SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

June 2024, Vol. 67, Iss. 6, 162204:1–162204:16
https://doi.org/10.1007/s11432-023-3924-x

Logit prototype learning with active multimodal representation for robust open-set recognition

Yimin FU, Zhunga LIU* & Zicheng WANG

School of Automation, Northwestern Polytechnical University, Xi'an 710072, *China*

Received 31 May 2023/Revised 30 August 2023/Accepted 18 November 2023/Published online 17 May 2024

Abstract Robust open-set recognition (OSR) performance has become a prerequisite for pattern recognition systems in real-world applications. However, the existing OSR methods are primarily implemented on the basis of single-modal perception, and their performance is limited when single-modal data fail to provide sufficient descriptions of the objects. Although multimodal data can provide more comprehensive information than single-modal data, the learning of decision boundaries can be affected by the feature representation gap between different modalities. To effectively integrate multimodal data for robust OSR performance, we propose logit prototype learning (LPL) with active multimodal representation. In LPL, the input multimodal data are transformed into the logit space, enabling a direct exploration of intermodal correlations without the impact of scale inconsistency. Then, the fusion weights of each modality are determined using an entropy-based uncertainty estimation method. This approach realizes adaptive adjustment of the fusion strategy to provide comprehensive descriptions in the presence of external disturbances. Moreover, the single-modal and multimodal representations are jointly optimized interactively to learn discriminative decision boundaries. Finally, a stepwise recognition rule is employed to reduce the misclassification risk and facilitate the distinction between known and unknown classes. Extensive experiments on three multimodal datasets have been done to demonstrate the effectiveness of the proposed method.

Keywords logit prototype learning, multimodal perception, open-set recognition, uncertainty estimation

1 Introduction

Pattern recognition systems inevitably encounter unknown class objects outside the training data in real-world applications [1]. However, traditional recognition methods often operate under an impractical closed-set setting, assuming that the training and test data share the same category information. In this case, the unknown objects are at risk of being misclassified into known classes, leading to unreliable recognition results. Consequently, although deep learning has outperformed human perception on various recognition tasks [2–4], its real-world deployment remains constrained because of reliability concerns.

Open-set recognition (OSR) [5] has been proposed as the solution to this closed-set predicament. An OSR classifier is designed to classify known objects while simultaneously identifying unknown objects. As summarized in [6], two main approaches are taken to learn decision boundaries for OSR: optimizing the representation of known classes and estimating the distribution of unknown classes, corresponding to the focus of discriminative and generative OSR methods, respectively. Although the effectiveness of existing OSR methods in learning decision boundaries to distinguish between known and unknown classes has been widely demonstrated, their implementations are mainly based on single-modal image datasets such as MNIST [7], SVHN [8], and CIFAR [9]. The test data in these datasets are generally identically distributed to the training data and can provide sufficient descriptions of the objects. However, the aforementioned conditions cannot always be guaranteed in real-world applications because of the limitations of single-modal perception and external disturbances, resulting in the unsatisfactory OSR performance of existing methods based on single-modal perception.

Figure 1(a) illustrates a case of single-modal visual perception. The known and unknown classes cannot be effectively distinguished when single-modal data (front view) fail to provide sufficient descriptions of

* Corresponding author (email: liuzhunga@nwpu.edu.cn)

© Science China Press 2024

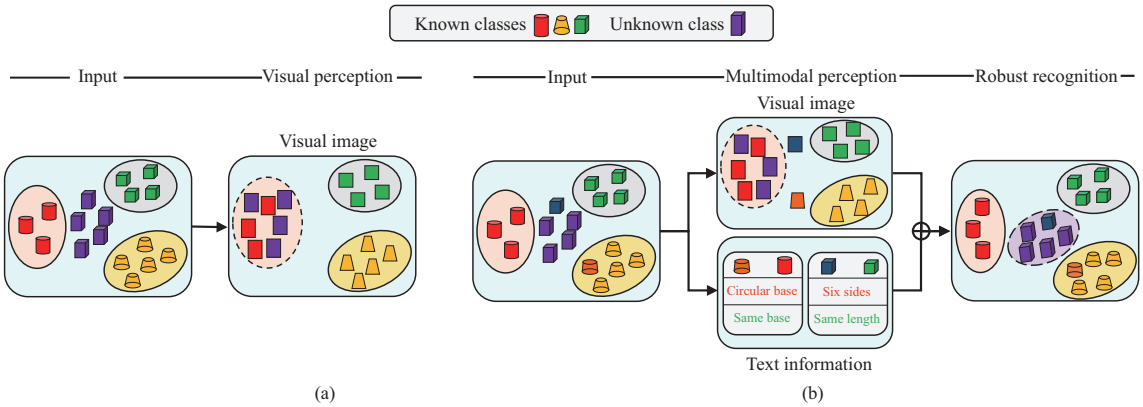

Figure 1 (Color online) (a) Limitations of the OSR methods based on single-modal visual perception; (b) to achieve robust OSR, active assessment of multimodal perception information and adaptive adjustment of the fusion strategy are required.

the objects. In this case, the misclassification is predominantly due to the limited information in the input data rather than deficiencies in OSR method design. Conversely, this confusion can be effectively resolved by supplementing the visual image with additional textual information, such as the surface shape. Therefore, the integration of multimodal perception data is indispensable for achieving robust OSR in real-world applications.

Multimodal perception [10] has been widely explored in various tasks, such as image classification [11], robot perception [12], and earth observation [13]. By integrating perception information from different modalities, more comprehensive descriptions of objects can be obtained to improve recognition system performance. The core problem of multimodal perception is multimodal fusion, which mainly includes feature-level and decision-level fusion methods [14]. The robustness of the model cannot be ensured because of the susceptibility of feature-level fusion methods to external disturbances. In addition, the feature representations of heterogeneous multimodal data are inconsistent in scales. Consequently, the representations of small-scale modalities may be underoptimized because the optimization is dominated by large-scale modalities, which can severely affect the learning of decision boundaries. In comparison, decision-level fusion methods can uniformly represent data from different modalities, demonstrating the potential applicability to OSR with multimodal data. However, complementary information between different modalities may not be fully exploited due to the lack of direct intermodal correlations at the decision level. Moreover, the weight assignments of existing fusion methods are mostly determined on the basis of predefined fusion rules [15] or data fitting [16] and cannot be adaptively adjusted when external disturbances occur during testing, resulting in non-robust OSR performance. As in the multimodal case shown in Figure 1(b), the cylinders and frustums can normally be distinguished by visual images, and textual information about the surface shape is redundant in this case. However, for the orange frustum, whose top and bottom bases are close in size, the discrimination of text information about the surface area is obviously better than visual images. Therefore, the model should actively assess the perception information of different modalities and adjust the fusion strategy during recognition.

Motivated by the analysis above, we propose logit prototype learning (LPL) with active multimodal representation, which incorporates four inspiring insights to achieve robust OSR performance with multimodal perception data. First, the multimodal data are transformed into the logit space, where the representations are directly associated with the classification results. As the logit representations of different modalities are consistent with scale, the intermodal correlations can be effectively explored without being affected by the feature representation gap. Then, an entropy-based uncertainty estimation method is proposed to determine the weights of each modality according to the distribution of their corresponding representations. Therefore, the fusion strategy can adaptively adjust to actively provide comprehensive descriptions of the objects. Moreover, the single-modal and multimodal representations are jointly and interactively optimized, which can simultaneously avoid redundant optimization and preserve modality-specific information during the learning process of decision boundaries. Finally, we extend the OSR framework with an additional fuzzy class to reduce the misclassification risk and obtain the final recognition result based on the scaled distance in a stepwise manner.

We evaluate the proposed method using three multimodal datasets involving data from various modalities, such as visual images, audio recordings, and textual information. The experiments show that the

proposed method can outperform previous OSR methods and maintain robust OSR performance when external disturbances occur during testing. Our main contributions can be summarized as follows:

(1) We propose logit prototype learning to integrate multimodal perception information for robust OSR performance. The representation in the logit space enables directly exploring intermodal correlations while effectively avoiding the feature representation gap between different modalities.

(2) We propose an entropy-based uncertainty estimation method to determine the weight assignments of different single-modal representations. The adaptive adjustment ability of the fusion strategy ensures the comprehensiveness of the descriptive information.

