Chinese Journal of Electronics
Vol.32, No.1, Jan. 2023

Cross Modal Adaptive Few-Shot Learning Based on Task Dependence

DAI Leichao, FENG Lin, SHANG Xinglin, and SU Han

(*School of Computer Science, Sichuan Normal University, Chengdu 610101, China*)

Abstract — Few-shot learning (FSL) is a new machine learning method that applies the prior knowledge from some different domains tasks. The existing FSL models of metric-based learning have some drawbacks, such as the extracted features cannot reflect the true data distribution and the generalization ability is weak. In order to solve the problem in the present, we developed a model named cross modal adaptive few-shot learning based on task dependence (COOPERATE for short). A feature extraction and task representation method based on task condition network and auxiliary co-training is proposed. Semantic representation is added to each task by combining both visual and textual features. The measurement scale is adjusted to change the property of parameter update of the algorithm. The experimental results show that the COOPERATE has the better performance comparing with all approaches of the monomode and modal alignment FSL.

Key words — Meta-learning, Few-shot learning, Metric learning, Cross modal.

I. Introduction

Deep learning has made great achievements in many applications such as natural language processing, speech recognition, and graphics and image processing. However, it needs to learn a good model from a large amount of labeled data to ensure the efficiency of deep learning. Aiming to solve the problems that there are few labeled data in machine learning, few-shot learning (FSL) is introduced [1].

Based on meta-learning, few-shot learning can be generalized to new tasks quickly in recent progress by using priori knowledge. The model is divided into different episodes during meta-training, generating specific classifiers on each scenario. At test time it gives the machine the ability to learn based on the effects generated by new tasks that do not overlap with those at training time. However, when there is insufficient data in a single modality, another shape would assist the model in training and produce better classification (e.g., people can rely on limited visual information to learn new concepts in a small amount of information and can combine visual, auditory, contextual and other knowledge to achieve good discrimination in the real world). This suggests that semantic features are useful information for pattern recognition [2].

In traditional unimodal meta-learning, the model can be meta-trained with only a little fine-tuning to accomplish the generalization to new tasks. The most representative algorithm is the model-agnostic meta-learning (MAML) [3] method. The advantage of this algorithm is that it has less iterative steps, good generalization performance, no need to worry about the form of the model or add new parameters, so it can be easily fine-tuned. Mishra [4] adopted simple neural attentive learner (SNAIL) to achieve his objective of meta-learning. The meta-learner LSTM [5] is a meta learning method with direct coding and rapid adaptation. Transductive propagation network (TPN) [6] used a label propagation way to make the model to learn how to from the tag samples to the spread of unmarked samples label. Ren [7] proposed that attention attractor networks would learn to assign a new weight to a new category to balance the basic category performance with that of the new category. The XtarNet, which learns to extract task-adaptive representation, proposed by Sung [8] is an incremental model with few samples, and learns to construct new representations with information features to identify base classes and new classes. In meta-learning, few-shot learning based on measurement is an important learning method.

Manuscript Received Mar. 16, 2021; Accepted Feb. 28, 2022. This work was supported by the National Natural Science Foundation of China (61876158) and Fundamental Research Funds for the Central Universities (2682021ZTPY030).
 DOI:10.23919/cje.2021.00.093

Through measurement, the distance between the same class keeps getting closer and the gap between different classes is as far as possible. Matching network [9] and prototypical networks [10] are the most representative methods. The former is based on the attention mechanism and uses the nearest neighbor framework as the similarity measure. The latter learns a metric space so that the sample query set is close to the prototype of the same category and away from the different classes. Xiang's conditional class-aware meta-learning (CAML) [11] combines the advantages of MAML and prototypical network to perform conditional conversion embedding in the metric space. Ren proposed a semi-supervised k-means learning method (prototypical network with soft k-means, ProtoNets W soft k-means) [12], which uses unmarked samples to improve the prototypical network. The relation network proposed by Sung *et al.* [13] introduces a separate relation module, which can be learned and modified in the training process to measure similarity. Yu *et al.* [14] hybrid trained the network and get the better generalization ability and knowledge transfer effect by using the pseudo label data of the target domain and the real label data of the source domain.

The image classification algorithms using auxiliary modal assist model are mainly used in zero-shot learning (ZSL). Moreover, the existing methods mainly rely on mode alignment, which aligns two modes of the same class of samples to obtain the same semantic structure. Andrea put forward of the deep visual-semantic embedding model (DeViSE) [15]. It uses the representation space to map the visual representation space to the semantic representation space. The model uses the semantic information obtained from marked image data and unmarked text to recognize visual objects. Yao's robust semi-supervised visual-semantic embeddings model (ReViSE) [16] minimized the maximum average difference between the two spatial distributions in terms of distribution representation alignment, to align the model. Edgar proposed cross-and distribution aligned variational autoencoders (CADA-VAE) [17]. It uses two VAEs to embed two types of modal information, shares image and semantic information in hidden space, and generates hidden features for invisible classes.

In the unimodal meta-learning, the information of this modality is often local and tends to make the model fall into overfitting with few visually supported images. The semantic representation information can be used as a priori knowledge and context to assist model training. However, the traditional approach of using auxiliary modalities to help model training is generally applied in the field of ZSL. Modal alignment is most common method, and mostly utilized to force the two modes to map together. Therefore, a cross modal adaptive few-shot learning based on task dependence (COOPERATE for short) was proposed. Firstly, a task cluster was adopted as task representation to reduce the supporting set dimension and parameters. It used the mean of each class of support set samples as a prototype to make the same category task cluster more closely in the representation space. Secondly, the task-specific feature extractor is conditioned to be more task-dependent using task conditioning networks. At the same time, the role between task conditioning and auxiliary co-training reduces the difficulty of tuning both networks together. Thirdly, in terms of modal mixing, the algorithm regards the two modes as two independent sources of knowledge and performs the classification task in an adaptive convex combination. Finally, in the metric space, we use scaling to improve the performance of similarity further. The results showed that COOPERATE adaptively adjusts both modalities' focus according to the merits of different spaces. And the model characterization capability is significantly enhanced based on a task-dependent approach. We have achieved state-of-the-art classification accuracy on two publicly data sets of few-shot image classification.

