Pattern Recognition 172 (2026) 112566

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Contrastive learning with auxiliary model

Linyi Xu [a], Patrick P.K. Chan [b,*], Jingwen Deng [c], Xuanming Liang [c], Daniel S. Yeung [d]

[a] School of Computer Science and Engineering, South China University of Technology, Guangzhou, 511442, China
[b] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, 511442, China
[c] School of Future Technology, South China University of Technology, Guangzhou, 511442, China
[d] Hong Kong, China

ARTICLE INFO

Keywords:
Contrastive learning
False negative sample identification
Auxiliary model
Self-supervised learning

ABSTRACT

Contrastive Learning (CL) achieves excellent performance to reduce the requirements of annotation in many domains recently. Most CL methods treat all samples other than an anchor as negative samples and move them far away in training. However, these samples may be semantically similar and should belong to the same class. Recent false negative sample identification methods only rely on a unique embedding space, and the decisions may be biased. This study proposes a false negative sample identification method with an auxiliary model shared with the backbone of the main CL model in order to provide an additional point of view on the decisions. The auxiliary model is trained by minimizing the standard contrastive loss while maximizing diversity with the main model. False negative samples are identified by the fusion of the similarity measures in both feature spaces learned by the two models. The standard datasets are used to evaluate the effectiveness of our method. Experimental results confirm that our method achieves more reliable and accurate performance compared with the state-of-the-art CL methods. Using auxiliary models improves the false negative sample identification significantly. Our proposed false negative sample identification method improves the accuracy and stability of contrastive learning. Code is available at https://github.com/czzerone/Dual-Contrastive-Learning.

1. Introduction

The excellent performance of Deep Learning (DL) closely relates to the volume and quality of annotated samples [1,2]. However, the data collection and annotation are expensive and time-consuming, especially in medical related applications [3–5]. In order to reduce the demand of DL on the samples, the contrastive learning (CL) [6–8], which is one of the self-supervised learning models, aims to pre-train a model for supervised learning tasks without label information. Many studies confirm the excellent performance of the model pre-trained by CL in various applications [9–12]. CL follows the main idea of metric learning [13] which minimizes the distance between samples in the same type (positive samples) while maximizing samples of different types (negative samples). Different from metric learning, due to lack of annotation, for each sample, its augmented sample is treated as the positive sample while all other samples are the negative samples in CL [6,14,15]. One drawback of CL is that some negative samples may be semantically similar to the anchor sample, called false negative samples. Moving these samples far away from the anchor sample will discard the semantic relationship and obstruct learning convergence.

The issue of false negative sample identification, where semantically similar samples are misclassified as negatives, has become a significant focus in CL. Current methods to address this problem can be divided into two main categories: grouping-based and similarity-based approaches. Grouping-based methods, such as InterCLR [16] and IFND [17], utilize clustering techniques to identify false negatives. InterCLR applies k-means clustering to construct contrastive pairs, while IFND identifies false negatives as samples within the same cluster that have high confidence. On the other hand, similarity-based methods like Ring [18] and False Negative Cancellation (FNC) [19] rely on cosine similarity to assess the relationships between samples within the embedding space. However, both approaches are heavily dependent on the embedding space learned by the CL model, which is often suboptimal and biased. This is because the CL framework provides only minimal guidance and lacks label supervision, limiting its ability to effectively steer the learning process. Studies [20,21] also show that deep learning models tend

* Corresponding author.
E-mail addresses: czzerone@qq.com (L. Xu), patrickchan@ieee.org (P.P.K. Chan), jingwen_deng@foxmail.com (J. Deng), liangxuanmingcj@163.com (X. Liang), danyeung@ieee.org (D.S. Yeung).

https://doi.org/10.1016/j.patcog.2025.112566
Received 30 September 2024; Received in revised form 8 June 2025; Accepted 3 October 2025
Available online 8 November 2025
0031-3203/

