# A Multi-View Multi-Scale Neural Network for Multi-Label ECG Classification

Shunxiang Yang, Cheng Lian, *Member, IEEE*, Zhigang Zeng, *Fellow, IEEE*, Bingrong Xu, Junbin Zang, and Zhidong Zhang

***Abstract*—The 12-lead electrocardiogram (ECG) is a common method used to diagnose cardiovascular diseases. Recently, ECG classification using deep neural networks has been more accurate and efficient than traditional methods. Most ECG classification methods usually connect the 12-lead ECG into a matrix and then input this matrix into a deep neural network. We propose a multi-view and multi-scale deep neural network for ECG classification tasks considering different leads as different views, taking full advantage of the diversity of different lead features in a 12-lead ECG. The proposed network utilizes a multi-view approach to effectively fuse different lead features, and uses a multi-scale convolutional neural network structure to obtain the temporal features of an ECG at different scales. In addition, the spatial information and channel relationships of ECG features are captured by coordinate attention to enhance the feature representation of the network. Since our network contains six view networks, to reduce the size of the network, we also explore the distillation of dark knowledge from the multi-view network into a single-view network. Experimental results on multiple multi-label datasets show that our multi-view network outperforms existing state-of-the-art networks in multiple tasks.**

***Index Terms*—ECG classification, multi-view, multi-scale, attention, knowledge distillation.**

## I. Introduction

CARDIOVASCULAR disease is one of the deadliest diseases in the world. According to statistics, in 2017, cardiovascular diseases caused approximately 17.8 million deaths worldwide [1]. A 12-lead ECG is one of the most essential diagnostic tools for screening and evaluating cardiac abnormalities. This inexpensive and noninvasive detection can provide a wealth of information that can help diagnose abnormalities associated with the electrical activity of the heart [2]. However, ECG annotation is time-consuming and labour-intensive and is performed by a specialized technician. In recent years, with the rapid development of smart medicine, automatic ECG-based arrhythmia detection can assist physicians in the diagnosis of ECG. Whereas limited achievements have been made in this area, the key to the problem is how to improve the accuracy of automatic arrhythmia detection.

Over the last decade, machine learning technologies have been used extensively in ECG signal classification. Among them, support vector machines, artificial neural networks and hidden Markov models have been widely used for the automatic detection of arrhythmias, improving their detection accuracy to a certain extent [3]. Traditional machine learning methods include three steps: preprocessing, feature extraction and classification. Of these, feature extraction is the most important, and it requires expertise in constructing manual features. This has led to traditional machine learning methods not achieving the desired accuracy in ECG detection.

Recently, deep learning, which is an important part of machine learning, has excelled in the fields of computer vision and natural language processing, and this has led to its gradual application in the detection of ECG signals. Neutral networks in deep learning are generally effective in extracting data features, making deep neural networks more accurate than general machine learning methods in performing automatic arrhythmia detection [4]. Convolutional neural networks (CNNs) [5], [6], [7], recurrent neural networks (RNNs) [8], [9] and their variants are widely used for ECG signal detection and analysis. The transformer, the most popular deep neural network structure, is also used for ECG signal classification [10]. All these deep neural network models have improved the accuracy of automatic arrhythmia detection and, thus, can assist cardiologists in the analysis of ECGs.

Most deep learning networks connect the 12-lead signals from a 12-lead ECG into a matrix [11], [12], [13], which is then input into a deep neural network to extract features. Although these networks have good ECG classification performance, they all directly use the ECG of all 12 leads, ignoring the diversity among the leads. The internationally accepted conventional 12-lead ECG is composed of standard leads I, II, and III, pressurized unipolar limb leads aVR, aVL, and aVF, and unipolar chest wall leads V1, V2, V3, V4, V5, and V6. ECGs with different leads measure different heart locations and provide different views [14]. This reflects the diversity of the 12-lead ECG, as shown in Fig. 1. In addition, the 12-lead measures the whole

Manuscript received 16 August 2022; revised 26 November 2022 and 15 December 2022; accepted 4 January 2023. Date of publication 19 January 2023; date of current version 25 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62176193, 61876219 and 62206204, and in part by the Key R&D Program of Shanxi Province under Grant 202102020101010. *(Corresponding Author: Cheng Lian.)*

Shunxiang Yang, Cheng Lian, and Bingrong Xu are with the School of Automation, Wuhan University of Technology, Wuhan 430074, China (e-mail: shunxiangyang@whut.edu.cn; chenglian@whut.edu.cn; bingrongxu@whut.edu.cn).

Zhigang Zeng is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, and Key Laboratory of Image Processing and Intelligent Control of Education Ministry of China, Wuhan 430074, China (e-mail: zgzeng@hust.edu.cn).

Junbin Zang and Zhidong Zhang are with the School of Instrument and Electronics, North University of China, Taiyuan 038507, China (e-mail: zangjunbin@163.com; zdzhang@nuc.edu.cn).

Digital Object Identifier 10.1109/TETCI.2023.3235374

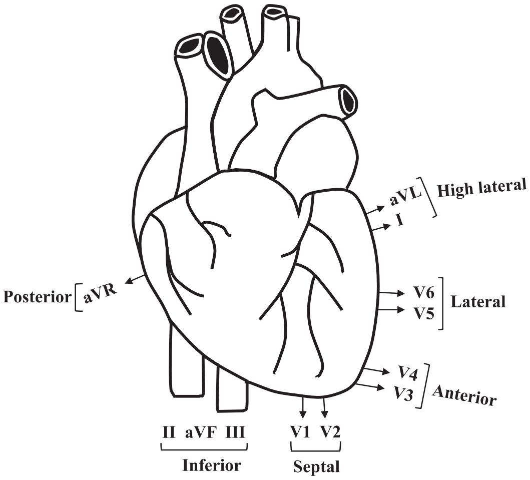

Fig. 1. Different leads of the ECG measure different locations of the heart, just like a camera taking pictures of the heart from different angles. The 12 leads are divided into 6 different locations.

heart, so it is a reflection of the whole heart. This reflects the integrity of the 12-lead ECG. To further improve the accuracy of the network in the detection of arrhythmias, the integrity of a multiple lead ECG can be exploited along with its diversity [15].

In consideration of the above discussion, this paper investigates how to fully utilize the integrity and diversity of 12-lead ECG signals to obtain more complete feature information and thus achieve higher classification accuracy. The main contributions of this paper are summarized as follows.

- We propose a deep neural network based on multi-view learning. Our multi-view network is composed of six view networks, and the division of the views is determined mainly according to the leads at different locations, thus making full use of the diversity of ECG signals. We use a specialized subnetwork to adaptively determine the multi-view fusion parameters and achieve a more effective multi-view fusion method.
- In each of the view networks, we use a multi-scale CNN structure to extract features at different scales of the ECG signal, which are essential for understanding the objects in the classification task. We also use coordinate attention to capture the spatial information and channel relationships of the ECG features to enhance the feature representation of our multi-view network.
- Since our network contains a total of six views, to reduce the size of the network, we explore the use of knowledge distillation to extract the dark knowledge of our multi-view network into a lightweight single-view network. Compared with the multi-view network, the distilled single-view network has fewer parameters and better performance.
- Experiments on three multi-label ECG datasets show that our network outperforms state-of-the-art networks on multiple classification tasks.

The remainder of this paper includes the following sections. In Section II, we introduce the ECG classification algorithm using machine learning and the ECG classification algorithm using deep learning. Section III describes our proposed multi-view and multi-scale network and the knowledge distillation process in detail. Section IV describes the datasets used in our experiments, some comparison networks, the experimental setup, and the model evaluation methods. The experimental results, ablation studies, impact of knowledge distillation on the performance of single-view networks, and hyperparametric studies in knowledge distillation are described in detail in Section V. Finally, in Section VI, we summarize the work of this paper and propose future research ideas.

## II. Related Work

The classification of ECG signals based on deep learning is a challenging and meaningful endeavor, and there are many studies that have examined this task. In this section, we present the application of traditional machine learning algorithms to ECG classification and deep learning to ECG classification.

### *A. Machine Learning Algorithms for ECG Classification*

In traditional machine learning algorithms, the first step is the manual extraction of features from the ECG signal. These features can be divided into morphological and statistical features. For example, the amplitude and width of QRS group waves are morphological features [16] while the features obtained by using a wavelet transform, hidden Markov chain, etc. are statistical features [17], [18]. When the dimensionality of the obtained features is high, principal component analysis, linear discriminant analysis and independent component analysis [19] are generally used to reduce the dimensionality of the features. The next step is to classify the features. Classification algorithms such as multilayer perceptron, K-nearest neighbors, and support vector machines [20], [21], [22] are often used in ECG classification tasks. For example, in [23], principal component analysis was first used to reduce the dimensionality of the ECG signal and then the bag of visual words is used for subsequent processing. The bag of visual words involves applying feature patch extraction, codebook construction with the aid of a clustering algorithm, and a pooling strategy for creating the final feature vectors, which are passed to a support vector machine for performing a classification task. In [24], discrete wavelet transformer coefficients are used for feature extraction, and the fully connected layer is used as a classifier for classification. Sree et al. [25] used higher-order spectral analysis to extract features, Student's *t*-tests to select features, and a random forest to classify ECGs. Due to the limitations of traditional machine learning methods that require tedious data preprocessing and manual feature production, it is difficult to obtain good performance. Many researchers have turned to studying the application of deep learning networks to ECG classification.