The main contributions were summarized as follows: i) The metric scaling method is introduced into the entropy loss function to improve the similarity measurement criteria between the output of the training sample and its true category, so as to improve the performance of the Few-Shot learning. ii) A task-dependent network was created to construct task representation based on the support set of each scene to improve feature extraction capability. iii) The auxiliary tasks were uesd to perform collaborative training on the feature extraction network to improve the generalization effect of the model. iv) The cross-modal adaptive few-shot classification mode is used to combine the two modal structures, which proves that it is more advanced than (single-modal and cross-modal) few-shot algorithms in different shot numbers of two data sets.

The structure of this paper is as follows: Section II is the problem definition, Section III is the specific explanation of the model, Section IV uses experiments to verify the performance of the model, and Section V summarizes the full text.

II. Problem Definition

To facilitate the description, the basic concepts related to FSL are defined in this section using the mathematical formalism with reference to several authoritative literatures [3], [9], [10] and [13]. Table 1 is the notation table related to the definition.

Table 1. Notation table related to the definition

Notation	Meaning
D	Data set
X, Y	Matrix
$\boldsymbol{x_i}$	I-th vector
$\{\cdots\}$	Set
$\lvert\cdots\rvert$	The potential of a set
$g(\cdot)$	Information function
C	The number of task categories for few-shot classification
K	The number of support set instances in each class
T	Few-shot training tasks
R	Few-shot testing tasks
q	The number of instances of query set for training task with few-shot learning
p	The number of instances of query set for testing task with Few-Shot learning
$Tasks = \{(\cdot,\cdot)\}$	Training task set
F^*	Classification function
l_k	The loss function for a single task
$L(\cdot)$	Loss function

Definition 1 (Few-shot dataset) Let D the data set be a two-tuple $D = \{X, Y\}$, where X is the input space. It consists of $|X|$ input instances $\{x_1, x_2, ..., x_{|X|}\}$. In this paper, $\forall x_i \in X$, x_i represent input image instances. Y is the input space. It consists of $|Y|$ input instances $\{y_1, y_2, ..., y_{|Y|}\}$. $g : X \to Y$ is an information function that specifies the category tag value for each input instance in X, i.e., $\forall x_i \in X, \exists y_j \in Y$, so $g(x_i) = y_j$. $\forall y_j \in Y$, if $g^{-1}(y_j) = \{x_i \in X | g(x_i) = y_j\}$, $g^{-1}(y_j)$ is called the set of instances of class label y_j.

Specifically, if $|Y| = C$ and $|g^{-1}(y_j)| = K$, D is called C way K shot Few-Shot data set when we obtain a small K. Where $|\cdot|$ is the potential of the set, $i = 1, 2, ..., |X|$, $j = 1, 2, ..., |Y|$.

In few-shot learning, multiple C-way K-shot few-shot sets need to be sampled on a large source domain data set $D_s = (X_s, Y_s)$ according to a certain method. And then the classification model is trained on these sets. The classification algorithm is transferred to target domain data $D_t = (X_t, Y_t)$, $D_s = (X_s, Y_s)$ and $D_t = (X_t, Y_t)$ need to meet $Y_s \cap Y_t = \emptyset$.

Definition 2 (Training task support and query set) Given$D_s = (X_s, Y_s), S_{tr} = (X_S, Y_S)$ and$Q_{tr} = (X_Q, Y_Q)$, where S_{tr} and Q_{tr} are respectively called the training task support set and query set, classes are randomly selected from Y_S, i.e., $\{y_j | j = 1, 2, ..., C\}$. C-way K-shot few-shot training task is defined on $T = (S_{tr}, Q_{tr})$, which meets the following requirements:

1) $Y_s = Y_Q = \{y_j | j = 1, 2, ..., C\}$;
2) $\forall y_m \in Y_S, |g^{-1}(y_m)| = K$;
3) $\forall y_n \in Y_Q, |g^{-1}(y_n)| = q$;
4) $X_s \cap X_Q = \emptyset$.

Definition 3 (Testing task support and query set) Given $D_t = (X_t, Y_t)$, $S_{te} = (X_e, Y_e)$ and $Q_{te} = (X_h, Y_h)$, where S_{te} and Q_{te} are respectively called the testing task support set and query set, C classes are randomly selected from Y_t, i.e., $\{y_j | j = 1,2, ..., C\}$. C-way K-shot few-shot testing task is defined on $R = (S_{te}, Q_{te})$, which meets the following requirements:

1) $Y_e = Y_h = \{y_j | j = 1, 2, ..., C\}$;
2) $\forall y_m \in Y_e, |g^{-1}(y_m)| = K$;
3) $\forall y_n \in Y_h, |g^{-1}(y_n)| = p$;
4) $S_e \cap Q_h = \emptyset$.

Definition 4 (Few-shot learning) Given training task set $T = \{(S_{tr}, Q_{tr})\}$ and testing task set $R = \{(S_{te}, Q_{te})\}$. The C-way K-shot FSL task aim to learn a classification function F^* on the data of multiple training tasks, and to learn a classification function $f^* = F^*(S_{te})$ on the support set S_{te} of the testing task, so that f^* can complete the classification of the query set in the test task. The learning process is as follows:

1) Training: A set of learning classification functions F and loss function L is defined. For each training task $k \in T$, the support set S_{tr} was used to generate $f_k = F(S_{tr})$, and loss l_k of f_k was calculated on the query set Q_{tr}. Then the loss function $L(F) = \Sigma_{k=1}^{|Tasks|} l_k$ on the training task followed. By minimizing $L(F)$, a classification model $F^* = \arg\min_F L(F)$ on the training task set is generated.

2) Testing: For the testing task, the classification model $f^* = F^*(S_{te})$ is generated by using the supporting set S_{te}. And then the model evaluation the f^* by

using query set Q_{te}.

In particular, if the adjustment of model super parameters is involved in the process of training F^*, the training task set can also be divided into training task set and verification task set.

For the definition of few-shot in Definition 4, there are two special cases: If $K = 1$, this task is called one-shot learning. Only the single picture of one class is given as the support set for each training. And the remaining pictures are used as the query set. Another is zero-shot learning. ZSL does not mean that training samples are not required at all. Instead, it is studied to train the model with training set samples and corresponding auxiliary text description and attribute feature information of samples when specific training cases are missing.