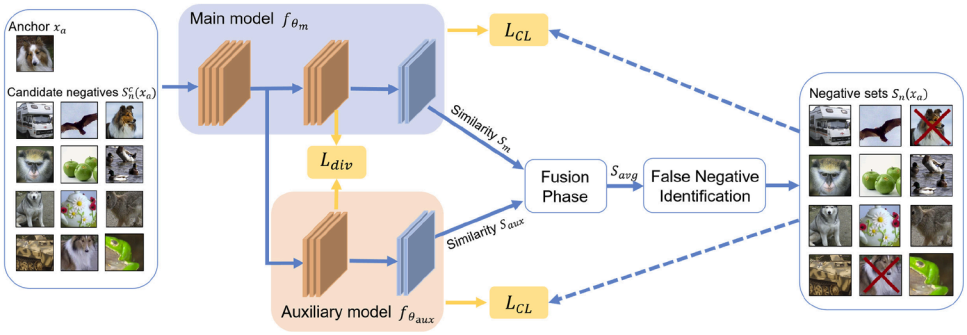


Fig. 1. Overview of the proposed method framework.

to capture incomplete representations of the full data distribution due to the randomness inherent in training. When the embedding space is inaccurate, both grouping and similarity measures become unreliable, making false negative identification unstable and error-prone. These limitations emphasize the need for more robust methods for false negative sample identification.

In order to avoid the domination of the embedding space, this study aims to identify the false negative samples according to the decisions of diverse multi-models. Besides the model in CL, an auxiliary model is trained aiming to minimize the CL loss and also maximize the diversity of all models in our method to guarantee that different and desirable high-dimension feature spaces are obtained. The similarity between samples is measured according to embedding spaces of the main and auxiliary models separately and the final decision is made by fusing all these results. It is expected that the bias of false negative sample identification caused by the unique embedding space is reduced in our method. Our proposed method is then evaluated and compared with state-of-the-art (SOTA) methods empirically in the standard dataset.

The contributions of the proposed method are summarized as follows:

- To address the bias in false negative sample identification caused by relying solely on a single embedding space in contrastive learning (CL), our method incorporates an auxiliary model alongside the main CL model. By sharing the backbone of the main model, the auxiliary model maintains low complexity and computational cost.
- A diversity loss is introduced to ensure the auxiliary model learns a feature space distinct from that of the main model. Combined with the standard CL loss, this enables the auxiliary model to provide complementary knowledge to the main model, enhancing the overall performance of the CL framework.
- An average fusion strategy is proposed to improve false negative sample identification. This strategy combines similarity measures from the feature spaces of both the main and auxiliary models, leading to more robust and accurate decisions.
- Extensive experiments validate the effectiveness of our proposed method. Results demonstrate a significant improvement in false negative sample identification accuracy when using the auxiliary model, as well as superior performance of our CL framework compared to state-of-the-art (SOTA) CL methods.

The rest of the paper is organized as follows. The related concepts of our study are briefly introduced in Section 2. Section 3 devises and discusses our proposed model in detail, and the experimental results and discussion are given in Section 4. The conclusion is finally given and future work is discussed in Section 5.

2. Related work

The related concepts of this study including contrastive learning and its negative sampling methods are briefly introduced in this section. A DL model achieves excellent performance in a wide range of applications. However, one of its problems is the annotation of the training set, which is time-consuming and expensive [5,22], especially the training set of DL is usually huge. Recently, contrastive learning (CL) [6,15], which is one of the self-supervised learning frameworks, has drawn attention. A model is updated without relying on label information by moving similar samples close together and pushing away dissimilar samples in CL.