### *B. Deep Learning Algorithms for ECG Classification*

In recent years, with the rapid development of deep learning and its wide application in various fields of research, researchers have proposed a large number of deep neural networks for ECG classification tasks. Yıldırım et al. [5] proposed a 1D-CNN classification model for arrhythmia classification, which is a 16-layer

TABLE I
CORRESPONDENCE BETWEEN VIEWS AND LEADS

| View | Leads | Location |
|---|---|---|
| 1 | aVR | posterior |
| 2 | I, aVL | high lateral |
| 3 | V5, V6 | lateral |
| 4 | V3, V4 | anterior |
| 5 | V1, V2 | septal |
| 6 | II, III, aVF | inferior |

deep convolutional network with high classification accuracy, on the MIT-BIH Arrhythmia Database [26]. In [6], the 1D ECG is first changed into a 2D time-frequency ECG spectral image using the short-time Fourier transform, and then the obtained ECG spectral image is input into a 2D-CNN. This method transforms a 1D signal into a 2D signal so that the properties of the CNN can be fully utilized. Hou et al. [8] proposed a novel deep learning network that extracts the features of arrhythmia data based on a long short-term memory (LSTM) autoencoder network and classifies arrhythmia from the learned features using a support vector machine. Chen et al. [11] proposed a new network for ECG classification with good results, which consisted of five convolutional neural network blocks, a bidirectional gate recurrent unit (GRU) with an attention mechanism, and a fully connected classifier. The method proposed by Wang et al. [13] consists of a 33-layer CNN architecture and a non-local convolutional block attention module. The CNN architecture in this method is designed to extract the spatial and channel features of an ECG, and the nonlocal attention captures the long dependencies of representative features along the spatial and channel axes. Wagner et al. [27] used some state-of-the-art networks in time series classification to perform ECG classification, which included a full convolutional network, 1-dimensional residual network, InceptionTime, and other networks. These networks also perform well in ECG classification. Xiong et al. [28] found that simply combining multiple single-lead signals did not systematically exploit the correlation of inter lead signals, so they proposed a novel, densely connected convolutional network-based method for localizing multiple leads in myocardial infarction. The method improves the accuracy of myocardial infarct identification. In practice, each lead measures a different part of the heart, which is equivalent to observing the heartbeat from different perspective, so the diversity of multi-lead ECG can be considered from a multi-view perspective.

## III. METHODS

### A. Method Overview

This section describes our method in detail. As shown in Fig. 1, we divide the 12-lead ECG into six views of the signal according to the different locations of the heart reflected by different leads, and the correspondence between each view and lead is shown in Table I. Therefore, as shown in Fig. 2, we propose a novel multi-view multi-scale convolutional neural network. We use six individual networks to process the signals of the six views, and each network extracts the features of the leads at different locations and finally performs late fusion through a learnable fusion layer. Although good classification performance is obtained, this is also very time consuming, so we further distill the knowledge of multiple networks into one (single-view) network. The structure of single-view network is the same as the structure of each individual networks in the multi-view network, only the number of input channels is different. Our training process can be divided into two steps. In the first step, the multi-view network is trained. In the second step, the knowledge of the teacher network and the labels are used to train the single-view network.

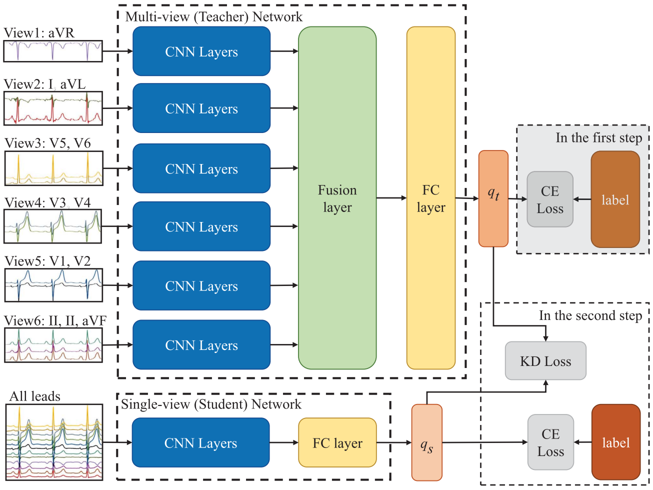

Fig. 2. Overview of our method. In the first step, our multi-view (teacher) network uses labeled information to learn ECG features. In the second step, the teacher network no longer learns. For each input $\mathbf{X}$, the probability distribution $q_t$ is output. Similarly, the single-view (student) network outputs the probability distribution $q_s$. The distillation loss (KD loss) is calculated between $q_t$ and $q_s$, and the cross-entropy loss (CE loss) is calculated between $q_s$ and the labels. Finally, these two losses are used to update the parameters of the student network.

TABLE II
THE NETWORK STRUCTURE OF THE $v$-TH VIEW

| Module | Layer Details | Feature Shape |
|---|---|---|
| Input | — | $1000 \times 1$ or<br>$1000 \times 2$ or<br>$1000 \times 3$ |
| Conv | $[conv, ks = 25, 64, stride = 1]$ | $976 \times 64$ |
| Cbca1 | $[conv\ block, ks = 15, 128, stride = 2]$<br>$coordinate\ attention$ | $488 \times 128$ |
| Cbca2 | $[conv\ block, ks = 15, 128, stride = 2]$<br>$coordinate\ attention$ | $244 \times 128$ |
| — | $adaptive\ avg\ pooling$<br>$reshape$ | $1 \times 128$<br>$128$ |

### B. Each View Network Structure

The 12-lead ECG is divided into six view signals, denoted as $\mathbf{X}^v \in \mathbb{R}^{n \times L}$, $v \in \{1, 2, ..., 6\}$, where $n$ is the number of leads and $L$ is the length of the signal. Then, $\mathbf{X}^v$ is input into the $v$-th view network with the same network structure for each view. The network structure of the $v$-th view is shown in Fig. 3. Table II shows some details of the network structure of the $v$-th view, which contains 1 convolutional layer and 2 submodules. Each submodule consists of a multi-scale convolutional block

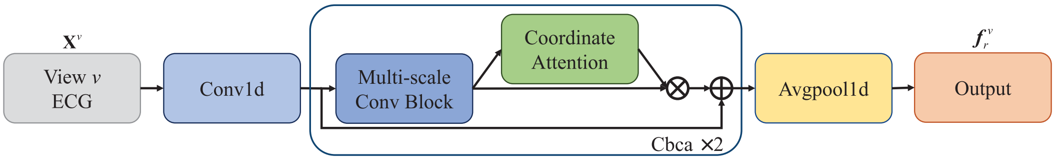

Fig. 3. The network structure of the $v$-th view, which contains 1 convolutional layer and 2 submodules. Each submodule consists of a multi-scale convolutional block and coordinate attention. Here, $\otimes$ denotes the tensor product, $\oplus$ denotes the tensor addition.

and coordinate attention. We obtain a specific representation of the view by

$$\mathbf{f}_{\mathbf{a}}^{v} = Conv\left(\mathbf{X}^{v}\right), \tag{1}$$

$$\mathbf{f}_{\mathbf{e}}^{v} = Cbca1\left(\mathbf{f}_{\mathbf{a}}^{v}\right), \tag{2}$$

$$\mathbf{f}_{\mathbf{g}}^{v} = Cbca2\left(\mathbf{f}_{\mathbf{e}}^{v}\right), \tag{3}$$

where $Conv$ represents a convolution layer. $Cbca1$ and $Cbca2$ denote the above two submodules. The difference between these two submodules is that the number of input feature channels is different. $\mathbf{f}_{\mathbf{a}}^{v}$, $\mathbf{f}_{\mathbf{e}}^{v}$, and $\mathbf{f}_{\mathbf{g}}^{v}$ denote the feature mapping of the output of these three modules of the $v$-th view network.

*1) Multi-Scale Convolution Block:* A reliable ECG classification network should have the property of capturing features at different time scales. Short-term ECG features reflect subtle changes in local regions, while long-term ECG features reflect overall trends, both of which play a key role in obtaining high-quality features. Few multi-scale neural networks have been applied to ECG classification, so we try to apply multi-scale convolutional blocks from computer vision tasks to ECG classification.

In this paper, we adopt a novel multi-scale building block called the Res2Net block [29]. Res2Net has no increase in computational load and more powerful feature extraction capability than ResNet [30]. Briefly, it replaces one 3 × 3 convolutional kernel in the residual block of the ResNet with a set of 3 × 3 convolutional kernels, and unlike the previous layer-by-layer multi-scale network, here, the multi-scale representation of the network is enhanced at a more fine-grained level.