(3) We propose a prototype-based joint optimization strategy to interactively optimize single-modal and multimodal representations, which can avoid redundancy in optimization and preserve modal-specific information during the learning process of decision boundaries.

(4) We propose a stepwise recognition rule to realize a more reasonable decision-making process. The extension of the OSR framework, incorporated with a scaling of the distance, effectively reduces the risk of misclassification and facilitates the distinction between known and unknown classes.

This paper is an extended version of our conference paper [17] in several aspects. First, we incorporate angle-based measurement into the preliminary distance-based joint metric learning in the logit space, further improving the discrimination of decision boundaries for OSR. Second, we extend the OSR framework with an additional fuzzy class to reduce misclassification risks and increase the reliability of recognition results. Finally, we conducted more extensive experiments on three multimodal datasets with various settings to prove the effectiveness and robustness of our proposed method.

2 Related work

2.1 Open-set recognition

2.1.1 *Discriminative OSR method*

Early research on OSR focused on adapting traditional machine learning classifiers to distinguish between known and unknown classes as a binary classification task. Scheirer et al. [5] proposed a formal definition of openness and used a 1-vs-set machine to balance empirical and open-space risk. They further proposed a Weibull-calibrated support vector machine [18] to limit the open space by calibrating the decision scores using extreme value theory. Bendale and Boult [19] extended the nearest class mean classifier to the nearest non-outlier, which measured the distance between samples and class means to reject unknown classes. In addition, Júnior et al. [20] proposed an open-set version nearest neighbor classifier based on the ratio of similarity scores.

Because of the powerful feature extraction and learning representation abilities of deep neural networks (DNNs), current research on OSR is mainly based on deep learning. Bendale et al. [21] replaced the Softmax layer with an OpenMax layer as the initial solution to extend DNNs to open-set scenarios. They computed the distances between the correctly classified samples and their mean activation vectors and calibrated the classification probabilities using a Weibull distribution. Inspired by this work, various deep learning-based OSR methods have been subsequently proposed for further improvement. Shu et al. [22] proposed a sigmoid-based 1-vs-rest final layer for OSR to avoid the careful hyperparameter selection of OpenMax. Yoshihashi et al. [23] proposed classification-reconstruction learning for OSR based on hierarchical latent representations. Jang and Kim [24] modeled the collective decision of multiple one-vs-rest networks to enhance the unknown rejection capability. Vaze et al. [25] used multiple image recognition techniques and the maximum logit score (MLS) during training and testing to enhance OSR performance.

2.1.2 *Generative OSR method*

Generative adversarial learning [26] has been widely used to generate unknown class samples. Ge et al. [27] employed incorrectly predicted samples from a generative adversarial network (GAN) [28] as unknown classes, but their method did not show substantial performance improvement on natural images. Neal et al. [29] proposed an encoder-decoder GAN architecture for counterfactual image generation, which was used as the synthetic unknown samples during training. Recently, Kong and Ramanan [30] proposed OpenGan, which trains an open-vs-closed classifier with adversarially synthesized fake open

data. However, the performance of generative OSR methods heavily depends on the quality of unknown estimation by GAN, making them unsuitable for real-world applications with external disturbances.

Despite continuous progress on single-modal datasets, existing methods ignore how to achieve robust OSR performance with multimodal perception data in real-world applications, particularly when single-modal perception data fails to provide sufficient descriptions of objects.

2.2 Prototype learning

The prototypes [31,32] can be considered overall representations or abstract memories of the corresponding class clusters. Early prototype learning methods were mainly based on update rules or loss functions. However, these methods require hand-designed features and cannot work end-to-end. Therefore, recent studies have focused more on integrating prototype learning into DNNs, including diverse explorations of OSR. Yang et al. [33] proposed generalized convolutional prototype learning (GCPL) to solve the open-world recognition problem by increasing the intraclass compactness. Chen et al. [34] proposed reciprocal point learning (RPL) to model the extraclass space of each class. They then incorporated angle direction into distance metrics and proposed adversarial reciprocal point learning (ARPL) [35] to learn more discriminative decision boundaries. Miller et al. [36] proposed class anchor clustering (CAC) loss to learn tightly clustered representations around anchored class centers in the logit space, unlike the above prototype learning methods in the feature space.

The training process of OSR with single-modal data is more stable in the logit space than in the feature space. More importantly, prototype learning in the logit space can avoid the differences in feature representations of multimodal data and enable the direct exploration of intermodal correlations. These properties are potentially valuable for obtaining uniform and comprehensive multimodal representations for robust OSR and have been ignored by previous methods.

2.3 Multimodal perception

Multimodal perception [10] processes and integrates multimodal data to achieve more accurate and reliable recognition performance, and it has been widely applied in various domains. The classification results of remote sensing images [37] can be effectively enhanced by integrating information from different modalities, e.g., hyper-spectral, multispectral, and synthetic aperture radar. In autonomous driving [38], RGB camera images can provide detailed perception of the surroundings but are sensitive to changes in light conditions. In contrast, LIDAR points are less affected by environmental factors but are limited in their range and resolution. Therefore, the fusion of camera images and LIDAR points can overcome the single-modal limitations and produce more accurate results. For human-machine interaction, integrating audio and visual information can provide more reliable and robust speech recognition results than single-modal audio information [39], which is easily affected by external noise. Based on complementary information from different modalities, the robustness of facial recognition [40] can also be improved by leveraging multimodal representations.

All of the above indicates that the effective integration of multimodal perception information is necessary to achieve robust OSR performance in real-world applications, which corresponds to the main objective of our proposed method.

3 Proposed method

We adaptively integrate multimodal perception information to achieve robust OSR performance. First, the features of each set of single-modal input data are extracted using the corresponding encoders and transformed into the logit space. Then, the weight assignments are determined using the entropy-based uncertainty estimation method, which enables the fusion strategy for multimodal representations to be actively adjusted. Subsequently, the single-modal and multimodal representations are jointly optimized to learn discriminative decision boundaries. Finally, the final recognition result can be obtained using the stepwise recognition rule during testing.

The overall framework of the proposed logit prototype learning with active multimodal representation is shown in Figure 2. In this section, we first provide a brief introduction to the problem setting and motivation for prototype learning in the logit space. We describe each crucial part of our method in Subsections 3.2–3.4.

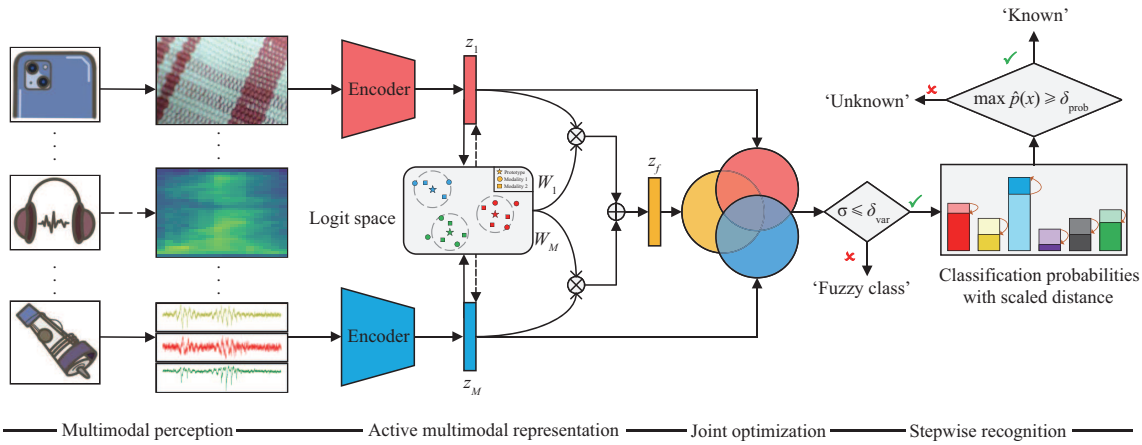

Figure 2 (Color online) Overview of the proposed LPL for OSR with multimodal perception data. The lighter and deeper colored histograms represent classification probabilities of each class before and after scaling the distance using single-modal reachabilities, respectively.