III. Model

1. Model architecture

The architecture of the COOPERATE is shown in Fig.1. In the figure, part A is the metric scaling and relationship module, B is the task condition module, C is the auxiliary collaborative training module, and part D is the semantic modal embedded module. Sections II–VI are detailed descriptions of each module of the model, and Section VII is the training strategy of the model.

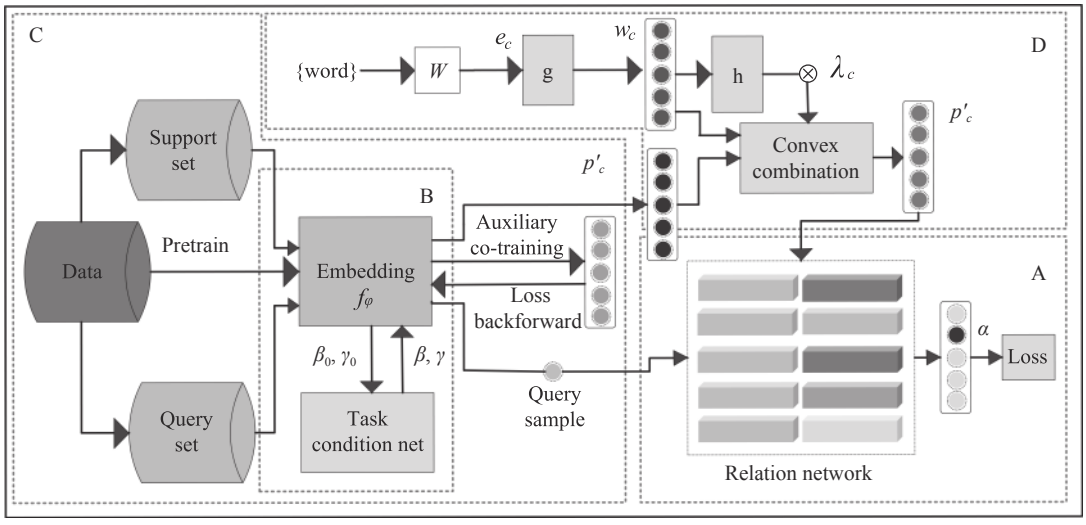

Fig. 1. The model architecture of COOPERATE.

2. Metric scaling

This section explains the metric scaling concept in part A of Fig.1. Scaling of the metric has been shown to be beneficial for the similarity measurement in entropy loss functions [18]. Define a super parameter α, softmax based metrics are represented as $p_{\varphi,\phi,\alpha}(y = k|X_Q) = softmax(-\alpha d_\phi(f_\varphi(X_Q), c_k))$. The model updates the embedded module and relational module parameters through the overall loss. As proved by [18], for small α value, the parameter will reduce the distance between the query sample and the corresponding prototype. Meanwhile, it also maximizes the distance between the query sample and the non-attributable category prototype. Large value of α (which are the same for the former) will maximize the distance of sample with the closest wrongly assigned prototype (if any). The latter is equivalent to learning only from the hardest examples. Thus, the two different modes of α are conducive to minimizing the overlap of sample distributions or correcting cluster attribution on the case.

The larger α is more directly related to the error correcting for the few-shot learning. However, with the improvement of optimization and classification accuracy, the number of samples of misclassification will decrease. As a result, the average effective batch size is reduced and more samples generate zero derivatives in this updating mode. Therefore, for a given data set, task, and metric, there is an optimal value for α. Section IV.3 proves the existence of the optimal value through experiments (i.e. scaling effect).

3. Prototypes and relational networks

In order to understand part A of Fig.1 , this section explains the principles of prototypes and relational networks. COOPERATE is based on meta-learning of metric. For each episode, the prototype network calculates the prototype for each category using the support set. Query samples are classified according to the distance to each prototype. The network computes each class M-dimensional prototype $c_k \in \mathbb{R}^M$ by embedding functions $f_\varphi : \mathbb{R}^D \to \mathbb{R}^M$, and parameters that can be updated by learning φ. The prototype is represented by the mean value of the embedded feature vectors of each class, which not only reduces the input dimension, but also is an effective way of task clustering, making the

task representation space clustering of the same class more compact. The prototype is calculated as (1).

$$c_k = \frac{1}{|S_k|} \sum (X_i, y_i) \in S_k f_\varphi(x_i) \tag{1}$$

Relational network [13] is a deep convolutional neural network model, whose structure includes embedded modules and relational modules. The embedded module is used to extract features, and the relational module serves to measure the image similarity. Different from [13], COOPERATE modified the embedded module as Resnet-12, which is a popular and widely used feature extraction network recently and has been proved effective in experiments. The relational module adopts a two-layer fully connected network and carries out L_2 regularization. ReLU as the activation function after the hidden layer. It is followed by a layer of dropout to reduce overfitting, with the drop rate parameter set to 0.3.

Relational score: X_S and X_Q are located in support set S_{tr} and query set Q_{tr}, respectively. Input them into the embedded module f_φ to form the feature matrix $f_\varphi(X_S)$ and $f_\varphi(X_Q)$ after mapping. The support set prototype c_k is calculated by (1). And then the model combines c_k and $f_\varphi(X_Q)$ with the operator $C(c_k, f_\varphi(X_Q))$. $C(\cdot,\cdot)$ is series in rows. The combination operator is input into the relation module g_φ to calculate the similarity of features. We represent it as a scalar from 0 to 1 (i.e. a relational score). The relational score $\eta_{k,Q}$ can be obtained by querying the set sample X_Q and the prototype c_k. The relational score is calculated as (2).

$$\eta_{k,Q} = g_\varphi(C(c_k, f_\varphi(X_Q))), k = 1, 2, ..., C \tag{2}$$

The mean square error loss in the model of [13] was adjusted to be the cross-entropy loss that is now widely used. Query sample X_Q to the softmax distance d of the embedded prototype to generate a distribution as shown in (3).

$$p(y = k|X_Q, S_{tr}, \varphi, \phi) = \frac{\exp(d(X_Q, c_k))}{\sum_n \exp(d(X_Q, c_n))} \tag{3}$$

where, $d(X_Q, c_k) = \eta_{k,Q}$ is the measurement of relational distance. The loss is calculated as the negative logarithmic likelihood of each query sample to the ground true value. The loss function is shown in (4).

$$L(\varphi, \phi) = \underset{(S_{tr}, Q_{tr})}{E} - \sum_{Q=1}^{Q_{tr}} \log p_{\varphi,\phi}(Y_Q|X_Q, S_{tr}) \tag{4}$$

where $(X_Q, Y_Q) \in Q_{tr}$ and S_{tr} are query set and support set sampled in each episode respectively. φ and ϕ are the model parameter. The model adopts the gradient descent feedback loss optimization network.