As shown in Fig. 4, in our multi-scale convolution block, after the convolutional layer with a convolution kernel size of 1, the feature map is divided into 4 parts. The first part $\mathbf{x}_1$ goes directly to $\mathbf{y}_1$ without going through a convolution layer. The second part $\mathbf{x}_2$ goes through a convolutional layer with a convolution kernel size of 15 and then divides into two routes; one propagates forward to $\mathbf{y}_2$ and the other to $\mathbf{x}_3$, so that the third part $\mathbf{x}_3$ obtains the information of the second part $\mathbf{x}_2$. The third part $\mathbf{x}_3$, the fourth part $\mathbf{x}_4$, and so on, the number of channels in each part is $n/s$, where $n$ is the total feature channel and $s$ is the number of feature groups. Finally, the feature maps obtained from each part need to be connected and then passed through a convolutional layer with a convolution kernel size of 1. In this multi-scale convolution block, we replace the 2D convolution with a convolution kernel size of 3 with a 1D convolution with a convolution kernel size of 15. The first reason is that the ECG data we need to deal with are 1D. The second reason is that ECG, unlike images, are more likely to use large convolution kernels. Large convolution kernels also have larger receptive fields and are more likely to capture features at higher semantic levels. In the following, we will detail the coordinate attention that has been added to the module.

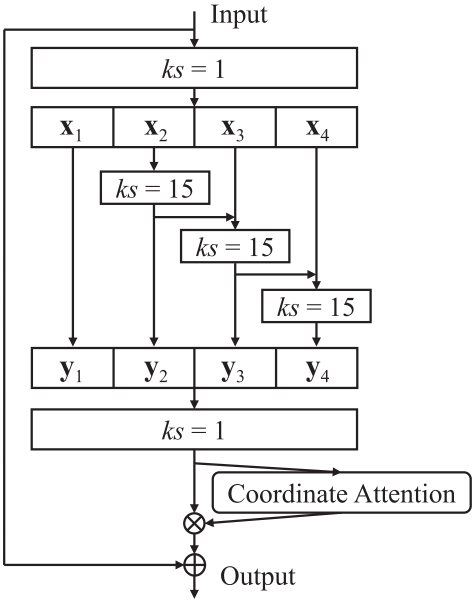

Fig. 4. Multi-scale convolutional block (number of feature groups $s = 4$) and coordination attention, where $ks$ denotes the convolutional kernel size.

*2) Coordinate Attention:* Attention has a significant role in network performance improvement; for example, Squeeze-and-Excitation (SE) attention [31] is widely used in various networks due to its small number of parameters and good performance. Generally, channel attention only constructs the relationship between each feature channel to reweight the importance of each channel while ignoring the spatial information in the feature space; however, the spatial information in the feature space is crucial to generate spatially selective attention maps.

In this paper, we use a novel attention mechanism called coordinate attention (CA), which both constructs relationships among feature channels to reweight the importance of each channel and obtains spatial information in the feature space [32]. Since the ECG is a 1D signal, global pooling of the input features along the vertical direction is not required, and all the convolutions used here are 1D. Our modified structure is shown in Fig. 5, where $C$ is the number of feature channels and $L$ is the feature length.

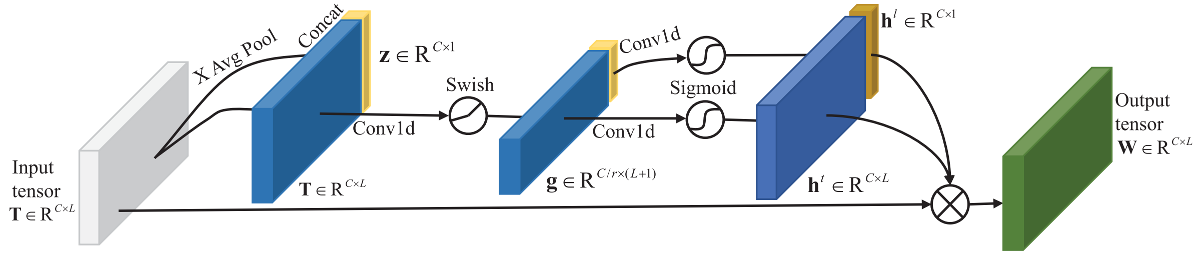

Fig. 5. Coordinate attention, which both constructs relationships among feature channels to reweight the importance of each channel and obtains spatial information in the feature space. Here, ‘X Avg Pool’ refers to 1D horizontal global pooling, $\otimes$ denotes the tensor product, ‘Swish’ and ‘Sigmoid’ are both activation functions.

Given the input $\mathbf{T}$, the information embedding of the $c$-th channel can be formulated as

$$z_c = \frac{1}{L} \sum_{0 \leq i < L} t_c(1,i), \tag{4}$$

where $t_c$ is the tensors of the $c$-th channel. The generation of coordinate attention can be described as

$$\mathbf{g} = \phi(F_1([\mathbf{T}||\mathbf{z}])), \tag{5}$$

where $[\cdot||\cdot]$ denotes the concatenation operation along the channel dimension, $\phi$ is the Swish activation function [33], $F_1$ is the convolutional layer with a convolutional kernel size of 1, and $\mathbf{g} \in \mathbb{R}^{C/r \times (L+1)}$ is the coordinate attention tensor. Then, along the channel dimension, $\mathbf{g}$ is separated into two independent tensors $\mathbf{g}^t \in \mathbb{R}^{C/r \times L}$ and $\mathbf{g}^l \in \mathbb{R}^{C/r \times 1}$, and these two tensors are transformed into the same number of channels as the input $\mathbf{T}$ using convolution, yielding

$$\mathbf{h}^t = \sigma(F_t(\mathbf{g}^t)), \tag{6}$$

$$\mathbf{h}^l = \sigma(F_l(\mathbf{g}^l)), \tag{7}$$

where $F_t$ and $F_l$ are both a convolutional layers with convolutional kernel size of 1 and $\sigma$ is the Sigmoid function. Finally, the output of coordinate attention block $\mathbf{W}$ can be written as

$$w_c(1,i) = t_c(1,i) \times h_c^l(1) \times h_c^t(i). \tag{8}$$

### C. Multi-View Fusion

Efficient fusion of features from multiple views is the key to improving network performance. In this paper, we propose an adaptive parameter fusion subnetwork to effectively fuse different view features. Our multi-view network performance is further improved by making full use of the diversity of different view features. The feature mapping $\mathbf{f}_{\mathbf{g}}^v$ of the $v$-th view is denoted as $f_{\mathbf{r}}^v \in \mathbb{R}^c$ after being processed by the adaptive average pooling layer and the reshaping function in turn, where $c$ is the number of feature channels. We get the multi-view fusion output $f_{\mathbf{o}}$ by

$$f_{\mathbf{o}} = \sum_{v=1}^{V} \sigma(F_c^v(f_{\mathbf{r}}^v)) \otimes f_{\mathbf{r}}^v, \tag{9}$$

where $V$ is the total number of views, $\sigma$ is the sigmoid function, and $F_c^v$ is a fully connected layer whose parameters are learned by a backpropagation algorithm. The subnetwork used for the fusion of $V$ view features is shown in Fig. 6. In this paper, $V = 6$. The fused features are obtained and can be input into the classifier, which is a fully connected layer here.

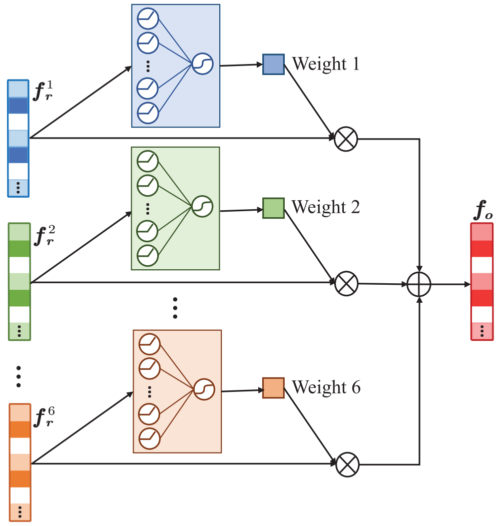

Fig. 6. The subnetwork for fusion. The weights of six view features are calculated and multiplied with their corresponding features. The weighted features of each view feature are summed.

### D. Knowledge Distillation

Knowledge distillation [34] is a model compression method, which is a training method based on the idea of a “teacher-student network” in which the knowledge contained in the trained teacher network is distilled and extracted into the student network. As shown in step 2 in Fig. 2, the teacher network is our multi-view network, while the student network is a single-view network. $q_t$ is the output probability distribution of the teacher network, while $q_s$ is the output probability distribution of the student network. The process of knowledge distillation is divided into two steps. First, the teacher network is trained, which is characterized by a relatively complex network and can also be integrated by several separately trained networks. The only requirement is that for input $\mathbf{X}$, it can output $q_t$, where $q_t$ is the output probability distribution after the mapping of softmax. Next, the student network is trained, which is a single

TABLE III
NUMBER OF ECG RECORDINGS FOR TRAINING SET, VALIDATION SET AND TEST SET IN DIFFERENT TASKS ON THE PTB-XL DATASET

| Task | # Classes | # Train | # Val | # Test | # Total |
|---|---|---|---|---|---|
| all | 71 | 17441 | 2193 | 2203 | 21837 |
| diag. | 44 | 17111 | 2156 | 2163 | 21430 |
| sub-diag. | 23 | 17111 | 2156 | 2163 | 21430 |
| super-diag. | 5 | 17111 | 2156 | 2163 | 21430 |
| form | 19 | 7202 | 904 | 882 | 8988 |
| rhythm | 12 | 16854 | 2109 | 2103 | 21066 |

network with a small number of parameters and a relatively simple network structure. Again, input $\mathbf{X}$ can output $q_s$, which is the corresponding output probability distribution after softmax mapping. The second stage also becomes induction training, where the teacher network is not longer trained, but the student network needs to be induced to train. The loss of the student network [35] can be defined as

$$L_s(\theta_s) = \alpha\tau^2 \cdot KL(q_s/\tau, q_t/\tau) + (1-\alpha) \cdot CE(q_s, y), \quad (10)$$

where $\theta_s$ denotes the parameter of the student network, $\alpha$ is a hyperparameter, $\tau$ is a temperature parameter, $KL(\cdot,\cdot)$ denotes the KL divergence, $CE(\cdot,\cdot)$ denotes the cross-entropy loss, and $y$ denotes the instance label.

