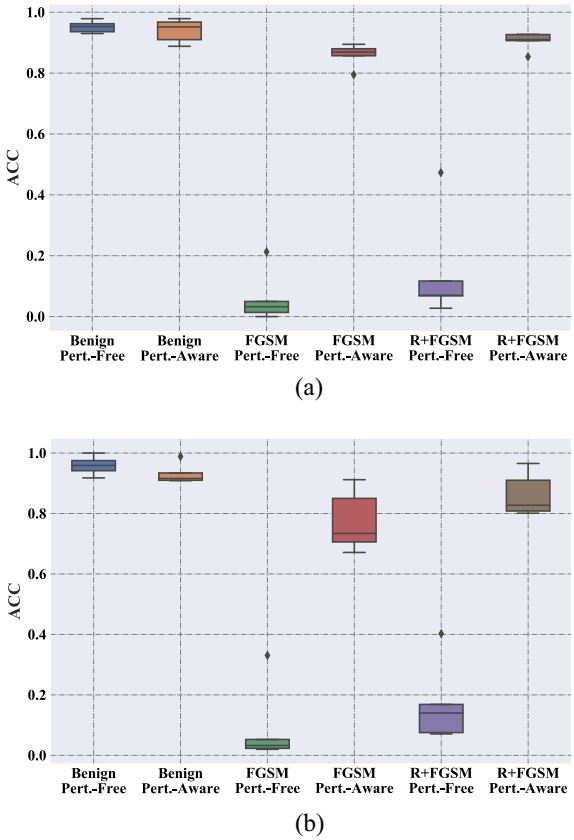


Fig. 8. Distribution characteristics of accuracy on benign facial expressions, perturbed facial expressions based on FGSM, and perturbed facial expressions based on R+FGSM. (a) *VGG16*-based feature extraction. (b) *AlexNet*-based feature extraction.

Fig. 7 shows the accuracy on the entire testing dataset in the form of *histogram*, while Fig. 8 leverages *boxplot* to describe the distribution of accuracy for each category. Specifically, each box represents the distribution of five accuracy values for five emotional states. Regarding the *x*-axis in Fig. 8, Benign, FGSM, and R+FGSM, respectively, refer to that all facial expressions are benign, all facial expressions are perturbed via FGSM, and all facial expressions are perturbed via R+FGSM. Pert.-Free refers to the MEC-enabled hierarchical emotion recognition system without any defense mechanism, while Pert.-Aware means we apply the proactive perturbation-aware defense mechanism. According to Fig. 8, it can be demonstrated that the performance of our proposed MEC-enabled hierarchical emotion recognition system for multiple

TABLE IV
ACCURACY DETAILS AND CORRESPONDING PERFORMANCE IMPROVEMENTS WITH *AlexNet*-BASED FEATURE EXTRACTION MODULE

	Accuracy with benign facial expressions: 93.603% (↓ 1.563%)					
	$\epsilon_{test} = 0.005$		$\epsilon_{test} = 0.01$		$\epsilon_{test} = 0.015$	
FGSM	85.101%	(↑ 303.514%)	75.939%	(↑ 473.211%)	67.396%	(↑ 498.172%)
R+FGSM	89.311%	(↑ 148.736%)	85.101%	(↑ 418.087%)	80.933%	(↑ 787.326%)
Average	89.338%	(↑ 76.226%)	84.881%	(↑ 104.100%)	80.644%	(↑ 109.508%)

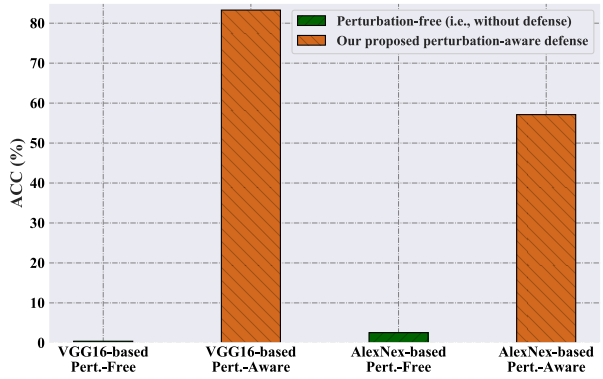


Fig. 9. Comparison of accuracy on facial expressions with perturbation generated by unknown method (i.e., PGD).