3.1 Problem setting and motivation

3.1.1 *Problem setting*

Given a set of training data $D_{\text{Train}} = \{X, Y\}$ with N known classes, the labels of input samples X are represented as Y. The logit prototypes $\boldsymbol{O} \in \mathbb{R}^{N\times N}$ of known classes are predefined as nontrainable one-hot style vectors in the N-dimensional logit space.

$$\boldsymbol{O} = [o_1, \ldots, o_N] = [\alpha \cdot \boldsymbol{e}_1, \ldots, \alpha \cdot \boldsymbol{e}_N], \tag{1}$$

$$\boldsymbol{e}_1 = [1, 0, \ldots, 0], \quad \boldsymbol{e}_N = [0, \ldots, 0, 1], \tag{2}$$

where hyperparameter α describes the magnitude of the logit prototypes, and $\boldsymbol{e}$ are unit one-hot vectors corresponding to the ground truth labels of each known class.

During training, the feature of a single-modal input sample x is extracted by the encoder f and transformed as the logit embedding $z \in \mathbb{R}^N$. In the logit space, the Euclidean distances between the logit vector and each prototype are represented as $\boldsymbol{d} \in \mathbb{R}^N$:

$$\boldsymbol{d} = [d_1, \ldots, d_N] = [\|z - o_1\|_2, \ldots, \|z - o_N\|_2]. \tag{3}$$

Then, the classification result of x can be calculated using the distance-based Softmin function:

$$p(y = n \mid x) = \frac{\mathrm{e}^{-d_n}}{\sum_{i=1}^{N} \mathrm{e}^{-d_i}}, \ n \in \{1, \ldots, N\}, \tag{4}$$

meaning that the closer the logit embedding z is to the prototype o_n, the higher the probability that the sample x will belong to class n.

During testing, for a set of testing data $D_{\text{Test}} = \{D_k \cup D_u\}$ with N known classes and U unknown classes, OSR aims to correctly classify samples in D_k and label samples in D_u as unknown. The distinction between known and unknown classes is conducted with a probability threshold, which is usually determined by assuming a certain percentage of training samples as outliers.

3.1.2 *Motivation*

Integrating multimodal representations through a proper fusion method is crucial for prototype learning-based OSR with multimodal data. However, existing commonly used fusion methods have several limitations. First, the representation gap between different single-modal feature embeddings is considerable, primarily because of the distinct perception mechanisms of sensors in different modalities. This inconsistency in the feature space can severely affect the distance measurement between feature embeddings and prototypes, which is crucial for the optimization and recognition processes of prototype learning-based OSR methods. Second, although decision-level representations (i.e., classification probabilities) of

multimodal data exhibit high consistency, direct intermodal correlations for decision fusion methods are lacking, making the complementary information between different modalities difficult to fully capture.

To simultaneously overcome the shortcomings at the feature and decision levels, we adapt prototype learning in the logit space for OSR with multimodal data. The representations in the logit space (logit embeddings) serve as feature representations directly associated with the classification results. Therefore, prototype learning in the logit space can effectively bridge the feature representation gap and enable flexible interaction of multimodal data to explore intermodal correlations.

3.2 Active multimodal representation

The foundation for improving recognition performance with multimodal data is to obtain comprehensive descriptions of the objects. However, the fusion strategies of existing methods cannot be actively adjusted during testing, leading to non-robust recognition performance when subjected to external disturbances. In the context of logit prototype learning with multimodal data, the proximity of each single-modal logit embedding to different prototypes is directly related to the certainty of its corresponding classification result. Particularly, if the logit embedding of a modality is close to one prototype and distant from the others, the classification result will have high certainty, indicating sufficient descriptions of the object. Conversely, a classification result based on the logit embedding of a modality similar to distance to different prototypes is highly uncertain.

Inspired by this outcome, we propose an entropy-based uncertainty estimation method to dynamically determine the weights of different modalities for active multimodal representations, which can provide comprehensive descriptions of the objects. For an input sample $\boldsymbol{x} = [x_1, \ldots, x_M]$ with the perception information of M modalities, we first extract the latent features of different modalities using encoders $\boldsymbol{f} = [f_1, \ldots, f_M]$, and adopt linear layers $\boldsymbol{l} = [l_1, \ldots, l_M]$ to transform them as logit embeddings $\boldsymbol{z} \in \mathbb{R}^{M \times N}$:

$$\boldsymbol{z} = [z_1, \ldots, z_M] = [l_1(f_1(x_1)), \ldots, l_M(f_M(x_M))]\,. \tag{5}$$

Next, we calculate the Euclidean distances $\boldsymbol{d} = [d_1, \ldots, d_M]$ between different single-modal logit embeddings and the logit prototypes of known classes:

$$d_m = \left[d_m^1, \ldots, d_m^N\right] = \left[\|z_m - o_1\|_2, \ldots, \|z_m - o_N\|_2\right],\ m \in \{1, \ldots, M\}. \tag{6}$$

Then, we perform class-wise L1 normalization on the reciprocal distances between each single-modal logit embedding and different prototypes:

$$\hat{d}_m = \frac{d_m^{-1}}{\max(\|d_m^{-1}\|_1, \epsilon)}, \tag{7}$$

where ϵ is a small constant to avoid division by zero. This strategy enables positively correlating the proximity to different prototypes with the probabilities of belonging to the corresponding classes. Next, we estimate the uncertainty of each piece of single-modal perception information from the information entropy $\boldsymbol{H} = [H_1, \ldots, H_M]$, which is expressed as follows:

$$H_m = -\sum_{n=1}^{N} \hat{d}_m^n \log \hat{d}_m^n. \tag{8}$$

When fusing different single-modal representations into a multimodal representation, it is reasonable to assign a large weight to the modality with low uncertainty. Accordingly, the weight assignments $\boldsymbol{W} = [W_1, \ldots, W_M]$ of each modality can be determined as follows:

$$W_m = \frac{\mathrm{e}^{-H_m}}{\sum_{j=1}^{M} \mathrm{e}^{-H_j}}. \tag{9}$$

Finally, the multimodal logit embedding z_f can be represented as follows:

$$z_f = \sum_{m=1}^{M} W_m z_m. \tag{10}$$

Therefore, the input $\boldsymbol{x}$ is represented as a joint form of single-modal and multimodal logit embeddings:

$$\tilde{\boldsymbol{z}} = [z_f, z_1, \ldots, z_M]\,. \tag{11}$$

3.3 Prototype-based joint optimization strategy

To effectively distinguish between known and unknown classes, the model learns discriminative decision boundaries by increasing intraclass compactness and interclass separability. However, the optimization strategies of existing OSR methods are mainly based on an individual single-modal representation and cannot be directly applied to multimodal scenarios. On the one hand, optimizing single-modal and multimodal representations with the same objective is redundant, and may limit the exploration of intermodal correlations. On the other hand, optimizing only the multimodal representation will lose the modality-specific information inherent in different modalities.

To this end, we propose a prototype-based joint optimization strategy in the logit space. First, multimodal logit embedding is essentially a weighted expression of different single-modal logit embeddings. In this case, if the single-modal logit embeddings of an object are tightly clustered, the multimodal logit embedding will naturally match the corresponding prototype. Therefore, we directly optimize the Euclidean distances and cosine similarities between single-modal logit embeddings and the corresponding prototypes to increase intraclass compactness. The measurement loss can be expressed as

$$L_{\text{mea}} = \frac{1}{M}\sum_{m=1}^{M}\left(\|z_m - o_{n_c}\|_2 - \beta\frac{z_m \cdot o_{n_c}}{\|z_m\|\|o_{n_c}\|}\right), \tag{12}$$

where n_c is the corresponding class of x, and β is the coefficient of the angle-based measurement. The representation discrimination between different classes can be effectively improved by combining Euclidean distance and cosine similarity. Then, the multimodal logit embedding is used to calculate the classification probabilities by Softmin function based on the sum-to-one property:

$$p\left(y = n \mid x\right) = \frac{\mathrm{e}^{(\beta\frac{z_m \cdot o_y}{\|z_m\|\|o_y\|} - \|z_m - o_n\|_2)}}{\sum_{i=1}^{N}\mathrm{e}^{(\beta\frac{z_m \cdot o_y}{\|z_m\|\|o_y\|} - \|z_m - o_i\|_2)}},\ n \in \{1, \ldots, N\}. \tag{13}$$

Next, the classification probabilities are used to optimize the classification loss, which is developed on the basis of the negative log-probability of the ground-truth label:

$$L_{\text{cls}} = -\log p(y = n_c \mid x), \tag{14}$$

where a small value of L_{cls} refers to a large distance gap between the representation of the sample and the corresponding prototype and other prototypes, and the interclass separability can be inherently improved. Finally, the discrimination of decision boundaries can be effectively improved with the joint supervision of L_{cls} and L_{mea}:

$$L_{\text{joint}} = L_{\text{cls}} + \lambda L_{\text{mea}}, \tag{15}$$

where λ is a weight parameter for the measurement loss. On the basis of the joint optimization of representations interactively, the model can learn discriminative decision boundaries without optimization redundancy. In addition, the specific information of each modality can also be well preserved.