## IV. EXPERIMENTS

### A. Datasets

*1) PTB-XL Dataset [36]:* PTB-XL is a recently published large accessible 12-lead ECG dataset that contains 21,837 clinical ECG recordings of 10 seconds length from 18,885 patients. The raw data are annotated by up to two ECG experts, each ECG data has multiple ECG statements, i.e., multiple labels, and each ECG statement meets the SCP-ECG criteria. This ECG dataset has a total of 71 different ECG statements (all), which can be divided into 44 diagnostic statements (diag.), 19 formal statements (form), and 12 rhythmic statements (rhythm), of which 4 formal and 4 rhythmic statements are consistent. The diagnostic statements can be further classified into 24 subclasses (sub-diag.) and 5 coarse superclasses (super-diag.) according to their hierarchical structure. Therefore, we perform six classification tasks on this dataset. Table III shows the number of ECG recordings in different tasks on this dataset. From the original PTB-XL paper, we also learn that this dataset is an imbalanced dataset. Since these data are sampled at 100 Hz and 500 Hz, and we use the 100 Hz data here. The dataset is relatively clean, so only data normalization is needed, and no extra preprocessing is required. We divide the dataset in the way recommended in the original PTB-XL paper, where groups 1 to 8 are the training set, group 9 is the validation set and group 10 is the test set.

*2) CPSC 2018 Dataset [37]:* This multi-label dataset was derived from the China Physiological Signalling Challenge 2018 (CPSC 2018), which contains 6,877 12-lead ECG recordings ranging from 6 to 60 seconds. The dataset has 9 classes with a sampling frequency of 500 Hz. First, we reduce the sampling frequency of the data to 100 Hz. Second, due to the varying lengths of the data, the data longer than 10 seconds were cropped, and those smaller than 10 seconds were upsampled. In this way, we obtain the final data all 10 seconds in length and divide them into 10 groups, with the 9th group as the validation set, the 10th group as the test set, and the rest as the training set.

*3) Hefei High-Tech Cup (HFHC) Dataset [38]:* This multi-label dataset contains 20,335 medical ECG samples provided by the Engineering Research Centre of the Ministry of Education for Mobile Health Management System of Hangzhou Normal University. The dataset has 34 classes. Each sample has 8 leads, which are I, II, V1, V2, V3, V4, V5 and V6. The data of the remaining 4 leads can be calculated from the data of the first 8 leads. The sampling frequency of each data point is 500 Hz, and the length is 10 seconds. We reduce the sampling frequency of the data to 100 Hz and then perform the data normalization process. The training set, validation set, and test set of this dataset account for 60%, 20%, and 20%, respectively.

### B. Introduction of the Comparison Network

*1) Baseline Network:* (1) Fcnwang [39]: This network is a 3-layer convolutional network with convolutional kernels of sizes 8, 5, and 3, respectively. (2) Resnet1dwang [39]: This network is composed of three residual blocks with 10 convolutional layers and 1 fully connected layer. (3) LSTM [40]: Long short-term memory is a special type of RNN mainly used to solve the gradient disappearance and gradient explosion problems during the training of long sequences. Compared with a normal RNN, LSTM can achieve better performance in longer sequences. (4) BiLSTM [41]: Bi-directional long short-term memory is a combination of forward LSTM and backward LSTM. Since LSTM cannot encode backward-to-forward information, BiLSTM can better capture the semantic dependencies in both directions. (5) ViT [42]: The vision transformer (ViT) combines computer vision and natural language processing domain knowledge to chunk original images, flatten them into sequences, and input them into the encoder part of the original transformer model and finally link a fully connected layer to classify the images. In this paper, the patch size is set to 50, the number of encoder blocks is set to 2, and the number of multi-head attention modules is set to 12.

*2) State-of-The-Art Network:* (1) InceptionTime [43]: InceptionTime is based on InceptionV4, which has six improved Inception modules. Each Inception module is concatenated by four convolutional layers with convolutional kernel sizes of 39, 19, 9 and 1, respectively. (2) Xresnet1d101 [44]: This network is an improvement on ResNet. It adds a 2×2 average pooling layer with a step of size 1 to the downsampling block in the residual block of ResNet and adjusts the step size of the convolution kernel. (3) ACNet [11]: The network consists of five convolutional blocks, a bidirectional GRU with attention, and a fully connected layer. (4) ATI-CNN [12]: The network was constructed integrating a fully convolutional neural network, LSTM layers and an attention module, where the convolutional layer is similar to the structure of the VGG network. (5) MiniRocket [45]: The model uses a random convolution kernel to obtain the transformed features of the time series data and then trains a linear classifier with the transformed features. (6) MobileNetV3 [46]: This network is a lightweight network that utilizes an automatic mobile neural architecture search method and a platform-aware algorithm for mobile applications. The network has a relatively small number of parameters and is fast and suitable for application in mobile intelligent platforms.

### C. Implementation Details

Our network is implemented using the PyTorch framework. For a fair comparison, the PyTorch framework is also used in the comparison networks. The training parameters of all networks in the experiments are set the same: the batch size is set to 64, the fixed learning rate is set to 0.001, the optimizer is the adaptive moment estimation (Adam) optimizer, and the total number of training epochs is 200. Due to the multi-label problem in these datasets, we use the $nn.BCELoss$ loss function here, which combines the sigmoid layer and binary cross entropy loss to achieve numerical stability. All of our experiments are performed on a desktop PC equipped with an Intel(R) Core(TM) i7-10700 K CPU, 32 GB of RAM and an RTX 2080Ti GPU with 11 GB of video memory. Our code is available at https://github.com/ysxGitHub/MVMS-net.

### D. Multi-Label Evaluation Metrics

The evaluation metrics required for multi-label classification are different from those for single-label classification. The evaluation metrics commonly used for multi-label classification are the Hamming loss, the mean average precision (mAP), and the area under the curve (AUC), which can be classified as example-based metrics and label-based metrics [47]. Since these datasets suffer from category imbalance, and the AUC is not sensitive to whether the sample categories are balanced, the evaluation metric we use here is the macro-AUC score. The AUC was defined as the area under the receiver operating characteristic (ROC) curve. The horizontal coordinate of the ROC curve is the false positive rate (FPR), and the vertical coordinate is the true positive rate (TPR).

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{11}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

where FP, TN, TP, and FN indicate the numbers of false positives, true negatives, true positives, and false negatives, respectively. The AUC is calculated using the approximation method to find the approximate value, which will not be described here. For the details, please refer to [48]. However, the AUC reflects too general information, and the problem of an unbalanced dataset causes the $F_1$ score to not be very applicable as a metric. Therefore, we also give the TPR, also known as the sensitivity (SEN), which is the percentage of patients who actually have the disease and are correctly judged as having the disease by the criteria of that screening test. The SEN reflects the ability of the screening test to detect patients.

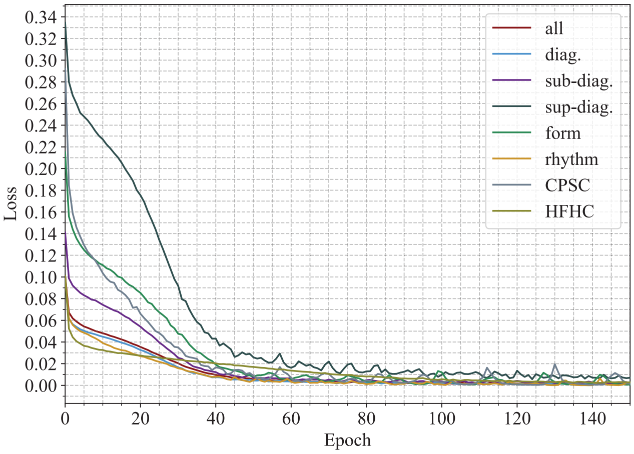

Fig. 7. Training losses vs. training epochs. Loss variation curves of our multi-view network during training on eight classification tasks for three datasets.