emotional states is consistent. Although the volume of data possessed by different emotional states in the dataset is different, the difference in accuracy for each emotional state is very small. This means that our method can cope with the imbalance of different categories in the dataset. In addition, we also find that while enhancing the ability to defend against perturbations, the accuracy difference among different emotional states is slightly enlarged. This is because our defense mechanism essentially improves the model's generalization ability to more emotions in different scenarios.

Fig. 9 shows the accuracy comparison of the emotion recognition system with or without the defense mechanism under the context of PGD-based perturbation. Note that for the *immune-driven data augmentation* in the training phase, we have $\mathcal{S} = \{\text{FGSM}, R + \text{FGSM}\}$ in (6), so PGD is a completely unknown perturbation generation method. Moreover, as introduced in Section IV-A2, PGD, as an iterative generation method, has stronger perturbation capabilities for DNN-based models. The evaluation in the context of PGD-based perturbation can further demonstrate the generalization ability of our proposed defense mechanism. For the MEC-enabled emotion recognition system with *VGG16*-based feature extraction, when the defense mechanism is not applied, it achieves 0.412% accuracy on PGD-based perturbed facial expressions and 7.222% accuracy on FGSM-based perturbed facial expressions. $0.412\% \ll 7.222\%$ can demonstrate that PGD has a stronger perturbation capability. After applying our perturbation-aware defense mechanism, its accuracy is as high as 83.285%. Regarding the comparison of the MEC-enabled emotion recognition system with *VGG16*-based feature extraction in this respect, it can be found from the two histograms on the right side of Fig. 9 that there are also consistent conclusions. Therefore, our defense mechanism can maintain the ability to defend against perturbation in completely unknown scenarios. In addition, as an iterative generation method, PGD requires more resources (e.g., time and computing) to generate the required perturbations. If we utilize PGD to generate perturbed facial expressions in the training phase, it will take up a lot of computing resources and increase time overhead. It is worth noting that training with $\mathcal{S} = \{\text{FGSM}, R + \text{FGSM}\}$ instead of $\mathcal{S} = \{\text{FGSM}, R + \text{FGSM}, PGD\}$ can defend against PGD-based perturbation. This can promote the deployment of our defense mechanisms in real-world resource-constrained scenarios.

VI. CONCLUSION

Since resource-constrained IoT devices cannot directly satisfy the resource requirements of DNN-based architectures in the process of optimizing parameters, in this article, we proposed the MEC-enabled hierarchical emotion recognition system, which can enable more resource-constrained IoT devices to serve emotion care in smart cities. Coupled with the localization modules for specific scenarios, the proposed emotion recognition system achieves the utilization of pretrained models on the remote cloud, as well as short-delay and high-performance computing services on the node of mobile edge network. Moreover, it can be compatible with different architectures as a feature extraction module. Extensive experimental results demonstrated that our proposed MEC-enabled hierarchical emotion recognition system for resource-constrained IoT devices can achieve state-of-the-art performance on the publicly available LIRIS-CSE dataset.

After clarifying that DNN-based emotion recognition system is vulnerable to perturbation, we further proposed the proactive perturbation-aware defense mechanism to enhance the robustness of our MEC-enabled hierarchical emotion recognition system. The newly proposed defense mechanism enables the emotion recognition system to basically maintain its original performance for benign facial expressions and can defend against the negative effects of both known and unknown perturbations. This reduces the requirements for training resources and is compatible with various perturbations in real-world scenarios.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their constructive comments and guidance.

REFERENCES

[1] M. Chen, Y. Miao, H. Gharavi, L. Hu, and I. Humar, "Intelligent traffic adaptive resource allocation for edge computing-based 5G networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 499–508, Jun. 2020.