After finishing the optimization process, we replace the shared one-hot style prototypes with the mean values of single-modal and multimodal logit embeddings, which can better measure the similarities between the objects and prototypes during testing. Therefore, the prototypes of each class are extended into different sets $\tilde{\boldsymbol{O}} = [\tilde{o}_1, \ldots, \tilde{o}_N]$ and can be expressed as follows:

$$\tilde{o}_n = \left[\tilde{o}_f^n, \tilde{o}_1^n, \ldots, \tilde{o}_M^n\right] = \frac{1}{|N_n|}\sum_{i=1}^{|N_n|}\left[z_f(i), z_1(i), \ldots, z_M(i)\right],\ n \in \{1, \ldots, N\}, \tag{16}$$

where N_n is the number of training samples that belong to class n. To capture the distribution properties of different classes across modalities in the logit space, we define the distances between the representations and corresponding prototypes as the radiuses of each single-modal class cluster $\boldsymbol{R} = [r_1, \ldots, r_N]$:

$$r_n = [r_1^n, \ldots, r_M^n] = \frac{1}{|N_n|}\sum_{i=1}^{|N_n|}\left[\|z_1(i) - \tilde{o}_1^n\|_2, \ldots, \|z_M(i) - \tilde{o}_M^n\|_2\right]. \tag{17}$$

The details of the proposed prototype-based joint optimization strategy are summarized in Algorithm 1.

Algorithm 1 Prototype-based joint optimization strategy in the logit space

Input: Training data x with class label y and the perception information of M modalities; initializing parameters θ for single-modal encoders, and logit prototytpes $\boldsymbol{O} = [o_1, \ldots, o_N]$; hyperparameters α, β, and λ; number of iterations $t \leftarrow 0$.
Output: Parameters θ, prototype sets $\tilde{\boldsymbol{O}}$, and radiuses $\boldsymbol{R}$ of single-modal clusters.

while not converged **do**
  $t \leftarrow t + 1$;
  Transform x into different single-modal logit embeddings $\boldsymbol{z}$;
  Determine the weights $\boldsymbol{W}$ of each single modality by entropy-based uncertainty estimation;
  Actively represent multimodal logit embedding as z_f and form the joint representation $\tilde{\boldsymbol{z}}$;
  Compute the loss of joint optimization by $L_{\text{joint}} = L_{\text{cls}} + \lambda L_{\text{mea}}$;
  Update the parameters θ for single-modal encoders;
end while
Calculate the prototype sets $\tilde{\boldsymbol{O}}$ and radiuses $\boldsymbol{R}$ of single-modal clusters;
Return θ, $\tilde{\boldsymbol{O}}$, and $\boldsymbol{R}$.

3.4 Stepwise recognition rule with distance measurement

During testing, the known and unknown classes can be distinguished on the basis of active multimodal representation with discriminative decision boundaries in most cases. However, the occurrence of some extremely ambiguous outliers is inevitable. For the OSR with two-modal data, if the representation of a sample can well match the prototype of a certain class in one modality but deviates considerably from that category in the other, the existing OSR framework will fail to provide reliable recognition results. On the one hand, the deviation may be caused by excessive external disturbance to a known class sample in that modality. On the other hand, the proximity to a known class prototype may result from similar characteristics of an unknown sample to that of a known class. In this case, arbitrarily classifying these ambiguous outliers into either known or unknown classes can result in high misclassification risks. Therefore, it is more reasonable to assign them into an additional subset and wait for further verification.

To solve this problem, we propose a stepwise recognition rule with distance measurement that can not only classify known classes while rejecting unknown classes but also reduce the misclassification risk caused by ambiguous outliers. Specifically, for each test sample x, we first measure the distance between its multimodal logit embedding z_f and different fusion prototypes $\tilde{\boldsymbol{o}}_f = [\tilde{o}_f^1, \ldots, \tilde{o}_f^N]$, and use the class label of the nearest fusion prototype as the initial classification result y_{init}:

$$y_{\text{init}} = \underset{n \in \{1,\ldots,N\}}{\arg\min} \left\| z_f - \tilde{o}_f^n \right\|_2. \tag{18}$$

Next, for each modality, we calculate the distance between the single-modal logit embedding and the prototype of the initial classification and define relative reachability $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_M]$ as the ratio of the radius of the corresponding single-modal class cluster to this distance:

$$\gamma_m = \frac{r_m^{y_{\text{init}}}}{\left\| z_m - \tilde{o}_m^{y_{\text{init}}} \right\|_2}. \tag{19}$$

We then scale the relative reachabilities of each modality by max normalization:

$$\boldsymbol{\gamma}' = [\gamma_1', \ldots, \gamma_M'] = \frac{\gamma}{\max(\gamma_m)}, \tag{20}$$

and calculate the variance σ of the normalized reachabilities:

$$\sigma = \frac{1}{M} \sum_{m=1}^{M} (\gamma_m' - \bar{\gamma'})^2, \tag{21}$$

which is used as the indicator of the fuzzy class with a variance threshold δ_{var}:

$$y = \begin{cases} \text{known/unknown classes}, & \text{if } \sigma \leqslant \delta_{\text{var}}, \\ \text{fuzzy class}, & \text{otherwise}. \end{cases} \tag{22}$$

Samples with consistent single-modal relative reachabilities (i.e., small variance) are classified into known or unknown classes, whereas samples with inconsistent single-modal relative reachabilities (i.e., large variance) are classified into the fuzzy class.

Algorithm 2 Stepwise recognition rule with distance measurement

Input: Testing data x, single-modal encoders with parameters θ. Prototype sets $\tilde{\boldsymbol{O}}$, radiuses $\boldsymbol{R}$ of single-modal clusters, variance threshold δ_{var}, and probability threshold δ_{prob}.
Output: Recognition result y.
Transform x as the joint representation of single-modal and multimodal logit embeddings $\tilde{\boldsymbol{z}}$;
Set the class label of the nearest fusion prototype to the multimodal logit embedding z_f as the initial classification decision y_{init};
Calculate the relative reachability $\boldsymbol{\gamma}$ and scale it into $\boldsymbol{\gamma}'$;
Calculate the variance σ of normalized reachabilities;
if $\sigma \leqslant \delta_{\text{var}}$ **then**
    Get the scaled distance $\hat{d}_f$ and use it to calculate the final probabilities $\hat{p}\,(x)$;
    if $\max \hat{p}(x) \geqslant \delta_{\text{prob}}$ **then**
        $y = \arg\max \hat{p}\,(y = n \mid x)$;
    else
        $y =$ unknown class;
    end if
else
    $y =$ fuzzy class;
end if
Return y.

Subsequently, the emphasis shifts toward further distinguishing between known and unknown classes among the remaining samples not assigned to the fuzzy class. To further enhance the distinction between known and unknown classes, we use relative reachabilities to scale the distances of multimodal representation to different fusion prototypes $\hat{d}_f = [\hat{d}_f^1, \ldots, \hat{d}_f^N]$:

$$\hat{d}_f^n = \left\| z_f - \tilde{o}_f^n \right\|_2 \prod_{m=1}^{M} \gamma_m. \tag{23}$$

A higher product value of relative reachabilities will widen the distance gap between the multimodal logit embedding and different fusion prototypes. In contrast, the separability between known and unknown classes will decrease. In this case, the unknown class is more distinct from the known classes based on the distance-based classification function. The final classification probabilities are calculated as

$$\hat{p}\,(y = n \mid x) = \frac{\mathrm{e}^{-\hat{d}_f^n}}{\sum_{i=1}^{N} \mathrm{e}^{-\hat{d}_f^i}}. \tag{24}$$

Finally, the known and unknown classes can be effectively distinguished by employing a probability threshold δ_{prob}:

$$y = \begin{cases} \arg\max \hat{p}(y = n \mid x), & \text{if } \sigma \leqslant \delta_{\text{var}} \text{ and } \max \hat{p}(x) \geqslant \delta_{\text{prob}}, \\ \text{unknown class}, & \text{if } \sigma \leqslant \delta_{\text{var}} \text{ and } \max \hat{p}(x) < \delta_{\text{prob}}, \\ \text{fuzzy class}, & \text{otherwise.} \end{cases} \tag{25}$$

Therefore, the OSR framework can be extended from $(N+1)$-way to $(N+2)$-way classification, which can mitigate the unreliability due to the potential misclassification risk caused by ambiguous outliers.