## V. Results and Discussion

### A. Experimental Results

We compare the macro-AUC and SEN scores of each network on the test set of each dataset. As seen in Tables IV and V, in terms of the macro-AUC score, our network, which is also the distilled single-view network, achieves the best scores in almost all tasks on all three datasets (except for the rhythm classification task on the PTB-XL dataset). Compared to the state-of-the-art network InceptionTime, macro-AUC score increased by about 1.51% on average (0.78%, 1.53%, 1.31%, 0.44%, 2.44%, and 0.62% under the six annotation levels of the PTB-XL dataset, respectively. 1.33% increase on the CPSC dataset and 3.63% increase on the HFHC dataset). This indicates that our network has better classification performance compared to other state-of-the-art networks. In terms of SEN score, our network achieves the best performance in all classification tasks for all three datasets except for the form classification task on the PTB-XL dataset. It also indicates that our network has a very good effect for detecting real diseases.

The number of parameters, floating point operations (FLOPs) values, and inference time for each network are shown in Table VI. Although the number of parameters of our network is close to the number of parameters of InceptionTime, the FLOPs of our network is much smaller than the FLOPs of InceptionTime. From the view of inference time, the inference speed of our network is faster than that of InceptionTime. The number of parameters, FLOPs values and inference time of ACNet are small, but its performance is moderate. In summary, our network has not only relatively good performance, but also relatively fast inference speed. Fig. 7 shows the variation of the training loss of our multi-view network on the three datasets (the first stage). The training loss decreases rapidly within the first 50 epochs and finally converges to zero.

### B. Ablation Studies

*1) Experiments on Convolutional Block Structure and Convolutional Kernel Size:* We compare the effect of the

TABLE IV
COMPARISON OF THE MACRO-AUC AND SEN SCORES OF SEVERAL DEEP NEURAL NETWORKS WITH OUR NETWORK ON MULTIPLE TASKS OF THE PTB-XL DATASET

| Network | all | | diag. | | sub-diag. | | super-diag. | | form | | rhythm | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN |
| fcn_wang [39] | 91.24 | 68.88 | 91.97 | 65.63 | 90.15 | 69.90 | 92.02 | 76.42 | 85.43 | 49.42 | 92.61 | 88.37 |
| resnet1d_wang [39] | 91.13 | 69.32 | 92.12 | 66.44 | 90.91 | 68.23 | 92.17 | 75.43 | 86.04 | 52.97 | 95.06 | 89.71 |
| LSTM [40] | 92.74 | 67.97 | 91.72 | 64.28 | 91.66 | 67.36 | 92.44 | 76.21 | 85.14 | 48.82 | 91.27 | 87.19 |
| BiLSTM [41] | 91.35 | 69.53 | 91.37 | 64.13 | 91.98 | 69.13 | 92.19 | 78.61 | 82.14 | 49.24 | 94.93 | 88.42 |
| InceptionTime [43] | 92.11 | 71.93 | 92.82 | 66.24 | 92.39 | 70.95 | 92.61 | 78.17 | 86.66 | 59.05 | 95.03 | 90.11 |
| Xresnet1d101 [44] | 90.61 | 68.89 | 91.42 | 64.69 | 89.98 | 67.94 | 91.95 | 75.27 | 82.02 | 51.35 | 94.90 | 91.15 |
| MiniRocket [45] | 60.38 | 67.21 | 57.24 | 67.41 | 62.06 | 69.31 | 73.00 | 69.45 | 54.05 | **60.72** | 51.71 | 62.96 |
| ViT [42] | 78.23 | 57.69 | 81.73 | 48.63 | 83.04 | 51.83 | 81.64 | 61.49 | 71.61 | 21.43 | 77.37 | 81.26 |
| MobileNetV3 [46] | 90.12 | 68.81 | 89.53 | 61.34 | 89.25 | 64.06 | 91.29 | 75.40 | 81.80 | 48.25 | 96.22 | 91.39 |
| ACNet [11] | 89.54 | 67.81 | 89.51 | 66.27 | 89.69 | 67.45 | 92.33 | 78.14 | 83.12 | 47.65 | 95.90 | 77.82 |
| ATI-CNN [12] | 89.51 | 70.23 | 90.83 | 67.17 | 90.80 | 70.34 | 91.81 | 77.77 | 83.54 | 55.58 | **96.69** | 91.17 |
| Ours‡ | 92.78 | 71.13 | 93.84 | 67.62 | 93.66 | 71.35 | 93.08 | 78.13 | 87.40 | 51.18 | 95.15 | 89.90 |
| Ours | **92.89** | **73.66** | **94.35** | **68.97** | **93.70** | **74.21** | **93.25** | **81.29** | **89.10** | 56.16 | 95.65 | **91.83** |

The best and second best results for each dataset are bolded and underlined, respectively. ‡: our multi-view network. Ours: distilled single-view network (unit: %).

TABLE V
COMPARISON OF THE MACRO-AUC AND SEN SCORES OF SEVERAL DEEP NEURAL NETWORKS AND OUR NETWORK ON THE CPSC 2018 DATASET AND HFHC DATASET

| Network | CPSC | | HFHC | |
|---|---|---|---|---|
| | AUC | SEN | AUC | SEN |
| fcn_wang [39] | 91.85 | 70.59 | 89.08 | 84.71 |
| resnet1d_wang [39] | 94.72 | 76.74 | 92.18 | 91.23 |
| LSTM [40] | 94.81 | 74.93 | 89.82 | 87.31 |
| BiLSTM [41] | 95.06 | 77.75 | 91.21 | 88.43 |
| InceptionTime [43] | 94.47 | 79.28 | 92.00 | 90.89 |
| Xresnet1d101 [44] | 95.22 | 82.61 | 90.88 | 90.87 |
| MiniRocket [45] | 73.23 | 60.49 | 67.39 | 62.42 |
| ViT [42] | 82.58 | 47.10 | 61.17 | 35.70 |
| MobileNetV3 [46] | 95.23 | 81.52 | 92.77 | 91.55 |
| ACNet [11] | 94.66 | 80.94 | 91.46 | 90.73 |
| ATI-CNN[12] | 94.73 | 83.26 | 91.85 | 92.47 |
| Ours‡ | 95.32 | 80.07 | 93.50 | 90.29 |
| Ours | **95.80** | **84.98** | **95.63** | **95.47** |

TABLE VI
THE NUMBER OF PARAMETERS, FLOPS VALUE, AND INFERENCE TIME FOR EACH NETWORK

| Network | Params ($10^6$) | FLOPs ($10^6$) | Infer (ms) |
|---|---|---|---|
| fcn_wang [39] | 0.28 | 137.91 | 4.37 |
| resnet1d_wang [39] | 0.29 | 16.64 | **1.70** |
| Xresnet1d101 [44] | 1.53 | 69.73 | 19.08 |
| InceptionTime [43] | 0.47 | 237.58 | 15.13 |
| LSTM [40] | 0.81 | 404.10 | 35.18 |
| BiLSTM [41] | 2.14 | 2140.68 | 70.51 |
| ViT [42] | 9.34 | 94.54 | 3.21 |
| MobileNetV3 [46] | 1.48 | 10.16 | 9.04 |
| ACNet [11] | **0.03** | **2.66** | 3.55 |
| ATI-CNN [12] | 5.00 | 287.34 | 13.42 |
| Ours‡ | 2.24 | 446.75 | 52.68 |
| Ours | 0.39 | 82.27 | 8.91 |

Infer indicates that we measure the inference time for 64 ECGS on the PTB-XL dataset. MS is milliseconds.

multi-scale convolutional blocks used in this paper with the effect of the residual blocks in ResNets on our multi-view network performance on the diagnostic statement (diag.) classification task in the PTB-XL dataset, and also investigated the effect of the convolutional kernel size on the ECG classification task.

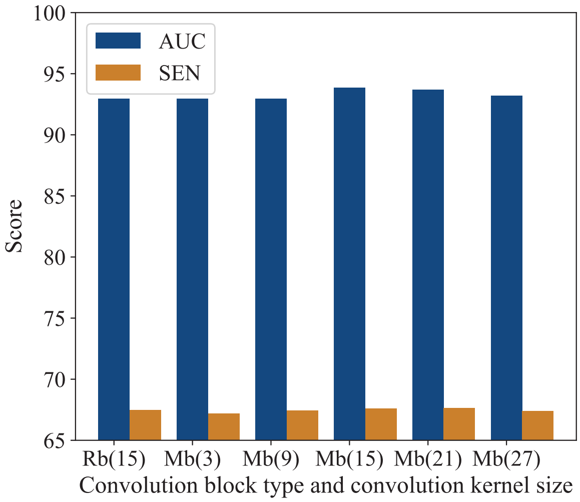

Fig. 8. The macro-AUC and SEN scores for different convolution block types and convolution kernel sizes of our multi-view network, where “Rb(15)” denotes a residual convolution block with a convolution kernel size of 15, “Mb(3)” denotes a multi-scale convolution block with a convolution kernel size of 3, and the rest are also the same.

As shown in Fig. 8, the macro-AUC score and the SEN score of our multi-view network are both improved when replacing the residual blocks with the multi-scale convolutional blocks, which also indicates that the multi-scale convolutional blocks also perform well in the ECG classification task. Fig. 8 shows that the convolutional kernel size also has a great influence on the performance of our multi-view network, the best performance of which is achieved when the convolutional kernel size is set to 15 in this paper. The size of the convolution kernel used in general ECG classification networks is usually larger than that used in image classification networks.