[2] Y. Zhao, K. Xu, H. Wang, B. Li, and R. Jia, "Stability-based analysis and defense against backdoor attacks on edge computing services," *IEEE Netw.*, vol. 35, no. 1, pp. 163–169, Jan./Feb. 2021.

[3] M. Ke, Z. Gao, Y. Wu, X. Gao, and K.-K. Wong, "Massive access in cell-free massive MIMO-based Internet of Things: Cloud computing and edge computing paradigms," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 756–772, Mar. 2021.

[4] Y. Zhao *et al.*, "TDFI: Two-stage deep learning framework for friendship inference via multi-source information," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2019, pp. 1981–1989.

[5] C. Gong, F. Lin, X. Gong, and Y. Lu, "Intelligent cooperative edge computing in Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9372–9382, Oct. 2020.

[6] M. Chen *et al.*, "Living with I-fabric: Smart living powered by intelligent fabric and deep analytics," *IEEE Netw.*, vol. 34, no. 5, pp. 156–163, Sep./Oct. 2020.

[7] Y. Qian, D. Wu, W. Bao, and P. Lorenz, "The Internet of Things for smart cities: Technologies and applications," *IEEE Netw.*, vol. 33, no. 2, pp. 4–5, Mar./Apr. 2019.

[8] M. Chen, W. Xiao, J. Ma, Y. Zhang, L. Hu, and G. Tao, "Cognitive wearable robotics for autism perception enhancement," *ACM Trans. Internet Technol.*, to be published.

[9] M. Chen and Y. Hao, "Label-less learning for emotion cognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2430–2440, Jul. 2020.

[10] H. Kim, J. Ben-Othman, S. Cho, and L. Mokdad, "A framework for IoT-enabled virtual emotion detection in advanced smart cities," *IEEE Netw.*, vol. 33, no. 5, pp. 142–148, Sep./Oct. 2019.

[11] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, Aug. 2018.

[12] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2755–2766, Nov. 2020.

[13] S. Kwon *et al.*, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2020.

[14] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2013, pp. 223–232.

[15] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3359–3368.

[16] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on hilbert hypersphere with conditional Wasserstein generative adversarial nets," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 15, 2020, doi: 10.1109/TPAMI.2020.3002500.

[17] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[18] K. T. Co, L. Muñoz-González, S. de Maupeou, and E. C. Lupu, "Procedural noise adversarial examples for black-box attacks on deep convolutional networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security (CCS)*, 2019, pp. 275–289.

[19] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–9.

[20] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P.-V. Mekikis, A. Antonopoulos, and C. Verikoukis, "Online VNF lifecycle management in an MEC-enabled 5G IoT architecture," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4183–4194, May 2020.

[21] S. Shi, Z. Tang, X. Chu, C. Liu, W. Wang, and B. Li, "A quantitative survey of communication optimizations in distributed deep learning," *IEEE Netw.*, early access, Dec. 2, 2020, doi: 10.1109/MNET.011.2000530.

[22] A. Solanas *et al.*, "Smart health: A context-aware health paradigm within smart cities," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 74–81, Aug. 2014.

[23] V. Moustaka, A. Maitis, A. Vakali, and L. G. Anthopoulos, "CityDNA dynamics: A model for smart city maturity and performance benchmarking," in *Proc. Web Conf. (WWW)*, 2020, pp. 829–833.

[24] G. Muhammad and M. S. Hossain, "Emotion recognition for cognitive edge computing using deep learning," *IEEE Internet Things J.*, early access, Feb. 10, 2021, doi: 10.1109/JIOT.2021.3058587.

[25] Z. Tariq, S. K. Shah, and Y. Lee, "Speech emotion detection using IoT based deep learning for health care," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2019, pp. 4191–4196.

[26] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5683–5692.

[27] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.

[28] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 62–68, Jun. 2019.

[29] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 10143–10152.