4. Task conditioning

Task condition network is located in Part B of Fig.1. It is assumed that the tasks of feature extractor f_φ are independent of each other. We define a dynamic feature extractor $f_\varphi(x, \Gamma)$, x as the sample, and Γ as the parameter set represented by the task. The performance of $f_\varphi(x, \Gamma)$ is optimized under the condition of given task sample D_s. According to FILM conditioning layer [19] and conditional batch normalization, $h_{l+1} = \gamma \odot h_l + \beta$ is defined, where γ and β are scaling and shift vectors of h_l layer. According to Section III.3, the mean value of the class is used as the class prototype, and then the mean value of the class prototype is used as the task representation, i.e. $\bar{c} = \frac{1}{K} \sum_k c_k$. This network is called a task embedded network. The task embedded network is used to predict the layer-level element-wise scale and shift vectors γ, β of each convolutional layer in the feature extractor. Mean value of the prototype as task representation can reduce the dimension of task embedded network, and better clustering effect can be achieved without the establishment of complex work such as CNN, RNN or attention modeling [20].

Following the terminology in [18], task embedded network create two separate fully connected residual networks to generate γ, β. And the number of layers is 3. The first layer represents the task as the target width, and the remaining layers operate according to the target width (each layer containing a jump connection). γ_0, β_0 are multiplied by the task code after the L_2 penalty. The L_2 regularization weights of γ_0 and β_0 are cross-verified at each layer. γ_0, β_0 are the prior experience of γ, β. And the restriction on them is critical for task embedding network. Without them, training tends to sink into local minima, which is bad for overall performance. The mathematical equations are $\gamma = \gamma_0 h\phi(\bar{c}) + 1$, $\beta = \beta_0 g_\theta(\bar{c})$, where, h_φ, g_θ are the predictors of γ, β. In order to maximize its contribution, the task embedded network is injected into the layer before the maximum pool separately and the injection of the last layer is retained according to [18].

5. Auxiliary task co-training

This section explain the auxiliary task co-training network in part C of Fig.1. The insertion of task conditioning layer after feature extraction increases the complexity of model training. This makes the training feature extraction network and task conditioning network too hard. Thus, a separate predictor is used to auxiliary co-training. The auxiliary task uses the loss function alone to adjust the parameters backward. Auxiliary tasks were classified 64-way on the miniImageNet [21] data set and 351-way on the tieredImageNet [12]. The model initial attenuation exponent is chosen to be 0.9, and the number of attenuation steps is 20. In Section

IV.5, it can be seen that auxiliary tasks have a positive effect on model training. The FSL of auxiliary task co-training is related to the total training task volume. That is a simple structure in the whole complex model structure, which can help to intensify extract features.

6. Cross modal

The text embedding module is shown in part D of Fig.1. In few-shot learning, due to the limited samples of the support set, it is difficult to improve the classification effect by relying only on the visual information. Furthermore, if the target domain data is a more granular image [22], such as "dog" and "wolf," it is easy to distinguish with semantic information [2]. In the ZSL, the test task has not support set of image information, and the model relies entirely on semantic information. Few-shot learning is somewhere between zero-shot and traditional learning.

Assume that both visual information and semantic information are effective for FSL. But two modes are different in spatial structure. We hope to train a model to make use of both modal information adaptively. When visual information is insufficient, text information can be relied on to achieve good recognition. On the contrary, it can make full use of image information to participate in training.

The text mode of COOPERATE adopts the language structure word embedded W which is pre-trained in large unsupervised text corpus. It contains labels embedded $D_s \cup D_t$ for all categories. The new prototype is a convex combination of the two representations, i.e., for each category C, the adjusted calculation as shown in (5).

$$p'_c = \lambda_c \cdot p_c + (1 - \lambda_c) \cdot w_c \tag{5}$$

where, λ_c is the adaptive mixing coefficient and p_c is the prototype. $w_c = g(e_c)$ is the label embedding of class c, e_c is the word embedding of W pre-training tag. The mapping transformation $g : \mathbb{R}^{n_w} \to \mathbb{R}^{n_p}$ uses the parameter θ_g to map the two representation spaces to the same dimension space $\mathbb{R}^{n_p}$. The calculation of coefficient T is based on the condition of category, and the calculation formula is shown as (6).

$$\lambda_t = \frac{1}{1 + \exp(-h(w_c))} \tag{6}$$

where, h is adaptive hybrid network, and the parameter is θ_h. At this point, d is the distance between the query sample and the cross-modal prototype p'_c, as shown in (7).

$$p_\theta(y = c | X_Q, S_{tr}, W) = \frac{\exp(d(X_Q, p'_c))}{\sum_k \exp(d(X_Q, p'_k))} \tag{7}$$

where θ is the model parameter set. The model is trained by minimizing (4). The probability is also related to the word embedded W.

7. Training strategy

This section mainly introduces the training strategy of each episode loss of the COOPERATE. The input of the algorithm is the training set, where each episode randomly selects the support and the query set according to Definition 2. The training process is end-to-end training, and finally the model loss is updated. The update method of the loss is shown in Algorithm 1, where s is the total number of classes in the training set, C is the number of classes in every episode, K is the number of supports for eachclass, q is the number of queries for each class, and W is the pretrained label embedding dictionary.

Algorithm 1 Training strategy of each episode loss of COOPERATE

Input: Train set $D_s = \{(X_s, Y_s)\}_s, Y_s \in \{1, ..., C\}$

Output: Episodic loss $L(\theta)$.