*2) Comparative Experiment of Attention Mechanisms:* Table VII shows the macro-AUC and SEN scores of our multi-view network with and without attention on all tasks of PTB-XL dataset. Additionally, we compare the performance of several types of attention. Among them, the convolutional block

TABLE VII
THE EFFECT OF ATTENTION ON NETWORK PERFORMANCE

| Attention | | ✗ | CBAM | SE | CA |
|---|---|---|---|---|---|
| all | AUC | 92.21 | 92.28 | **92.93** | 92.78 |
| | SEN | 70.71 | 70.99 | 70.89 | **71.13** |
| diag. | AUC | 92.93 | 92.76 | 93.81 | **93.84** |
| | SEN | 66.80 | 67.06 | 67.26 | **67.62** |
| sub-diag. | AUC | 93.04 | 92.99 | 93.15 | **93.66** |
| | SEN | 68.62 | 69.82 | 69.07 | **71.35** |
| super-diag. | AUC | 92.90 | 92.75 | 92.73 | **93.08** |
| | SEN | 78.06 | 78.60 | **79.55** | 78.13 |
| form | AUC | 86.94 | 87.24 | 86.12 | **87.40** |
| | SEN | 49.73 | 50.28 | **51.94** | 51.18 |
| rhythm | AUC | 91.58 | 94.55 | **95.41** | 95.15 |
| | SEN | 89.30 | 89.75 | 89.37 | **89.90** |

The best and second best results for each dataset are bolded and underlined, respectively (unit: %).

TABLE VIII
COMPARISON AMONG OUR UNDISTILLED SINGLE-VIEW NETWORK, MULTI-VIEW NETWORK (OURS‡), AND LATE FUSION BY CONCATENATION ON THE PTB-XL DATASET

| task | single-view | | concatenation | | Ours‡ | |
|---|---|---|---|---|---|---|
| | AUC | SEN | AUC | SEN | AUC | SEN |
| all | 91.38 | 64.74 | 92.37 | 70.99 | **92.78** | **71.13** |
| diag. | 93.10 | 57.49 | 93.15 | 67.43 | **93.84** | **67.62** |
| sub-diag. | 91.97 | 60.61 | 93.43 | 70.24 | **93.66** | **71.35** |
| super-diag. | 92.24 | 71.86 | 92.65 | **78.14** | **93.08** | 78.13 |
| form | 84.27 | 42.14 | 86.85 | 51.05 | **87.40** | **51.18** |
| rhythm | 94.56 | 90.34 | 94.93 | 89.51 | **95.15** | **89.90** |

The best and second best results for each dataset are bolded and underlined, respectively (unit: %).

attention module (CBAM) performs spatial self-attention first and then channel self-attention [49]. SE attention only performs channel self-attention. CA performs both channel self-attention and spatial self-attention. As seen in Table VII, a suitable attention added to our multi-view network can improve its performance. CBAM has limited improvement in our multi-view network performance relative to SE attention and CA. Our multi-view network performance using SE attention is the best on a few classification tasks. Overall, our multi-view network performance using CA is the best, with improvements in all tasks relative to the no attention mechanisms. The use of attention mechanisms in ECG classification is also an important way to improve network performance, as can also be understood from [50], [51], where many ECG classification networks are now used as attention mechanisms.

*3) Multi-View Fusion Analysis:* There are many methods for multi-view fusion, such as late fusion by concatenation, mid fusion by concatenation, Squeeze-and-Excitement gate, and Non-Local gate [52]. Compared with the traditional three modal fusion of images, text and speech, our multi-view network has six views, and many of the fusion methods cannot be used directly on our multi-view network. Here, we compare our approach with the most commonly used multi-view fusion method, i.e., late fusion by concatenation. We also compare the performance of our multi-view network with our undistilled single-view network on the full task. The structure of the undistilled single-view network is the same as that of each individual network in the multi-view network, only the number of input channels is different. As shown in Table VIII, our multi-view network outperforms our undistilled single-view network, which illustrates that our multi-view network can take full advantage of the integrity of the ECG and the diversity of the different leads of the ECG. Table VIII shows that our proposed adaptive parameter fusion subnetwork outperforms the late fusion by concatenation, which also reflects that our multi-view fusion method may be able to fuse the features of each view more effectively.

## C. Effect of Distillation on the Network

Tables IV and V show that our distilled single-view network performs better than our undistilled single-view network, in terms of both the macro-AUC and SEN scores; and better than our multi-view network. This shows that using knowledge distillation to extract teacher network knowledge into the student network is helpful to improve the performance of the student network. Table VI shows that the FLOPs value and the inference time of our distilled single-view network are much smaller than those of our multi-view network. With knowledge distillation, we obtain a network with good performance and fast inference.

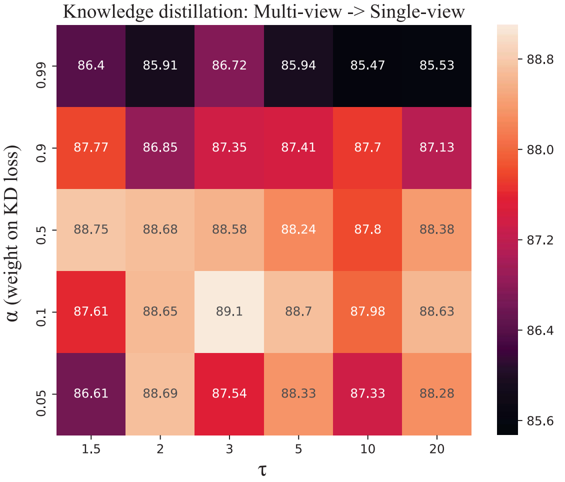

Fig. 9. On the rhythm classification task of the PTB-XL dataset, the impact of hyperparameters $\tau$ and $\alpha$ of knowledge distillation on model performance was evaluated. The lighter the color, the larger the macro-AUC.

The performance of the student network after knowledge distillation in the computer vision task is usually somewhat worse than that of the teacher network, but it is indeed surprising to see different results from knowledge distillation in computer vision in Tables IV and V. We try to explain this phenomenon and thus distill knowledge from three more larger networks into smaller networks with the same structure, and the results are shown in Table IX. Distilling the knowledge of Xresnet1d101 into Xresnet1d18, the performance of the distilled Xresnet1d18 is also higher than that of Xresnet1d101 on most tasks. Our smaller single-view network, i.e., a network with only 2 multi-scale convolutional blocks, also performs better after distillation than a single-view network with 6 multi-scale convolutional blocks. However, the distilled InceptionTime with

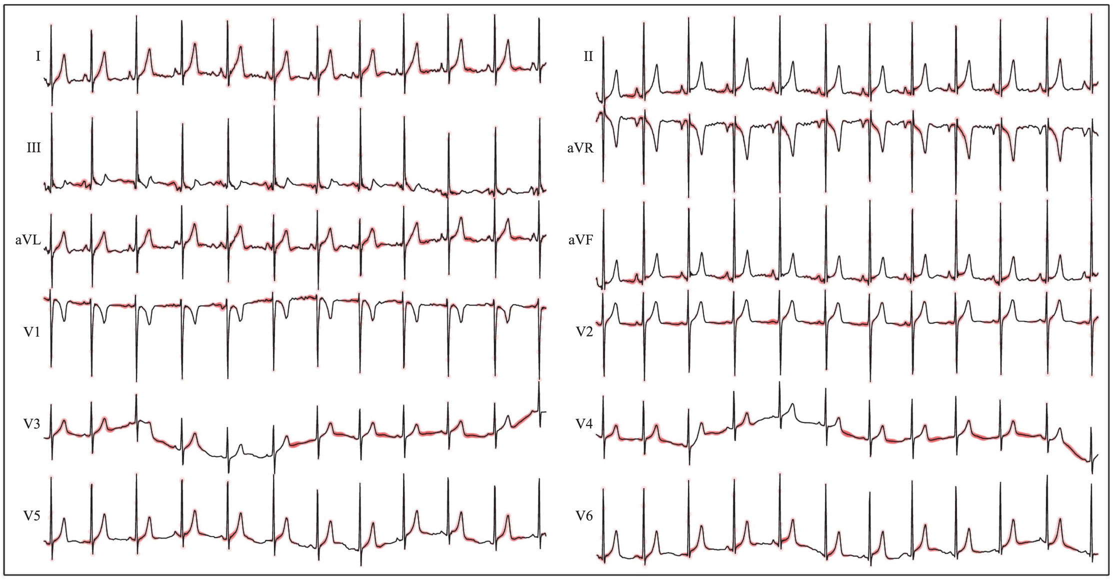

Fig. 10. Visualization of 12-lead ECG features. Since the ECG is a 1-dimensional signal, we use the red dot size to indicate the intensity of attention to that part of the ECG rather than the heatmap.