Since the annotation is not available, the studies of CL mainly focus on the formulation of positive and negative sample pairs. Some works design effective data transformation to construct positive samples, *e.g.,* ContrastiveCrop [23] and InfoMin Aug. [24]. Hard positive samples are generated by using data mixing or adversarial samples, Un-mix [25], CsMI [26] and *e.g.,* CLAE [27]. However, only a few research investigates on the formulation of negative sample pairs. Most methods, *e.g.,* SimCLR [15] and MoCo [6], treat all samples except the anchor sample in a mini-batch and dataset as negative samples. However, treating all samples as negative samples may mislead the training since some samples are semantically similar to the anchor. The learned representation may discard semantic information contained in samples. As a result, some works focus on the identification of false negative samples defined as samples which share similar semantic information with the anchor. DCL [28] reduces the influence of false negative samples by adjusting the weights of the positive and negative samples. False negative samples are removed by analyzing the structure of data, *e.g.,* InterCLR [16] and IFND [17]. Similarity measures, FNC [19] and Ring [18], are proposed to quantify the samples more precisely. Moreover, several methods have leveraged intra-class relationships to enhance contrastive learning. For example, CVC [29] integrates clustering mechanisms into contrastive learning by grouping high-confidence instances and refining low-confidence assignments, thereby improving the quality of the learned representation space.

3. Proposed method

The performance of contrastive learning may decrease by using false negative samples in learning. Most current methods identify the false negative samples by structure analysis with a similarity measure. However, the decisions made relying on the unique embedding space are easily misled and biased since the embedding space may only capture the partial feature information. Our study proposes a false negative sample identification method according to the decisions of diverse models. In addition to the main CL model (Section 3.1), an auxiliary model (Section 3.2) is trained by minimizing the CL loss and maximizing the diversity to the main model in order to provide a different angle of view on the sample similarity. The false negative samples are finally determined by combining all decisions (Section 3.3). The framework of our method is shown in Fig. 1.

3.1. Main model

A main CL model f_{θ_m} is the key component in our framework. f_{θ_m} includes a feature encoder, *e.g.,* the ResNet network, to extract the feature from the data and a multi-layer perceptron (MLP) as the projection head to map the extracted feature into the embedding space. f_{θ_m} is trained by minimizing the standard contrastive loss, *i.e.,* the InfoNCE loss [30]. Given an anchor x_a, positive sample x_p and its negative set $S_n(x_a)$, θ_m is optimized by:

$$L_{CL}(x_a, x_p, S_n(x_a), \theta_m) = -log \frac{e^{(f_{\theta_m}(x_a)\cdot f_{\theta_m}(x_p)/\tau)}}{e^{(f_{\theta_m}(x_a)\cdot f_{\theta_m}(x_p)/\tau)} + \sum_{x_n \in S_n(x_a)} e^{(f_{\theta_m}(x_a)\cdot f_{\theta_m}(x_n)/\tau)}} \quad (1)$$

where $\cdot$ represents the dot product and τ denotes the temperature parameter. The positive sample x_p is generated by applying another augmentation on the same image as the anchor. The negative set $S_n(x_a)$ is obtained by the latest batches of data to form a dynamic queue [6] after discarding some false negative samples, which will be discussed in Section 3.3 in detail. The trained main model is expected to learn a good semantic-aware representation.

3.2. Auxiliary model

In addition to f_{θ_m}, an auxiliary model $f_{\theta_{aux}}$ is also trained simultaneously. To reduce the training complexity, the low-layer feature encoder of f_{θ_m} is shared with $f_{\theta_{aux}}$. It should be noted that while the auxiliary and primary models share the same encoder architecture, their parameters are updated independently during training. Gradients are not propagated between the two models, ensuring that the auxiliary model learns a distinct feature space.