{select C class for episode e}

$T \leftarrow RandomSample(\{1, 2, ..., |Y_i|\}, C)$

{compute cross modal prototypes}

For i in T do

  For j in $T_{pretrain}$ do

    $image, label \leftarrow RandomSample(D_s, co\_trainBatch)$

    %Auxiliary co-training

    $\varphi \leftarrow co\_trainNet(image, label)$

    Co_training Loss(φ)

  End for

  $S_{tr} \leftarrow RandomSample(D_s, K)$ %Select support set

  $Q_{tr} \leftarrow RandomSample(D_s \setminus S_{tr}, q)$ %Select query set

  $c_i \leftarrow \frac{1}{|S_{tr}|} \sum_{(X_S, Y_S) \in S_{tr}} f_\varphi(X_s)$ %Task condition

  $\beta, \gamma \leftarrow TaskConditionNet(c_i)$

  $c'_i \leftarrow \frac{1}{|S_{tr}|} \sum_{(X_S, Y_S) \in S_{tr}} f_{\varphi,\beta,\gamma}(X_s)$

  $e_i \leftarrow LookUp(i, W)$ %Word embedding

  $w_i \leftarrow g(e_i)$

  $\lambda_i \leftarrow \frac{1}{1+\exp(-h(w_i))}$

  $p_i \leftarrow \lambda_i \cdot c'_i + (1 - \lambda_i) \cdot w_i$ %Convex combination

End for

{compute loss}

For i in T do

  For (X_Q, Y_Q) in Q_{tr} do

    $\eta_{i,Q} \leftarrow g_\phi(C(p_i, f_\varphi(X_Q)))$

    %Calculate the relation score

    $L(\theta) \leftarrow L(\theta) + \frac{1}{C \cdot K}[-\alpha\eta_{i,Q} + \log \sum_k \exp(-\alpha\eta_{k,Q})]$

    %Scaling metric

  End for

End for

IV. Experiments

In order to verify the effect of the work, the experiment is divided into five parts. 1) Comparing COOPERATE with two baselines (the single-modal and the modal alignment FSL model) to analyze the generalization effect. 2) Quantitatively analyzing the impact of scaling in COOPERATE on accuracy. 3) Analyzing the cross-modal adaptive coefficients and influencing factors. 4) The different feature input methods of the adaptive mechanism is explored. 5) Doing ablation study on each module.

Table 2 is the evaluation-related notation parameter table.

Table 2. Evaluation-related notation parameter table

Notation	Meaning
Backbone	Feature extraction network
Test-accuracy	Test accuracy
α	Metric scaling factor
λ	Modal adaptive mixing coefficient
$h(o)$	The original output characteristics of the GloVe model are used for adjustment the λ
$h(v)$	The features of the support set samples are used to adjust the λ
$h(w,q)$	The combination of query sample features and semantic features is used to adjust the λ
$h(w)$	The output of the GloVe model is adjusted and deformed by the network to adjust the λ
ST	The effect of scale transformation on model generalization
AT	The effect of auxiliary collaborative training on model generalization
TC	The effect of task conditioning on model generalization
CM	The effect of cross-modal on model generalization

1. Experimental setup

Dataset The evaluation of the experiment uses two widely used datasets with few-shot learning: miniImageNet [21] and tieredImageNet [12]. The miniImageNet dataset contains 100 random sampling categories. Each class has 600 images in total of 60,000 84×84 images. For comparison, we follow the segmentation method of [13], i.e., divide them into 64 training, 16 validation and 20 testing classes. TieredImageNet is larger than miniImageNet, contains 779,165 images, corresponding indicators are 351, 97 and 160. Each class has more than 1,000 images (more challenging).

Word embeddeding GloVe [23], embedding network trained with large unsupervised text corpora, is used to extract words embedded in category labels of images in FSL data sets. GloVe is an unsupervised learning method based on word-word co-occurrence statistics in a large text corpus. This fixed output of this part is 300. When there are multiple synonym comments for a category, the first one is selected as the category. If the first class does not appear in GloVe's vocabulary, the second category is selected. If a category is not marked in the dictionary, each dimension of the embedding is randomly selected from the uniform distribution of $(-1, 1)$. When the annotation contains more than one word, the average value of them is selected as the embedding.

Baselines In order to better prove the validity, two types of models were chosen for comparison. The first is uni-modality few-shot learning method, such as matching network [9], prototypical network [10], meta-learner LSTM [5], MAML [3], ProtoNets W soft k-means [12], relation network [13], TPN [6], attention attractor networks [7], XtarNet [8], SNAIL [4], and CAML [11]. In these method, relation network is the closest to our work. The second fold of modal aligned multimodal learning methods, respectively is: DeViSE [15], ReViSE [16], CADA-VAE [17]. Among them, CADA-VAE got the best performance in both zero-shot and few-shot learning.

Implementation The experimental configuration uses Linux operating system. We use Python language programming. TensorFlow deep learning framework and NVIDIA Tesla V100 GPU did computing power acceleration for us.

2. Results

Tables 3 and 4 respectively show the comparison of COOPERATE classification accuracy in miniImageNet and tieredImageNet. In these two tables, top of the table are the single mode few-shot models, the middle are modal alignment baselines, and the bottom are COOPERATE and its backbone results. In all test cases, our report was superior to relation network. Focusing on embedding network, the Resnet can improve the ability of model feature extraction w.r.t. classification capacity is better than that of convolutional network. COOPERATE is also built over the backbone of the Resnet. But on miniImageNet dataset, COOPERATE's classification accuracy is 12.28% higher than that of convolutional embedded network for 1-shot tasks and 10.24% better than that of relation network for Resnet backbone. It improves 5.29% and 1.41% on 5-shot mis-

sions, respectively. On the tieredImageNet, the classification accuracy of 1-shot task is 5.02% higher than that of relation network. Performance is improved by 2.46% in a 5-shot scenario. This shows that COOPERATE can effectively improve the performance of the model in metric-based few-shot learning. Through the above analysis people will ask two questions. Because the fewer shots, the better the scene. We speculate that the feature weight of the text mode will decrease as the number of shots increases. When there is less visual information, the model may have better adaptability. These questions will be verified in Section IV.4.