[30] Z. Xi, Y. Niu, J. Chen, X. Kan, and H. Liu, "Facial expression recognition of industrial Internet of Things by parallel neural networks combining texture features," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2784–2793, Apr. 2021.

[31] X. Ling *et al.*, "DEEPSEC: A uniform platform for security analysis of deep learning model," in *Proc. IEEE Symp. Security Privacy (Oakland)*, 2019, pp. 673–690.

[32] M. Dai, Z. Su, R. Li, Y. Wang, J. Ni, and D. Fang, "An edge-driven security framework for intelligent Internet of Things," *IEEE Netw.*, vol. 34, no. 5, pp. 39–45, Sep./Oct. 2020.

[33] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[34] K. Huang, X. Liu, S. Fu, D. Guo, and M. Xu, "A lightweight privacy-preserving CNN feature extraction framework for mobile sensing," *IEEE Trans. Depend. Secure Comput.*, early access, Apr. 26, 2019, doi: 10.1109/TDSC.2019.2913362.

[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–6.

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

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

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[39] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," *Image Vis. Comput.*, vol. 83, pp. 61–69, Mar./Apr. 2019.

[40] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–9.

[41] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–9.

[42] S. A. Taghanaki, K. Abhishek, S. Azizi, and G. Hamarneh, "A kernelized manifold mapping to diminish the effect of adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 11340–11349.

[43] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–6.

[44] S. C. Widen and J. A. Russell, "Children's recognition of disgust in others," *Psychol. Bull.*, vol. 139, no. 2, p. 271, 2013.

Yi Zhao (Graduate Student Member, IEEE) received the B.Eng. degree from the School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an, China, in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China.

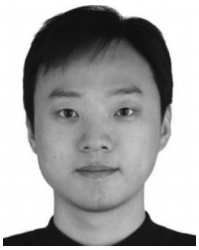
His research interests include network economics, machine learning, social network, and robust learning algorithms.

Mr. Zhao is a Student Member of ACM.

Ke Xu (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001.

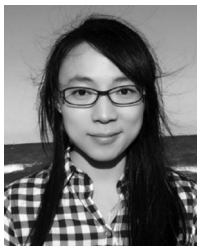
He serves as a Full Professor with Tsinghua University. He has published more than 200 technical papers and holds 11 U.S. patents in the research areas of next-generation Internet, blockchain systems, Internet of Things, and network security.

Prof. Xu has guest-edited several special issues in IEEE and Springer Journals, and also served as a Steering Committee Chair of IEEE/ACM IWQoS. He is an Editor of IEEE INTERNET OF THINGS JOURNAL. He is a member of ACM.



Haiyang Wang (Member, IEEE) received the Ph.D. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, in 2013.

He is currently an Associate Professor with the Department of Computer Science, University of Minnesota at Duluth, Duluth, MN, USA. His research interests include cloud computing, peer-to-peer networking, social networking, big data, and multimedia communications.



Bo Li (Member, IEEE) received the Ph.D. degree from Vanderbilt University, Nashville, TN, USA, in 2016.

She is currently an Assistant Professor with the Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL, USA. Her research focuses on machine learning, security, privacy, and game theory.

Dr. Li is the recipient of an MIT Technology Review TR-35 Award and a Symantec Fellowship Award.

Meina Qiao received the M.D. degree with the School of Automation Science and Control Engineering, Beihang University, Beijing, China, in 2019.

She is currently work with the Department of Computer Vision Technology, Baidu Inc., Beijing. Her research interests include computer vision, machine learning, and pattern recognition.

Haobin Shi received the Ph.D. degree in computer science and technology from the School of Computer Science, Northwestern Polytechnical University, Xi'an, China, in 2008.

He is a Professor of Computer Science and Technology with the School of Computer Science, Northwestern Polytechnical University, and a Visiting Scholar with the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan. He is the Director of the Chinese Association for Artificial Intelligence. His research interests include intelligent robots, decision support systems, artificial intelligence, multiagent systems, and machine learning.