In practice, a well-chosen probability plays a key role in OSR, but it is difficult and time-consuming to tune. For theoretical evaluation, the utilization of an arbitrary threshold for OSR without prior knowledge of unknown classes is unreasonable [29]. Therefore, evaluations based on threshold-independent metrics were also conducted (reported in Section 4). The details of the proposed stepwise recognition rule during testing are summarized in Algorithm 2.

4 Experiments

4.1 Experimental setup

4.1.1 *Datasets and class partition*

We use the concept of openness [5] $O^* = 1 - \sqrt{\frac{2 \times N_{\text{Train}}}{N_{\text{Test}} + N_{\text{Train}}}}$ to describe the proportion of known and unknown classes divisions of each dataset, where N_{Train} is the number of known classes during training, and N_{Test} is the total number of known and unknown classes during testing.

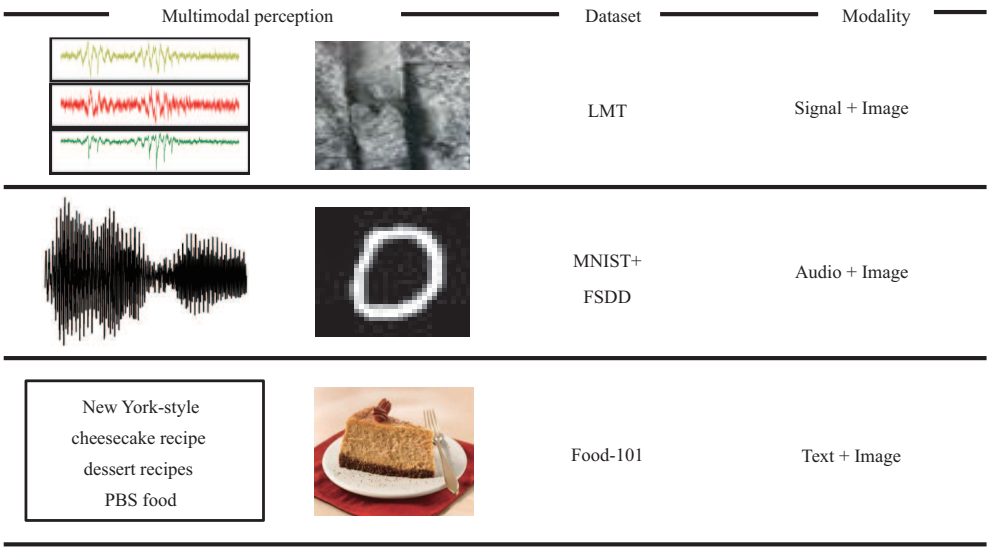

Figure 3 (Color online) Some samples in three multimodal datasets.

We evaluate the effectiveness of our proposed method for robust OSR on three multimodal datasets, and some representative samples are shown in Figure 3.

• LMT [41] is a multimodal dataset for surface material classification, which contains various perception information of 108 fine-grained classes in nine coarse-grained classes. We use visual images and tactile acceleration traces as the input, and two fine-grained classes in each coarse-grained class are randomly selected and set as known and unknown classes, and $O^* = 18.35\%$.

• MNIST [7]+ FSDD[1)] are image and speech datasets of digit numbers, respectively. We pair the data based on class labels for image-audio multimodal recognition. Five known and unknown classes are randomly chosen, and $O^* = 18.35\%$.

• Food-101 [42] is a multimodal classification dataset containing images and texts of 101 food classes. Twenty classes are randomly selected as known classes, and the remaining classes are regarded as unknown classes, and $O^* = 42.50\%$.

4.1.2 *Data preprocessing and implementation details*

To facilitate the feature extraction process and explore diverse fusion strategies, the following preprocessing procedures are applied to the audio and signal data. First, following the process in [43], we convert the three-axes acceleration signals in the LMT dataset into 1-D signals, which are transformed into spectrogram images by short-time Fourier transform (STFT). Data augmentation operations, such as flip and rotate, are also used to make the training data of the LMT dataset more adequate. Then, we transform the audio files in the FSDD dataset into the corresponding spectrogram images. We retain the original data form of the Food-101 dataset, i.e., text and images. To thoroughly validate the effectiveness of the proposed method, we simulate different types of disturbances of various degrees. First, for the LMT dataset, we perform pixel-value exponential transformations on the testing images to simulate the change in brightness, where the exponent is randomized from (0.5, 1). Second, for MNIST+FSDD datasets, we add random noise with a random proportion from 10% to 50% to the testing images. Third, for the Food-101 dataset, we simultaneously use pixel-value exponential transformations to change the brightness and delete words from the text with a random probability from (0.1, 0.5). This test scenario is more challenging because the perceptual information of each modality is disturbed to various degrees.

4.1.3 *Implementation details*

We used the convolutional blocks of VGG13 [2] as the feature encoder for image and spectrogram data, and bidirectional encoder representations from transformers (BERT) [44] as the feature encoder for text data. The globally pooled convolutional features and BERT output embeddings are transformed into the logit space by a linear layer. All the methods are trained for 100 epochs with a batch size of 32 and

1) https://github.com/Jakobovski/free-spoken-digit-dataset.

Table 1 Comparison of the OSR performance of different methods on the LMT dataset[a)]

Method	Image		Spectrogram		Early fusion		Late fusion		Decision fusion	
	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR
Softmax	65.9±3.0	48.1±2.2	71.5±6.1	65.9±7.3	70.6±1.9	52.5±0.8	76.6±4.0	61.2±6.9	76.0±4.3	61.2±6.9
OpenMax [21]	56.9±5.6	31.6±1.8	43.1±7.6	22.2±6.7	59.1±5.0	34.5±1.1	50.4±2.3	24.9±2.7	40.2±3.2	24.8±2.9
G-OpenMax [27]	56.8±3.5	25.2±1.5	48.9±5.1	30.1±5.8	51.0±3.7	36.0±1.5	50.8±3.3	25.5±3.0	45.4±2.7	25.4±3.4
CROSR [23]	64.6±4.0	35.5±4.9	60.7±5.8	48.6±6.2	–	–	–	–	66.9±3.8	55.2±4.6
GCPL [33]	54.2±5.1	33.3±4.3	60.1±3.2	46.9±3.6	59.0±5.5	31.2±5.3	66.6±9.2	53.9±7.4	78.3±6.2	68.7±5.0
RPL [34]	57.7±6.5	36.9±4.1	54.6±5.3	39.3±2.4	60.0±8.5	33.2±5.2	56.3±3.8	33.4±8.2	63.6±9.1	40.5±4.1
ARPL [35]	59.1±5.0	43.3±4.4	66.5±1.9	61.9±3.4	60.9±5.6	44.0±4.1	63.0±0.7	51.3±3.5	67.6±2.2	59.4±2.6
CAC [36]	67.2±2.7	47.4±1.4	74.6±2.9	71.8±3.4	70.5±3.4	54.0±3.2	73.4±7.0	62.7±4.4	78.1±6.0	68.9±5.7
MLS [25]	66.5±3.7	46.3±1.3	70.1±1.9	67.3±3.4	76.6±3.2	67.6±2.0	75.8±5.1	69.8±5.0	76.6±3.9	68.2±5.9
MMOSR [17]	AUC/OSCR: 81.9±1.4/74.1±3.6									
LPL	AUC/OSCR: **84.2±1.2/79.4±2.4**									

a) The experimental results are averaged among three randomized trials. The best performance is in bold.

are optimized using the SGD optimizer with a momentum of 0.9. The initial learning rate of training is 0.01 and is divided by 10 every 30 epochs. For hyperparameter values of LPL, the magnitude of logit prototypes α, the coefficient of angle-based measurement β, and the weight of measurement loss λ_{mea} are set to 5, 1, and 0.1, respectively. The experiments are implemented with the support of NVIDIA RTX 3090 GPU in the Pytorch framework.