Table 3. Classification accuracy test on miniImageNet data set

Model	Backbone	Test-accuracy	
		5-way 1-shot	5-way 5-shot
Uni-modality few-shot learning baselines			
Matching network [9]	ConvNet	43.56±0.84	55.31±0.73
Prototypical network [10]	ConvNet	49.42±0.78	68.20±0.66
Meta-learner LSTM [5]	ConvNet	43.44±0.77	60.60±0.71
MAML [3]	ConvNet	48.70±1.84	63.11±0.92
ProtoNets W soft k-means [12]	ConvNet	50.41±0.31	69.88±0.20
TPN [6]	ConvNet	55.51±0.86	69.86±0.65
MAML [3]	ResNet	49.61±0.92	65.72±0.77
Matching network [9]	ResNet	52.91±0.88	68.88±0.69
Attention Attractor Networks [7]	ResNet	54.59±0.30	63.04±0.30
XtarNet [8]	ResNet	55.28±0.33	66.86±0.31
SNAIL [4]	ResNet	55.71±0.99	68.80±0.92
CAML [11]	ResNet	59.23±0.99	72.35±0.71
Modality alignment baselines			
DeViSE [15]	–	37.43±0.42	59.82±0.39
ReViSE [16]	–	43.20±0.87	66.53±0.68
CADA-VAE [17]	ResNet	58.92±1.36	**73.46±1.08**
Relation network [13]	ConvNet	50.44±0.82	65.32±0.70
Relation network [13]	ResNet	52.48±0.86	69.83±0.68
Ours	ResNet	**62.72±0.41**	71.24±0.33

Table 4. Classification accuracy test on tieredImageNet data set

Model	Backbone	Test-accuracy	
		5-way 1-shot	5-way 5-shot
Uni-modality few-shot learning baselines			
Prototypical network [10]	ConvNet	53.31±0.89	72.69±0.74
MAML [3]	ConvNet	51.67±1.81	70.30±1.75
ProtoNets W soft k-means [12]	ConvNet	53.31±0.89	72.69±0.74
TPN [6]	ConvNet	59.91±0.94	73.30±0.75
SNAIL [4]	ResNet	55.71±0.99	68.88±0.92
CAML [11]	ResNet	59.23±0.99	72.35±0.71
Attention attractor networks [7]	ResNet	56.11±0.33	65.52±0.31
XtarNet [8]	ResNet	**61.37±0.36**	69.58±0.32
Modality alignment baselines			
DeViSE [15]	–	49.05±0.92	68.27±0.73
ReViSE [16]	–	52.40±0.46	69.92±0.59
CADA-VAE [17]	ResNet	58.92±1.36	73.46±1.08
Relation network [13]	ConvNet	54.48±0.93	71.32±0.78
Ours	ResNet	59.5±0.43	**73.78±0.34**

Compared with the traditional monomodal few-shot learning models, COOPERATE has obvious effect under the 1-shot situation of miniImageNet and the 5-shot situation of tieredImageNet, preformed the best results than current state-of-the-art. The results are even better than the TPN model (using transductive inference). In other scenarios, although the highest accuracy rate is not achieved, the classification effect is still outstanding. Its performance is very close to theirs. Compared to the baseline of modal alignment, CO-

OPERATE is at the highest level except for a slightly lower than CADA-VAE in the 5-shot task of miniImageNet dataset. At the same time, it can be seen from the table that most of the methods of modal alignment are not as good as the current learning method of single modal in few-shot. The possible reason is that when the two modes are aligned, part of the information is lost because the model structure is forced aligned.

3. Scaling effect

In this section, the effect of scale transformation on the accuracy of generalization is studied through experiments, and the experimental results are shown in Fig.2. The curve in Fig.2 reflects the model verification and test accuracy with the change of zoom factor α in part A in Fig.1 (the scaling factor is from 1 to 10 times).

It can be seen from the four broken line graphs that under the same other conditions, the metric parameter α and the model verification and test accuracy rate are inverse U-shaped curves. This verifies the assumption mentioned in Section III.2 that the metric parameters have optimal values. For the miniImageNet, on the 5-way 1-shot task, when the zoom factor is 2, the test accuracy reaches the highest. In the 5-shot task, the highest point of the task is reached when the scaling factor reaches 5. For the tieredImageNet, on the 5-way 1-shot classification task, the classification accuracy is the best when the zoom factor is 5. For 5-shot tasks, this value is 3.

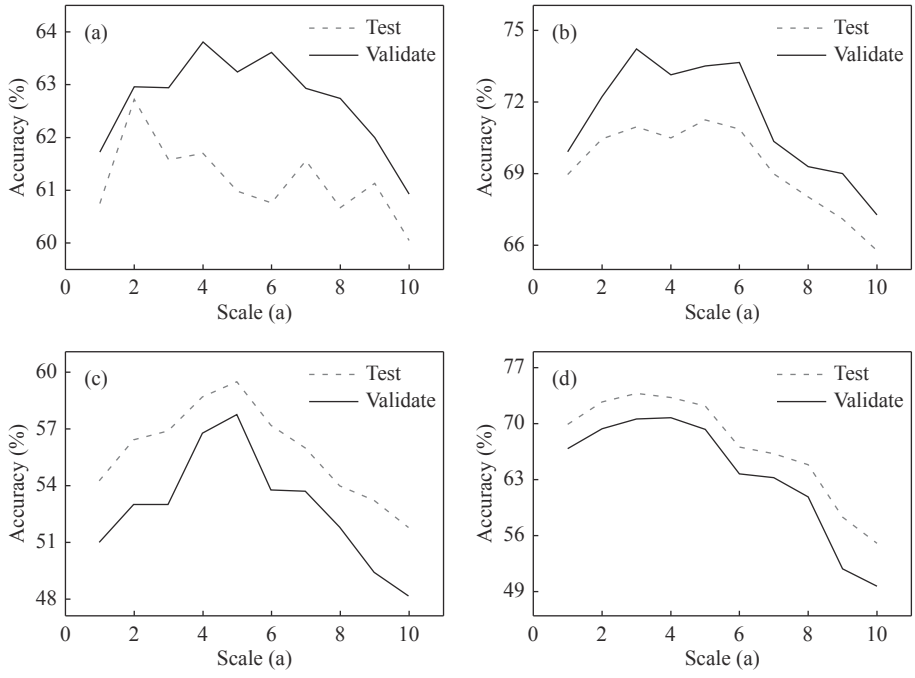

Fig. 2. Results of measurement scale parameters. (a) 5-way 1-shot on miniImageNet; (b) 5-way 5-shot on miniImageNet; (c) 5-way 1-shot on tieredImageNet; (d) 5-way 5-shot on tieredImageNet.

4. Adaptiveness analysis

In order to answer the questions raised in Section IV.2, the relationship between the adaptive coefficient and the number of shots is studied through experiments, and the influence of the adaptive mechanism on the performance improvement of the model is verified. In this section, experiments are designed to quantitatively verify that the COOPERATE adaptive mechanism can effectively and reasonably adjust the two modes.