The purpose of the auxiliary model is to provide complementary information to the main model, helping to mitigate biases in the learning process. Therefore, the auxiliary model must also possess the essential ability to measure similarity between samples. The contrastive loss (L_{CL}), introduced for the main model in Section 3.1 and defined in Eq. (1), is applied to the auxiliary model as well. In addition, the auxiliary model should also learn different knowledge to the main model, guaranteed by a diversity loss (L_{div}) is incorporated during training. Negative Correlation Learning (NCL) [31] is employed as the diversity loss, defined as follows:

$$L_{div}(x_a) = ||f_{\theta_{aux}}(x_a) - \left(\frac{f_{\theta_m}(x_a) + f_{\theta_{aux}}(x_a)}{2}\right)||^2 \quad (2)$$

where $f_{\theta_m}(x_a)$ and $f_{\theta_{aux}}(x_a)$ represent the features of x_a extracted by the main and auxiliary models, respectively. The diversity loss L_{div} quantifies the correlation between the outputs of the two models. A larger L_{div} value indicates that the auxiliary model's feature space is more distinct from the main model's, while a smaller value implies greater similarity. This design ensures that the auxiliary model contributes unique and complementary insights to enhance the overall learning process.

As a result, the auxiliary model's total loss comprises two components and is defined as:

$$L_{aux}(x_a) = L_{CL}(x_a, x_p, S_n(x_a), \theta_{aux}) - \lambda L_{div}(x_a) \quad (3)$$

where λ denotes the trade-off parameter, controlling the influence of the diversity loss on the auxiliary model's training objectives. This parameter ranges from 0 to 1, with larger values encouraging greater divergence between the feature spaces learned by the auxiliary and main models. A small value for λ is recommended, as the auxiliary model's performance primarily relies on minimizing the standard contrastive loss, while the diversity loss acts as a supplementary term to encourage the auxiliary model to learn a feature space distinct from that of the main model. Setting λ too large forces the auxiliary model to learn features that deviate excessively from the main model's, potentially reducing accuracy by steering the auxiliary model away from optimal feature representations.

3.3. False negative samples identification

Our method detects the false negative samples from the negative set by combining decisions of diverse and well-performed models in order to reduce the bias of using a unique embedding space.

Given an anchor x_a, its candidate negative sample set $S_n^c(x_a)$ containing M samples randomly selected from the training set [6]. In each update, the similarity score $s_{\theta_i}(x_a, x)$ between x_a and each sample x in $S_n^c(x_a)$ in the embedding space generated by a model θ_i based on the cosine similarity:

$$s_{\theta_i}(x_a, x) = cosine(f_{\theta_i}(x_a), f_{\theta_i}(x)),\ i = m, aux \text{ and } x \in S_n^c(x_a) \quad (4)$$

The range of the cosine similarity score is from -1 to 1. A sample pair with a larger similarity score means the samples are more similar. Combining similarity evaluation in different feature spaces increases reliability and reduces the bias of estimation. An average fusion strategy is applied to aggregate the similarity scores of samples in $S_n^c(x_a)$, denoted as $s_{x_a}(x)$:

$$s_{x_a}(x) = \frac{1}{2}(s_{\theta_m}(x_a, x) + s_{\theta_{aux}}(x_a, x)),\quad x \in S_n^c(x_a) \quad (5)$$

The samples with the top K average similarity scores are grouped as the set of false negative samples ($S_{FN}(x_a)$) since they are closest to the anchor in most embedding spaces. The negative set $S_n(x_a)$ is revised by subtracting $S_n^c(x_a)$ by $S_{FN}(x_a)$:

$$S_n(x_a) = S_n^c(x_a) - S_{FN}(x_a) \quad (6)$$

Our method removes the detected false negatives but does not treat them as positive samples. This decision stems from the fact that the main model and auxiliary model are trained without strong label supervision, making its identification of false negatives potentially unreliable. Misclassifying false negatives (i.e., true negatives) as positives in the contrastive learning framework, where each anchor is associated with only one positive sample, would introduce incorrect positive information, significantly disrupting the learning process. On the other hand, removing a small number of true negatives from the extensive negative set has only a minimal impact on overall performance.