4.1.4 *Evaluation metrics*

The following metrics are used to evaluate the OSR performance of the proposed method:

(1) The area under ROC curve (AUC) is the most commonly used evaluation metric for OSR; it is threshold-independent and can comprehensively measure the effect of all thresholds.

(2) The open-set classification rate (OSCR) indicates the recognition performance of unknown samples with the consideration of the classification accuracy of known classes based on the correct classification rate (CCR) and the false positive rate (FPR).

4.2 Comparison to the state-of-the-art methods

We compare the performance of LPL with that of various combinations of state-of-the-art OSR methods and commonly used fusion strategies to prove the effectiveness of our proposed method.

The experimental results on three multimodal datasets are shown in Tables 1–3. The first two of the five columns represent methods with single-modal input (i.e., image, spectrogram, and text), while the last three of the five columns represent methods with multimodal input based on commonly used feature-level and decision-level fusion strategies. Therefore, the AUC and OSCR performance of the comparison methods at each position in the table refers to the combination of the corresponding fusion strategy and the OSR method. For experiments on the LMT and MNIST+FSDD datasets, we do not report the performances of CROSR with feature-level fusion strategies because multimodal data are difficult to reconstruct from unique fused feature representations. In addition, we do not report the performance of G-OpenMax and CROSR on the Food-101 dataset because their generation and reconstruction approaches are oriented to image data.

According to the results, LPL substantially outperforms the other methods on all datasets, indicating that it can achieve robust OSR performance by effectively integrating multimodal perception information. In addition, OpenMax fails to provide significant improvements in OSR performance and may even degrade it compared with the baseline Softmax classification. When the description from certain sensory data is incomplete, the representations of known class samples deviate from the corresponding training information, making them confused with the unknown classes. Therefore, the classification probability calibration method based on the statistical characteristics (Weibull cumulative distribution function) of the training data is not applicable to OSR scenarios with external disturbances. Furthermore, the OSR performances of the generative OSR methods are also unsatisfactory.

Compared with the training data of known classes, the representations of disturbed test samples tend to overlap more with the estimated unknown samples because they are both outliers of known training samples. Similarly, such representation inconsistencies lead to large reconstruction errors, and the

Table 2 Comparison of the OSR performance of different methods on MNIST+FSDD datasets[a)]

Method	Image		Spectrogram		Early fusion		Late fusion		Decision fusion	
	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR
Softmax	69.4±9.1	48.7±9.9	66.8±6.6	64.9±7.6	79.9±4.6	58.4±6.5	76.8±10.1	57.6±11.3	72.1±6.8	55.8±9.5
OpenMax [21]	61.8±0.7	49.7±0.8	66.7±3.4	61.1±5.1	60.0±1.3	49.1±0.4	61.0±2.3	52.2±3.1	58.8±2.0	46.7±3.9
G-OpenMax [27]	62.4±0.8	47.9±1.9	60.8±1.3	55.2±5.4	61.2±1.6	47.0±1.8	62.1±1.7	47.7±3.0	63.8±2.2	49.1±2.8
CROSR [23]	58.3±2.8	42.8±7.2	67.2±2.4	62.7±5.3	–	–	–	–	61.2±4.6	47.9±1.1
GCPL [33]	61.8±4.7	34.3±6.2	41.5±8.9	36.2±7.0	58.8±1.6	30.8±2.1	62.9±3.2	37.4±3.1	62.4±1.8	44.7±12.1
RPL [34]	56.9±2.0	27.9±2.6	45.0±14.7	35.6±12.4	57.7±1.3	25.9±0.9	57.8±1.7	26.2±0.5	60.7±5.1	35.3±3.2
ARPL [35]	65.7±2.2	39.4±3.6	71.2±10.8	69.5±13.0	66.4±2.3	41.5±1.4	67.0±4.6	42.4±6.2	72.7±3.7	57.0±9.8
CAC [36]	77.6±0.4	56.1±1.1	76.5±4.6	74.7±5.8	80.6±1.9	61.2±1.3	81.6±3.8	67.8±5.8	78.4±5.0	62.8±4.1
MLS [25]	78.2±3.6	58.3±2.3	78.9±3.9	76.3±4.7	80.7±1.8	60.3±1.2	78.5±3.9	63.3±3.7	83.7±0.9	69.6±4.3
MMOSR [17]	AUC/OSCR: 90.6±3.9/89.5±5.0									
LPL	AUC/OSCR: **92.0±2.5/89.9±4.5**									

a) The experimental results are averaged among three randomized trials. The best performance is in bold.

Table 3 Comparison of the OSR performance of different methods on the Food-101 dataset[a)]

Method	Image		Text		Early fusion		Late fusion		Decision fusion	
	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR	AUC	OSCR
Softmax	70.1±0.9	52.7±1.7	80.5±1.0	67.5±1.4	82.6±0.5	72.1±0.5	82.5±0.5	72.7±0.8	83.1±0.9	74.6±1.2
OpenMax [21]	72.7±0.8	57.7±1.0	80.7±1.1	65.9±1.0	81.4±0.4	70.8±0.8	78.0±1.1	69.2±1.5	82.0±1.3	73.4±1.7
GCPL [33]	66.3±1.8	41.5±3.2	66.4±2.2	43.4±2.6	76.1±1.5	61.1±3.4	80.5±1.7	69.1±1.5	72.6±0.4	56.4±1.3
RPL [34]	58.8±0.7	36.1±1.7	78.3±2.8	63.2±2.6	79.8±1.9	64.7±1.6	78.9±1.4	65.7±0.8	80.4±1.8	66.3±0.8
ARPL [35]	74.1±0.3	58.9±1.7	78.8±1.0	65.9±1.1	85.0±1.1	77.1±0.7	82.9±1.1	72.9±0.8	78.4±6.0	65.5±9.8
CAC [36]	75.6±1.6	62.0±2.1	80.2±1.3	67.3±1.2	79.3±1.6	67.9±1.2	79.8±1.1	68.6±0.6	79.6±1.1	67.9±1.0
MLS [25]	72.9±0.7	58.1±1.4	79.2±0.9	65.7±0.7	83.5±0.7	72.6±0.4	82.1±1.3	70.5±1.1	83.2±1.1	72.7±1.2
MMOSR [17]	AUC/OSCR: 85.3±1.5/76.2±2.2									
LPL	AUC/OSCR: **90.8±0.9/84.8±0.8**									

a) The experimental results are averaged among three randomized trials. The best performance is in bold.

known and unknown classes cannot be accurately distinguished by reconstruction-based representation. Moreover, for prototype-based OSR methods, GCPL can achieve better OSR performance than RPL when the openness is small (LMT and MNIST+FSDD) but is inferior to RPL when the openness is large (Food-101). This result is obtained because the class clusters learned by GCPL may overlap with unknown classes when the OSR problem becomes more complex [35], leading to the increase in the open-space risk. The OSR performance of ARPL is better than that of RPL in most cases with different fusion strategies, indicating the superiority of incorporating angle and distance measurements to improve the discriminability of decision boundaries. In addition, for the performance of each OSR method with different fusion strategies, the combinations based on decision fusion lead to the best performance in most cases. This result demonstrates that higher-level representations with stronger consistency are more suitable for learning decision boundaries for OSR with multimodal data.

In addition to the above results based on OSR metrics, we report the closed-set accuracy of different methods on each dataset in Table 4. Our proposed method outperforms Softmax and other prototype-based OSR methods, indicating that our method can also improve the discrimination of feature representations for known classes. However, the model based on other prototype-based methods fails to maintain the closed-set classification ability in some cases with external disturbance.

4.3 Ablation studies

4.3.1 *Contribution of each part*

Our proposed method comprises three main parts: the active multimodal representation, the joint optimization strategy, and the stepwise recognition rule. We conducted ablation experiments on the LMT dataset by removing each of these three parts to prove their contributions to OSR with multimodal data. The performances are reported by AUC, OSCR, and Accuracy (ACC), as shown in Table 5. The baseline represents the results of Softmax classification with decision fusion. First, model performance can still be improved compared with the baseline when one of the three main parts is removed in most cases. Then,

Table 4 Closed-set accuracy on each dataset[a)]

Method	LMT	MNIST+FSDD	Food-101
Softmax	73.5	71.1	83.7
GCPL [33]	77.9	71.8	72.3
RPL [34]	67.4	61.5	72.5
ARPL [35]	71.5	78.3	75.8
CAC [36]	81.2	73.6	74.0
LPL	**89.1**	**95.0**	**89.2**

a) The best results for each dataset are in bold.