TABLE IX
Performance of the Teacher Network and the Student Network on the PTB-XL Dataset

| Network | all | | diag. | | sub-diag. | | super-diag. | | form | | rhythm | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN | AUC | SEN |
| $T$: Xresnet1d101 | 90.61 | 68.89 | 91.42 | 64.69 | 89.98 | 67.94 | 91.95 | 75.27 | 82.02 | **51.35** | 94.90 | **91.15** |
| $S$: Xresnet1d18† | **92.17** | **70.36** | 92.28 | 66.32 | 91.81 | 66.39 | **91.99** | 75.91 | **86.24** | 48.03 | **95.26** | 90.39 |
| $S$: Xresnet1d18 | 90.67 | 69.45 | **93.19** | **66.47** | **92.46** | **68.13** | 91.85 | **76.71** | 83.20 | 49.13 | 89.36 | 89.73 |
| $T$: single-view (6 blocks) | 90.90 | 70.33 | 91.18 | 64.87 | 90.62 | 66.03 | 91.29 | 77.75 | 83.67 | 46.63 | 95.63 | **91.58** |
| $S$: single-view† (2 blocks) | 91.13 | **72.56** | 91.88 | **76.45** | 91.17 | **73.76** | **92.53** | **78.45** | 83.17 | **66.93** | **96.10** | 91.01 |
| $S$: single-view (2 blocks) | **91.38** | 64.74 | **93.10** | 57.49 | **91.97** | 60.61 | 92.24 | 71.86 | **84.27** | 42.14 | 94.56 | 90.34 |
| $T$: InceptionTime (6 blocks) | **92.11** | **71.93** | **92.82** | 66.24 | 92.39 | **70.95** | 92.6 | **78.17** | **86.66** | 59.05 | **95.03** | **90.11** |
| $S$: InceptionTime† (3 blocks) | 90.00 | 61.72 | 92.23 | 56.52 | 92.47 | 58.38 | **92.79** | 75.89 | 84.64 | **60.43** | 90.16 | 88.87 |
| $S$: InceptionTime (3 blocks) | 91.42 | 71.21 | 92.11 | **66.53** | **93.03** | 68.95 | 92.31 | 76.86 | 85.77 | 52.30 | 89.98 | 89.61 |

$S$: Student Network. †: Distilled Network (unit: %).

3 convolutional blocks has worse performance than the original InceptionTime with 6 convolutional blocks. Why does this happen now? This may be caused by the network structure for this reason. The training process of the large network may reach suboptimal points or overfit some bad distributions. At this time, the small network is then trained, and since the small network has a teacher network to guide the training, and its training process also involves labels, it may bring the small network closer to the global optimal point. Why not just use the small model to do ECG classification? Similarly, we show the performance of the student network without distillation in Table IX. The student network performs better after distillation than the student network before distillation. Overall, the results presented in Table IX are subject to further investigation in our future work.

## D. Hyperparametric Analysis of Knowledge Distillation

This paper focuses on finding the optimal hyperparameters using the grid search method. The grid search records each possible combination of hyperparameters, trains them, and finds the optimal result corresponding to the hyperparameters. Fig. 9 shows the effect of the temperature parameter $\tau$ and hyperparameter $\alpha$ on the model macro-AUC score during knowledge distillation on the rhythmicity (rhythm) statement classification task. The results show that the selection of these two parameters has a large impact on the experimental results. The CPSC 2018 dataset and the HFHC dataset are set with $\tau$ set to 2 and $\alpha$ set to 0.95 on the macro-AUC score. The settings of $\tau$ and $\alpha$ on the macro-AUC score for the PTB-XL dataset are shown in Table X.

TABLE X
The Settings of $\tau$ and $\alpha$ on the macro-AUC Score for the PTB-XL Dataset

| Parameters | all | diag. | sub-diag. | super-diag. | form | rhythm |
|---|---|---|---|---|---|---|
| $\tau$ | 2 | 5 | 5 | 3 | 3 | 2 |
| $\alpha$ | 0.1 | 0.9 | 0.1 | 0.9 | 0.1 | 0.5 |

## E. ECG Feature Visualization

Networks are often treated as black boxes in deep learning. Grad-CAM [53] is commonly used in computer vision to visualize image features to show what features are extracted by convolution. Grad-CAM is a weighted gradient-like activation mapping that allows the gradient of any target feature to pass through the last convolution layer to produce a roughly local feature map, highlighting regions of the image that are important for target prediction classification. In this paper, we use this method to visualize ECG features, which requires changes to this method since ECG is a 1-dimensional signal. Fig. 10 demonstrates a feature visualization of an ECG. Our network pays more attention to QRS wave clusters on leads II, III, aVR (view 6), V1, V2 (view 5), and aVF (view 1), more attention to P and T waves on leads I, aVL (view 2), V3, and V4 (view 4), while the first few ECG beats on leads V5, V6 (view 3) pays attention to T waves while also in V5 and V6 (view 3) leads, and the QRS wave group is also noticed along with the T wave in the first few beats. The features extracted by different view networks are diverse, and the fusion of multiple views can better classify the ECG.

## VI. Conclusion

In this paper, we propose a novel deep neural network based on multi-view learning to solve the problem of multi-label ECG classification. In our network, multi-scale convolutional blocks and coordinate attention modules are integrated to obtain high quality ECG features. The network uses a multi-view approach to obtain the diversity of ECG features in different leads, and then the final ECG features are obtained by an effective multi-view fusion approach. Since our network contains six view networks, to reduce the size of our network, we try to compress our network using knowledge distillation. Compared with the multi-view network, the distilled single-view network has fewer parameters and better performance.

In future work, we will further explore the specific reasons for the network performance improvement due to knowledge distillation in this paper, and how to address the classification of ECG datasets with noise, multiple labels, and data imbalance problems more effectively. Most actual ECG data are noisy due to medical hardware devices. Certain arrhythmia cases are rare, which will cause imbalance problems, and some cases have different arrhythmia problems themselves. In our experiments, we found that the imbalance problem in ECG datasets seriously affects the performance of our network and becomes a crossing problem when the multi-label problem and the imbalance problem occur in a single ECG dataset at the same time, which is more difficult to solve. We hope that solving these problems can further improve network performance.

## Appendix
## Analysis of Each View and Each Class

As seen from Tables XI and XII, the performance of the network for each view is worse for each classification task of the three datasets compared to our multi-view network and single-view network with 12 input channels in Tables IV and V. Table I shows that each view corresponds to a different channel. This means that each view utilizes only a portion of the 12-lead ECG and does not utilize the full information of the 12-lead ECG. This is the reason why each view has poor network performance.

TABLE XI
The macro-AUC Scores for Each View on Each of the Six Classification Tasks in the PTB-XL Dataset (Unit: %)

| view | all | diag. | sub-diag. | super-diag. | form | rhythm |
|---|---|---|---|---|---|---|
| 1 | 87.67 | 86.96 | 87.96 | 86.25 | 79.80 | 89.42 |
| 2 | 89.01 | **90.58** | **91.18** | 89.02 | 81.15 | 94.45 |
| 3 | 88.22 | 88.79 | 86.88 | 85.44 | 79.56 | 89.11 |
| 4 | 88.18 | 88.54 | 87.39 | 86.16 | 81.18 | **95.48** |
| 5 | 88.81 | 88.05 | 85.86 | 87.48 | **82.68** | 88.87 |
| 6 | **89.54** | 88.78 | 90.83 | **89.20** | 78.79 | 74.55 |

TABLE XII
The macro-AUC Scores for Each View on the CPSC 2018 Dataset and HFHC Dataset

| dataset | view1 | view2 | view3 | view4 | view5 | view6 |
|---|---|---|---|---|---|---|
| CPSC | 94.86 | 93.52 | 93.95 | 93.93 | 94.13 | **94.88** |
| HFHC | 88.99 | 90.98 | 90.90 | **90.99** | 90.27 | 89.59 |

TABLE XIII
The macro-AUC Scores of Our Multi-View Network for Each Class on the Superclass Diagnostic (super-diag.) Classification Task in the PTB-XL Dataset (Unit: %)

| Superclass | # Records | Description | AUC |
|---|---|---|---|
| NORM | 9528 | Normal ECG | **94.89** |
| CD | 4907 | Conduction Disturbance | 92.87 |
| MI | 5486 | Myocardial Infarction | 93.15 |
| HYP | 2655 | Hypertrophy | 90.90 |
| STTC | 5250 | ST/T Change | 93.62 |

TABLE XIV
The macro-AUC Scores of Our Multi-View Network for Each Class on the CPSC 2018 Dataset (Unit: %)

| Type | # Records | Description | AUC |
|---|---|---|---|
| NORM | 918 | Normal ECG | 95.32 |
| AF | 1098 | Atrial Fibrillation | 98.28 |
| I-AVB | 704 | First-degree Atrioventricular Block | 98.52 |
| LBBB | 207 | Left Bundle Brunch Block | **99.86** |
| RBBB | 1695 | Right Bundle Brunch Block | 97.49 |
| PAC | 574 | Premature Atrial Contraction | 85.44 |
| PVC | 653 | Premature Ventricular Contraction | 95.34 |
| STD | 826 | ST-segment Depression | 95.57 |
| STE | 202 | ST-segment Elevated | 92.05 |

Since some classification tasks in the three datasets we used have more classes (for example, 34 classes in the HFHC dataset, 71 classes in the PTB-XL dataset for all ECG classifications, etc.), we select the classification tasks with fewer classes to save space. Table XIII shows the macro-AUC of our multi-view network for each class on the superclass diagnostic classification task on the PTB-XL dataset. Table XIV shows the macro-AUC scores of our multi-view network for each class on the CSPC 2018 dataset.