Fig.3 shows the test accuracy of the model for 1–10 shot tasks on miniImageNet and tieredImageNet. As shown in figure, with the increase of shot number, the accuracy rate increases gradually. This indicates that the support of visual samples is increased and the classification effect of models is improved significantly.

Fig.3 also shows the mean and std. of different shot numbers and the mixing coefficient λ. First of all, it can be seen that the mean value of λ is related to shots. As shots decrease, the amount of visual data decreases, and COOPERATE has a greater weight for text modes and a smaller weight for visual modes. This shows that when visual information is low, the model can automatically adjust the focus to the text mode to help the model classify. Secondly, it can also be observed that the 10-fold variance of λ is correlated with the performance of COOPERATE. The variance of λ decreases as the number of shots increases and the performance improves accordingly. It shows that the algorithm's adaptability at the level of category plays a very important

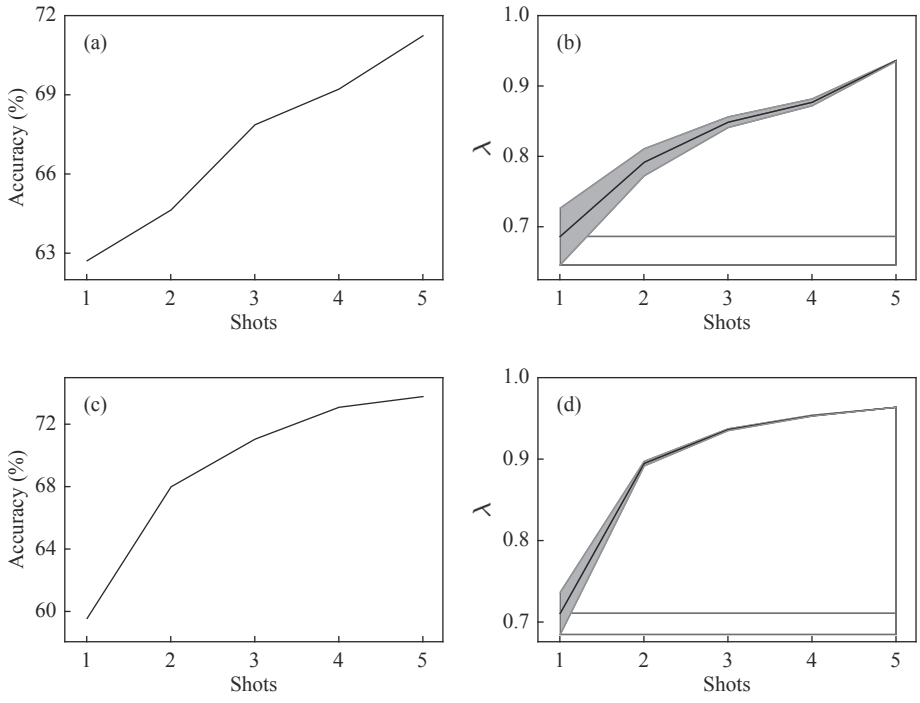

Fig. 3. COOPERATE accuracy and mean value under different shot Numbers. (a) The relationship between accuracy and shots on mini-ImageNet; (b) The relationship between λ and shots on miniImage-Net; (c) The relationship between accuracy and shots on tieredImag-eNet; (d) The relationship between λ and shots on tieredImageNet.

role in improving performance.

5. The study of the input of adaptive mechanism

In order to study the operation effect of the adaptive mechanism under different input features, the ablation study of the adaptive mechanism is carried out through experiments. Table 5 shows the results of 4 different inputs of the adaptive hybrid network h: i) Original input form of GloVe ($h(o)$); ii) Visual input representation of support set samples ($h(v)$); iii) Connection combination of query samples and semantic embedding; and iv) The GloVe deformed semantic input mode ($h(w)$), which is the method adopted by COOPERATE. The result shows in the last row of the table is the best result of the $h(w)$ feature input method. By comparing with the first three rows, it can be seen that the effect of inputting the adaptive mixing network after deforming GloVe is the best. The visual spatial input form in line 2 and the connection method of query samples and semantic embedding in row 3 increase a lot of computational overhead due to the increase of image samples. Under the same hardware and software configuration, the optimization time is also extended. However, it does not exceed the effect of word feature conditions, which shows that semantic space is more suitable for adaptive mechanisms.

Table 5. The impact of different feature input ways of the adaptive mechanism on the model classification accuracy (%)

Method	miniImageNet		tieredImageNet	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
$h(o)$	61.30	70.59	58.8	73.27
$h(v)$	61.12	70.79	59.32	72.81
$h(w,q)$	61.93	70.86	58.36	72.95
$\boldsymbol{h(w)}$	**62.72**	**71.24**	**59.50**	**73.78**

6. Ablation studies

In this section, the effects of scaling, task conditioning, auxiliary co-training, and cross-modal on the generalization are studied through experiments. The results are shown in Table 6.

Firstly, the relationship between task conditioning and auxiliary task co-training (mentioned in Sections III.4 and III.5) is analyzed through experiments. As observed in lines 2 and 3 of the table, when there is no auxiliary cooperative training, task conditioning has little effect on feature extractor. It is difficult for task embedding network to extract features and filter them

Table 6. Influence of each module on model classification accuracy (%)

				miniImageNet		tieredImageNet	
ST	AT	TC	CM	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
	√	√	√	60.75	68.97	54.25	69.94
√		√	√	54.70	57.18	49.30	43.59
√	√		√	61.23	70.29	59.48	73.48
√	√	√		56.78	70.49	57.15	73.19

Note: ST: Scaling transformation; AT: Auxiliary co-training; TC: Task conditioning; CM: Cross modal.

effectively at the same time, which is easy to fall into local extremum. This problem can be greatly improved through collaborative training of auxiliary tasks. From the experimental results, it can be concluded that the contribution of the auxiliary collaborative training to the model is outstanding mainly for the following two reasons: i) the initial convergence value can be obtained by pre-training, and ii) the forced feature extractor performs well on both separate tasks, thus achieving good performance on the Few-Shot learning task. The coupling of the auxiliary task and the main task enables the task regulation to give full play.

Secondly, combined line 1 of the table with Section IV.3, it can be seen that task regulation plays a great role in COOPERATE. This is also of reference significance in other relevant researches in this field.