Table 5 Ablation experiments to prove the effectiveness of each part of the proposed method[a)]

Method	LMT			MNIST+FSDD			Food-101		
	AUC	OSCR	ACC	AUC	OSCR	ACC	AUC	OSCR	ACC
Baseline	76.0	61.2	73.5	72.1	55.8	71.1	83.1	74.6	83.7
w/o active representation	81.2	75.8	85.3	90.4	88.5	92.6	84.9	76.1	83.0
w/o joint optimization	79.6	68.8	76.7	78.1	71.7	80.9	82.4	73.0	79.9
w/o stepwise recognition	82.7	76.1	86.5	91.2	89.1	94.2	86.8	77.9	84.2
LPL	**84.2**	**79.4**	**89.1**	**92.0**	**89.9**	**95.1**	**90.8**	**84.8**	**89.2**

a) The best results for each dataset are in bold.

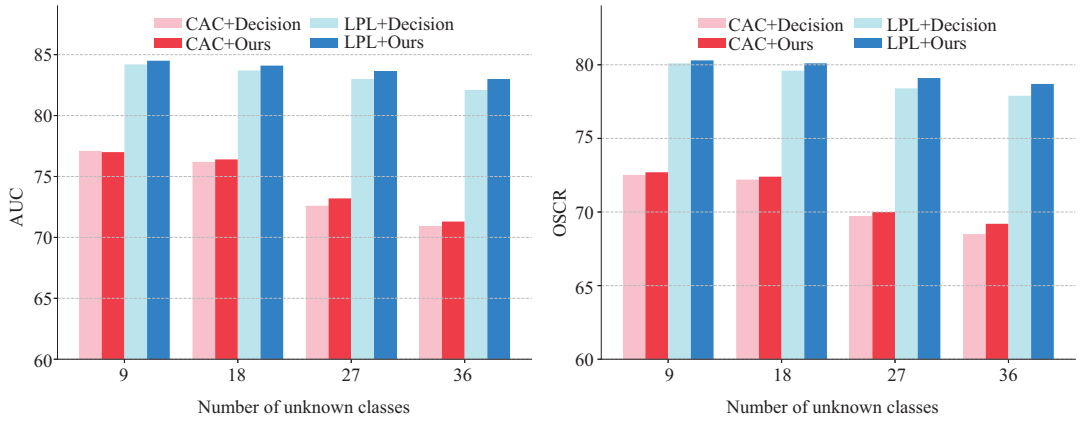

Figure 4 (Color online) Ablation experiments of the active multimodal representations on the LMT dataset.

the performance degradation over LPL is more considerable in the absence of the active multimodal representation or the joint optimization strategy. Moreover, the OSR performance can still be effectively improved based on the joint employment of these two parts without the stepwise recognition rule. This finding indicates that the discriminability of the decision boundaries is already sufficient through the joint optimization of comprehensively integrated multimodal representations. Finally, combining all these three parts into LPL leads to the best performance. Therefore, each part contributes to OSR with multimodal data, and their collaboration can improve OSR performance more effectively.

4.3.2 *Effectiveness of active multimodal representation*

We propose an entropy-based uncertainty estimation method to fuse multimodal data for active multimodal representations. To prove the effectiveness of the proposed fusion strategy, we compared the OSR performance of LPL with different fusion strategies. The experiments were conducted on the LMT dataset by selecting one known class and different numbers of unknown classes in each coarse-grained class. The performances are reported by AUC and OSCR, as shown in Figure 4. We see that the OSR performance of LPL with the proposed fusion strategy consistently surpasses that of the traditional decision-level fusion strategy. In addition, we conducted experiments by combining CAC with different fusion strategies to investigate the effectiveness of the proposed fusion strategy on other prototype-based OSR methods. The results show that the OSR performance of CAC can also be improved on the basis of active multimodal representations, further showing the effectiveness of our proposed fusion strategy.

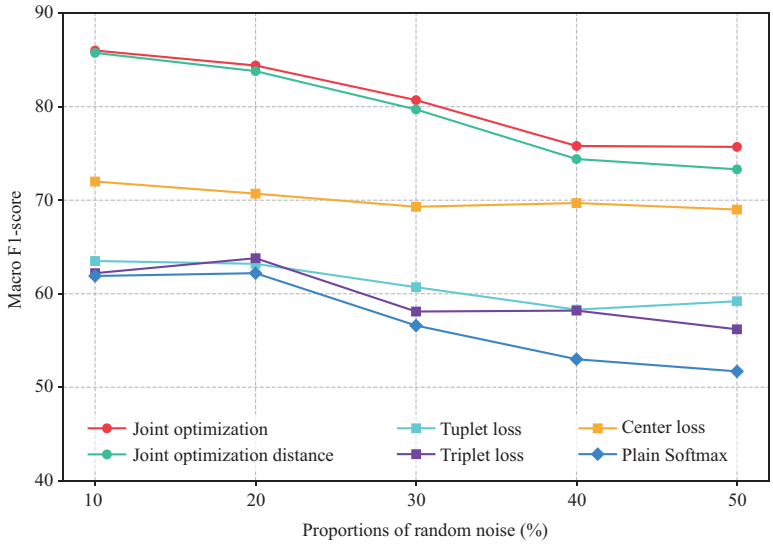

Figure 5 (Color online) Macro-average F1-scores on MNIST+FSDD datasets with various proportions of random noise.

4.3.3 *Effectiveness of joint optimization strategy*

The proposed joint optimization strategy is closely related to the distance-based loss functions. To prove the effectiveness of the proposed joint optimization strategy, we compare the OSR performance of networks trained by it with that of other distance-based loss functions [45–47]. The experiments were conducted on MNIST+FSDD datasets with various proportions of random noise. For a fair comparison, we adopt the decision-level fusion strategy to integrate multimodal representations of our method without the stepwise recognition rule. The performances are reported using the macro-average F1-score, as shown in Figure 5. The "Plain Softmax" refers to Softmax classification with the decision-level fusion strategy, and the "Joint optimization distance" refers to joint optimization with only distance-based measurement. First, the proposed joint optimization strategy considerably outperforms other distance-based loss functions. Then, training the network by center loss improves recognition performance better than training with triplet or tuple loss, demonstrating the applicability of prototype-based methods for OSR. Moreover, extending the joint optimization strategy with angle-based measurement can achieve better performance across various situations, particularly when disturbances are large.

4.4 Analysis of hyperparameter settings

In this experiment, we analyze the effect of the values of hyperparameters included in our method, i.e., the magnitude of logit prototypes α, the coefficient of angle-based measurement β, and the weight of measurement loss λ. The experiments were conducted on the LMT dataset, and the results are reported by the average AUC, as shown in Figure 6. First, we set the value of β to 0 and varied the value of λ based on the different magnitudes of the logit prototypes α. According to the results, the OSR performance can be effectively improved with the measurement loss and is relatively stable, except when λ is too large or too small. When α is small, the change in λ does not result in much improvement. When α becomes larger, a small λ will degrade the performance, which is caused by insufficient intraclass compactness. We then fix the value of λ to 0.1 and vary the value of β based on different magnitudes of logit prototypes α. The results show that the proposed method is insensitive to the change in β. Finally, setting the magnitude of logit prototypes to a moderate value usually leads to the best results. According to the above analyses, a moderate value of λ can lead to more consistent performance improvements. Except when α is too large, setting the value of β to 1 provides the best performance. Therefore, we set the values of hyperparameters α, β, and λ to 5, 1, and 0.1, respectively.