## References

[1] G. A. Mensah, G. A. Roth, and V. Fuster, “The global burden of cardiovascular diseases and risk factors,” *J. Am Coll. Cardiol.*, vol. 74, no. 20, pp. 2529–2532, Oct. 2019.

[2] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, “Artificial intelligence-enhanced electrocardiography in cardiovascular disease management,” *Nature Rev. Cardiol.*, vol. 18, pp. 465–478, 2021.

[3] G. Goovaerts, S. Padhy, B. Vandenberk, C. Varon, R. Willems, and S. V. Huffel, "A machine-learning approach for detection and quantification of QRS fragmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 5, pp. 1980–1989, Sep. 2019.

[4] A. Mincholé and B. Rodriguez, "Artificial intelligence for the electrocardiogram," *Nature Med.*, vol. 25, pp. 22–23, Jan. 2019.

[5] Ö. P. Yıldırım, R. S. Pławiak Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Comput. Biol. Med.*, vol. 102, pp. 411–420, Nov. 2018.

[6] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019.

[7] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, pp. 65–69, Jan. 2019.

[8] B. Hou, J. Yang, P. Wang, and R. Yan, "LSTM-based auto-encoder model for ECG arrhythmias classification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1232–1240, Apr. 2020.

[9] W. Liu, F. Wang, Q. Huang, S. Chang, H. Wang, and J. He, "MFB-CBRNN: A hybrid network for MI detection using 12-Lead ECGs," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 503–514, Feb. 2020.

[10] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing transformer model with temporal features for ECG heartbeat classification," in *Proc. IEEE Int. Conf. Bioinform. Biomed.*, 2019, pp. 898–905.

[11] T. Chen, C. Huang, E. Shih, Y. Hu, and M. Hwang, "Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model," *iScience*, vol. 23, no. 3, Mar. 2020, Art. no. 100886.

[12] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network," *Inf. Fusion*, vol. 53, pp. 174–182, Jan. 2020.

[13] J. Wang et al., "Automated ECG classification using a non-local convolutional block attention module," *Comput. Meth. Prog. Biomed.*, vol. 203, May 2021, Art. no. 106006.

[14] M. Sajjan and E. V. S. Maben, "ECG leads," in *Learn ECG in a Day: A Systematic Approach*, 1st ed. London, U.K.: Jp Medical Ltd, 2013, ch. 4, pp. 8–10.

[15] W. Liu, Q. Huang, S. Chang, H. Wang, and J. He, "Multiple-feature-branch convolutional neural network for myocardial infarction diagnosis using electrocardiogram," *Biomed. Signal. Proces.*, vol. 45, pp. 22–32, Aug. 2018.

[16] I. Jekova, G. Bortolan, and I. Christov, "Assessment and comparison of different methods for heartbeat classification," *Med. Eng. Phys.*, vol. 30, no. 2, pp. 248–257, Mar. 2008.

[17] T. Ince*, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ECG signals," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 5, pp. 1415–1426, May 2009.

[18] R. V. Andreao, B. Dorizzi, and J. Boudy, "ECG signal analysis through hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 8, pp. 1541–1549, Aug. 2006.

[19] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomed. Signal. Process.Control*, vol. 8, no. 5, pp. 437–448, Sep. 2013.

[20] S. Karimifard, A. Ahmadian, M. Khoshnevisan, and M. S. Nambakhsh, "Morphological heart arrhythmia detection using hermitian basis functions and kNN classifier," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2006, pp. 1367–1370.

[21] R. Ghongade and A. A. Ghatol, "Performance analysis of feature extraction schemes for artificial neural network based ECG classification," in *Proc. IEEE Int. Conf. Comput. Intell. Multimedia Appl.*, 2007, pp. 486–490.

[22] D. Nagal and S. Sharma, "Simultaneous 12-lead QRS detection by k-means clustering algorithm," in *Proc. IEEE Int. Conf. Recent Adv. Innov. Eng.*, 2014, pp. 1–4.

[23] P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "A machine learning approach for the classification of cardiac arrhythmia," in *Proc. IEEE Int. Conf. Comput. Methodol. Commun.*, 2017, pp. 603–607.

[24] H. Lassoued and R. Ketata, "ECG multi-class classification using neural network as machine learning model," in *Proc. IEEE Int. Conf. Adv. Syst. Emergent Technol.*, 2018, pp. 473–478.

[25] V. Sree et al., "A novel machine learning framework for automated detection of arrhythmias in ECG segments," *J. Ambient Intell. Human. Comput.*, vol. 12, pp. 10145–10162, Jan. 2021.

[26] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol.*, vol. 20, no. 3, pp. 45–50, Jun. 2001.

[27] S. Nils, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, May 2021.

[28] P. Xiong et al., "Localization of myocardial infarction with multi-lead ECG based on DenseNet," *Comput. Meth. Prog. Biomed.*, vol. 203, May 2021, Art. no. 106024.

[29] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

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

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[32] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[33] P. Ramachandran, B. Zoph, and Q. V. Le and, "Searching for activation functions," 2017,*arXiv: 1710.05941*. [Online]. Available: https://arxiv.org/abs/1710.05941

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Conf. Neural Inf. Process. Syst. Deep Learn. Workshop*, 2015.

[35] H. Li, "Exploring knowledge distillation of deep neural networks for efficient hardware solutions," 2018. [Online]. Available: http://cs230.stanford.edu/files_winter_2018/projects/6940224.pdf

[36] P. Wagner, "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 154.

[37] F. Liu et al., "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *J. Med. Imag. Health Informat.*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018.

[38] Hefei high-tech cup dataset, 2019. [Online]. Available: https://tianchi.aliyun.com/competition/entrance/231754/information

[39] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. IEEE Int. Joint. Conf. Neural Netw.*, 2017, pp. 1578–1585.

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

[41] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proc. Pac. Asia Conf. Lang., Inf. Comput.*, 2015, pp. 73–78.

[42] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/pdf?id=YicbFdNTTy

[43] H. I. Fawaz et al., "InceptionTime: Finding AlexNet for time series classification.," *Data Min. Knowl. Disc.*, vol. 34, pp. 1936–1962, Sep. 2020.

[44] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 558–567.

[45] A. Dempster, D. F. Schmidt, and G. I. Webb, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2021, pp. 248–257.

[46] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[47] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[48] P. B. Andrew, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Nov. 1997.

[49] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[50] K. Jiang, S. Liang, L. Meng, Y. Zhang, P. Wang, and W. Wang, "A two-level attention-based sequence-to-sequence model for accurate inter-patient arrhythmia detection," in *Proc. IEEE Int. Conf. Bioinform. Biomed.*, 2020, pp. 1029–1033.

[51] X. Zhang, Y. Gao, J. Lin, and C. Lu, "TapNet: Multivariate time series classification with attentional prototypical network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 04, pp. 6845–6852, Apr. 2020.

[52] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12692–12702.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

**Shunxiang Yang** is currently working toward the M.S. degree with the School of Automation, Wuhan University of Technology, Wuhan, China. His research interests include machine learning and data mining.

**Cheng Lian** (Member, IEEE) received the B.S. degree in electrical engineering and automation, and the M.S. degree in control science and engineering from the School of Automation, Wuhan University of Technology, Wuhan, China, in 2008 and 2011, respectively, and the Ph.D. degree in control science and engineering, from the School of Automation, Huazhong University of Science and Technology, Wuhan, in 2014. He is currently a Professor with the School of Automation, Wuhan University of Technology. His research interests include machine learning, data mining, and pattern recognition.

**Zhigang Zeng** (Fellow, IEEE) received the Ph.D. degree in systems analysis and integration from the Huazhong University of Science and Technology, Wuhan, China, in 2003. He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, and also with the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, Wuhan. He has authored or coauthored more than 200 international journal papers. His research interests include theory of functional differential equations and differential equations with discontinuous right-hand sides, and their applications to dynamics of neural networks, memristive systems, and control systems. He has been an Associate Editor for the IEEE Transactions on Neural Networks (2010–2011), IEEE Transactions on Cybernetics (since 2014), IEEE Transactions on Fuzzy Systems (since 2016), and a Member of the Editorial Board of Neural Networks (since 2012), Cognitive Computation (since 2010), Applied Soft Computing (since 2013).

**Bingrong Xu** received the B.S. degree from the School of Automation, Wuhan University of Technology, Wuhan, China, in 2015, and the Ph.D. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, in 2021. She is currently an Associate Professor with the School of Automation, Wuhan University of Technology, Wuhan. Her research interests include zero-shot learning, transfer learning, sparse representation and low-rank representation.

**Junbin Zang** is currently working toward the Ph.D. degree with the School of instrument and Electronics, North University of China, Taiyuan, China. His research interests include machine learning and MEMS sensors.

**Zhidong Zhang** is currently an Associate Professor with the School of Instrument and Electronics, North University of China, Taiyuan, China. His research interests include testing and tensor technology, machine learning.