Finally, it is analyzed from the cross-modal perspective. As it be seen from line 4 of Table 6, semantic information play well in COOPERATE, especially for the 1-shot task. At this time, the image support sample feature is insufficient, which is compensated by the addition of semantic feature. Comparing the two data sets, it can be seen that semantic modes play a greater role on smaller data sets such as miniImageNet. This not only reflects the difficulty of tieredImageNet data set classification, but also indicates that the addition of auxiliary modes should consider the mode scale and model type. Because whether semantic modal features are useful for visual modal features is related to the selected word embedding model as well as the size of text corpus.

V. Conclusions

In this paper, a new method which can map adjustment to adapt to different task representations and use cross-modal information adaptively and effectively for the classification of few-shot learning is proposed. The effects of similarity metric scaling, auxiliary co-training, task conditioning and cross-modal information on the model are quantitatively analyzed. It was proved that scale factor plays an active role in parameter updating of similarity measure. It also be verified that the method of task conditioning representation can improve the performance of feature extractor in few-shot tasks, so as to design a more powerful task representation. In addition, by using unsupervised textual data, COOPERATE is greatly improved in terms of the classification of few-shot learning. When visual data is insufficient (such as one-shot), the semantic features of text have obvious effect on model classification. Moreover, quantitative experiments shown that our algorithm coule reasonably and effectively adjust the concerns of the two modes. In the future, we will consider using few-shot learning to realize automatic labeling.

References

[1] F. F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transaction on Pattern Analysis And Machine Intelligence*, vol.28, no.4, pp.594–611, 2006.

[2] C. Xing, N. Rostamzadeh, B. Oreshkin, *et al.*, "Adaptive cross-modal few-shot learning," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS'19)*, Vancouver City, Canada, pp.4847–4857, 2019.

[3] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, pp.1126–1135, 2017.

[4] M. Nikhil, R. Mostafa, X. Chen, *et al.*, "A simple neural attentive meta-learner," *arXiv preprint*, arXiv: 1707.03141, 2018.

[5] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," *The 5th International Conference on Learning Representations (ICLR 2017 Oral)*, Toulon, France, *https://openreview. net/forum?id=rJY0-Kcll*, 2017.

[6] Y. Liu, J. Lee, M. Park, *et al.*, "Learning to propagate labels: Transductive propagation network for few-shot learning," *arXiv preprint*, arXiv: 1805.10002, 2019.

[7] M. Ren, R. Liao, E. Fetaya, *et al.*, "Incremental few-shot learning with attention attractor networks," in *Proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver City, Canada, pp.5276–5286, 2019.

[8] W. Y. Sung, D. Y. Kim, S. Jun, *et al.*, "XtarNet: Learning to extract task-adaptive representation for incremental few-shot learning," *arXiv preprint* , arXiv: 2003.08561 , 2020.

[9] V. OriolL, B. Charles, L. Timothy, *et al.*, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*, Barcelona, Spain, pp.3630–3638, 2016.

[10] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, pp.4077–4087, 2017.

[11] X. Jiang, M. Havaei, F. Varno, *et al.*, "Learning to learn with conditional class dependencies," *The 7th International Conference on Learning Representations (ICLR (Poster) 2019)*, New Orleans, Louisiana, USA, *https://openreview.net/forum?id=BJfOXnActQ*, 2019.

[12] M. Ren, E. Triantafillou, S. Ravi, *et al.*, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint*, arXiv: 1803.00676, 2018.

[13] F. Sung, Y. Yang, L. Zhang, *et al.*, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, pp.1199–1208, 2018.

[14] Y. Yu, L. Feng, G. G. Wang, *et al.*, "A few-shot learning model based on semi-supervised with pseudo label," *Acta Electronica Sinica*, vol.47, no.11, pp.2284–2291, 2019.

[15] A. Frome, G. S. Corrado, J. Shlens, *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proceedings of the 27th Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, Lake Tahoe, USA, pp.2121–2129, 2013.

[16] T. Y. H. Hubert, L. K. Huang, and R. Salakhutdinov, "Learning robust visual-semantic embeddings," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.3571–3580, 2017.

[17] E. Schonfeld, S. Ebrahimi, S. Sinha, *et al.*, "Generalized zero-and few-shot learning via aligned variational autoencoders," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA, pp.8247–8255, 2019.

[18] B. Oreshkin, L. P. Rodriguez, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, pp.719–729, 2018.

[19] E. Perez, F. Strub, V. H. De, *et al.*, "FiLM: Visual reasoning with a general conditioning layer", *arXiv preprint*, arXiv: 1709.07871, 2018.

[20] F. Lyu, L. Y. Li, S. S. Victor, *et al.*, "Multi-label image classification via coarse-to-fine attention," *Chinese Journal of Electronics*, vol.28, no.6, pp.1118–1126, 2019.

[21] O. Vinyals, C. Blundell, T. Lillicrap, *et al.*, "Matching networks for one shot learning," in *Proceedings of the 30th Neural Information Processing Systems (NeurIPS)*, Barcelona, Spain, pp.3630–3638, 2016.

[22] X. S. Wang, Y. R. Li, and Y. H. Cheng, "Hyperspectral image classification based on unsupervised heterogeneous domain adaptation cycleGan," *Chinese Journal of Electronics*, vol.29, no.4, pp.608–614, 2020.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp.1532–1543, 2014.

DAI Leichao was born in 1994. He received the B.S. degree in engineering from Sichuan Normal University, China, in 2018 and is currently pursuing the M.S. degree in software engineering at Sichuan Normal University. His main research interests include computer vision and pattern recognition.
(Email: daileichao@gmail.com)

FENG Lin (corresponding author) received the Ph.D. degree from Southwest Jiaotong University, China. He is a Professor of School of Computer Science, Sichuan Normal University. His research interests include machine learning and data mining.
(Email: fenglin@sicnu.edu.cn)

SHANG Xinglin was born in 1995. She received the B.S. degree in engineering from Sichuan Normal University, China, in 2019 and is currently pursuing the M.S. degree in software engineering at Sichuan Normal University. Her main research interest includes machine learning.
(Email: 1250919363@qq.com)

SU Han received the Ph.D. degree from Harbin Engineering University. She is a Professor of School of Computer Science, Sichuan Normal University. Her research interests include pattern recognition and image processing.
(Email: jkxy_sh@sicnu.edu.cn)