5 Conclusion

In this paper, we proposed LPL with active multimodal representation for robust OSR. Specifically, a learning framework in the logit space was designed to avoid representation differences between different

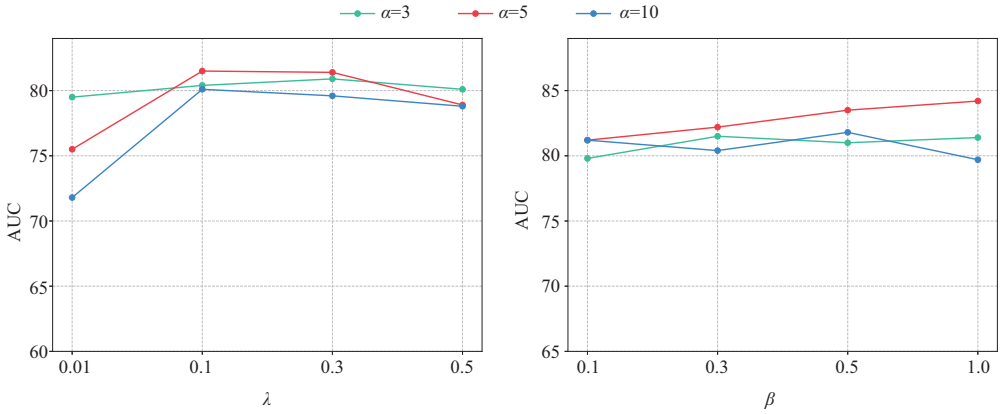

Figure 6 (Color online) OSR performance of the proposed method with different hyperparameter values.

modalities while exploring intermodal correlations. The proposed entropy-based uncertainty estimation method enables the active representation of multimodal data, overcoming the limitations of traditional fusion strategies. Therefore, the robustness of recognition systems can be guaranteed with comprehensive descriptions of the objects in the face of external disturbances. After learning discriminative decision boundaries using the joint optimization strategy, the misclassification risks can be further reduced based on the stepwise recognition rule with the fuzzy class. We conducted extensive experiments on multiple datasets to prove the effectiveness of the proposed method, and the experimental results show that LPL outperforms other state-of-the-art OSR methods in all cases. In the future, we will attempt to further explore OSR with other challenging problems in the open world, such as learning with incremental classes and learning from noisy labels.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. U20B2067), Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (Grant No. CX2023015), and Cultivation Foundation for Excellent Doctoral Dissertation of the School of Automation of Northwestern Polytechnical University.

References

1 Zhou Z H. Open-environment machine learning. Natl Sci Rev, 2022, 9: nwac123
2 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, 2015
3 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
4 Pan X Y, Fan Y-X, Jia J, et al. Identifying RNA-binding proteins using multi-label deep learning. Sci China Inf Sci, 2019, 62: 019103
5 Scheirer W J, Rocha A D R, Sapkota A, et al. Toward open set recognition. IEEE Trans Pattern Anal Mach Intell, 2015, 35: 1757–1772
6 Geng C, Huang S J, Chen S. Recent advances in open set recognition: a survey. IEEE Trans Pattern Anal Mach Intell, 2020, 43: 3614–3631
7 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
8 Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning. In: Proceedings of the Conference on Neural Information Processing Systems Workshops, 2011
9 Krizhevsky A, Hinton G. Learning Multiple Layers of Features From Tiny Images. Technical Report. Toronto: University of Toronto, 2009
10 Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell, 2018, 41: 423–443
11 Gong C, Tao D, Maybank S J, et al. Multi-modal curriculum learning for semi-supervised image classification. IEEE Trans Image Process, 2016, 25: 3249–3260
12 Zhang W C, Sun F C, Wu H, et al. A framework for the fusion of visual and tactile modalities for improving robot perception. Sci China Inf Sci, 2017, 60: 012201
13 Sun X, Tian Y, Lu W X, et al. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Sci China Inf Sci, 2023, 66: 140301
14 Mangai U G, Samanta S, Das S, et al. A survey of decision fusion and feature fusion strategies for pattern classification. IETE Tech Rev, 2010, 27: 293–307
15 Huang L Q, Liu Z G, Pan Q, et al. Evidential combination of augmented multi-source of information based on domain adaptation. Sci China Inf Sci, 2020, 63: 210203
16 Liu Z G, Ning L B, Zhang Z W. A new progressive multisource domain adaptation network with weighted decision fusion. IEEE Trans Neural Netw Learn Syst, 2024, 35: 1062–1072
17 Fu Y, Liu Z, Yang Y, et al. Adaptive open set recognition with multi-modal joint metric learning. In: Proceedings of the 5th Chinese Conference on Pattern Recognition and Computer Vision, 2022. 631–644
18 Scheirer W J, Jain L P, Boult T E. Probability models for open set recognition. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 2317–2324

19 Bendale A, Boult T. Towards open world recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1893–1902
20 Júnior P R M, de Souza R M, Werneck R O, et al. Nearest neighbors distance ratio open-set classifier. Mach Learn, 2017, 106: 359–386
21 Bendale A, Boult T E. Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1563–1572
22 Shu L, Xu H, Liu B. DOC: deep open classification of text documents. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017. 2911–2916
23 Yoshihashi R, Shao W, Kawakami R, et al. Classification-reconstruction learning for open-set recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 4016–4025
24 Jang J, Kim C O. Collective decision of one-vs-rest networks for open-set recognition. IEEE Trans Neural Netw Learn Syst, 2024, 35: 2327–2338
25 Vaze S, Han K, Vedaldi A, et al. Open-set recognition: a good closed-set classifier is all you need. In: Proceedings of the International Conference on Learning Representations, 2022
26 Gui J, Sun Z, Wen Y, et al. A review on generative adversarial networks: algorithms, theory, and applications. IEEE Trans Knowl Data Eng, 2021, 35: 3313–3332
27 Ge Z, Demyanov S, Chen Z, et al. Generative OpenMax for multi-class open set classification. In: Proceedings of British Machine Vision Conference, 2017
28 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. Commun ACM, 2020, 63: 139–144
29 Neal L, Olson M, Fern X, et al. Open set learning with counterfactual images. In: Proceedings of the European Conference on Computer Vision, 2018. 613–628
30 Kong S, Ramanan D. OpenGAN: open-set recognition via open data generation. IEEE Trans Pattern Anal Mach Intell, 2024. doi: 10.1109/TPAMI.2022.3184052
31 Kuncheva L I, Bezdek J C. Nearest prototype classification: clustering, genetic algorithms, or random search? IEEE Trans Syst Man Cybern C, 1998, 28: 160–164
32 Wei X-S, Xu S-L, Chen H, et al. Prototype-based classifier learning for long-tailed visual recognition. Sci China Inf Sci, 2022, 65: 160105
33 Yang H M, Zhang X Y, Yin F, et al. Convolutional prototype network for open set recognition. IEEE Trans Pattern Anal Mach Intell, 2020, 44: 2358–2370.
34 Chen G, Qiao L, Shi Y, et al. Learning open set network with discriminative reciprocal points. In: Proceedings of the European Conference on Computer Vision, 2020. 507–522
35 Chen G, Peng P, Wang X, et al. Adversarial reciprocal points learning for open set recognition. IEEE Trans Pattern Anal Mach Intell, 2021, 44: 8065–8081
36 Miller D, Sunderhauf N, Milford M, et al. Class anchor clustering: a loss for distance-based open set recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. 3570–3578
37 Gu Y F, Liu T Z, Gao G M, et al. Multimodal hyperspectral remote sensing: an overview and perspective. Sci China Inf Sci, 2021, 64: 121301
38 Feng D, Haase-Schutz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. IEEE Trans Intell Transp Syst, 2020, 22: 1341–1360
39 Song Q, Sun B, Li S. Multimodal sparse transformer network for audio-visual speech recognition. IEEE Trans Neural Netw Learn Syst, 2023, 34: 10028–10038
40 Ding C, Tao D. Robust face recognition via multimodal deep face representation. IEEE Trans Multimedia, 2015, 17: 2049–2058
41 Strese M, Schuwerk C, Iepure A, et al. Multimodal feature-based surface material classification. IEEE Trans Haptics, 2016, 10: 226–239
42 Wang X, Kumar D, Thome N, et al. Recipe recognition with large multimodal food dataset. In: Proceedings of IEEE International Conference on Multimedia & Expo Workshops, 2015. 1–6
43 Zheng H, Fang L, Ji M, et al. Deep learning for surface material classification using haptic and visual information. IEEE Trans Multimedia, 2016, 18: 2407–2416
44 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 4171–4186
45 Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision, 2016. 499–515
46 Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 815–823
47 Yu B, Tao D. Deep metric learning with tuplet margin loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 6490–6499