4. Experiment

4.1. Experimental settings

We learn the representation across different scales datasets: CIFAR10 [32], CIFAR100 [32], STL10 [33], Tiny-ImageNet [34] and ImageNet-100 [35] to validate the efficiency of our method. CIFAR10 with 10 classes contains 50,000 samples for training and 10,000 for testing and the size of each sample is 32×32, while CIFAR100 with 100 classes contains 50,000 samples for training and 10,000 for testing and the size of each sample is 32×32. STL10 contains 100,000 unlabeled samples and 13,000 labeled samples from 10 classes. 5000 labeled samples and all unlabeled samples are used for training and the size of each sample is 64×64. Tiny-ImageNet with 200 classes contains 100,000 samples for training and 10,000 for testing. The size of each sample is 64×64. ImageNet-100 with 100 classes contains 126,689 and 5000 samples in the train and test sets, and the size of each sample is 224×224.

A popular contrastive learning framework, *i.e.,* MoCo [6], is chosen as the baseline by following the settings of current CL studies [36,37]. The same general settings are used in all experiments for fair comparison. The model of MoCo contains the ResNet-18 network as the encoder followed by a 2-layer Multi-Layer Perceptron (MLP) as the projector. The dimension of the embedding vector is 128. The learning rate is set as 0.001 in the Adam optimizer and the temperature is $\tau = 0.07$. Each experiment is trained for 300 epochs and the batch-size is 256.

In addition to our proposed false negative sample identification method, the state-of-the-art methods, *i.e.,* Ring [18], FNC [19], and CVC [29], are also considered. A number of preliminary experiments have

Table 1
Average and standard deviation of the top-1 classification accuracy of MoCo, Ring, FNC, CVC and our method in different datasets in linear evaluation.

Methods	CIFAR10	CIFAR100	STL10	Tiny-ImageNet
MoCo	81.12 (±0.28)	56.82 (±0.50)	80.67 (±0.20)	43.56 (±0.30)
Ring	82.54 (±0.30)	57.43 (±0.46)	81.26 (±0.51)	44.58 (±0.32)
FNC	83.34 (±0.21)	57.99 (±0.23)	81.59 (±0.20)	44.34 (±0.21)
CVC	74.75 (±0.31)	47.37 (±0.25)	74.19 (±0.28)	38.32 (±0.29)
Ours	**85.85 (±0.17)**	**58.86 (±0.13)**	**82.45 (±0.19)**	**45.05 (±0.14)**

been carried out to decide suitable parameters for all methods. For Ring, the outer threshold anneals from 1 to 0.1 and the anneal epoch is set to 100. The inner threshold is set to 10 % of the outer threshold as in [18] during removing FN samples. The size of support set is 8, the top-k and threshold are set to 10 and 0.7 respectively as in [19] for detecting FN samples for FNC. To ensure a fair comparison, the original Vision Transformer (ViT) backbone in CVC was replaced with a ResNet-18 architecture, while all other settings were kept unchanged. The dimension of the embedding vector and tradeoff parameter λ of the auxiliary model in our method is set to 64 and 0.01. The K is set to 40 when identifying the FN samples. The size of the negative queue (*i.e.,* the candidate negative set in our method) is set to 65,536 with a momentum of 0.999. All experiments are conducted on a Pytorch framework running on a personal computer with NVIDIA GTX 1080.

4.2. Comparison with SOTA methods

4.2.1. Linear evaluation

The quality of the representation learned by different models is evaluated by using the same task of CL learning. A fully connected (FC) layer is concatenated to the encoder of CL model to form a classifier. The parameters of the FC layer are updated by the same dataset used in the contrastive learning phase in the framework of supervised learning. Each experiment is repeated 3 times independently using different random seeds. The performance of MoCo, Ring, FNC and our method in terms of the average and standard deviation of the top-1 classification accuracy on the datasets are reported in Table 1.

The average accuracy of all methods with false negative sample identification performs consistently better than MoCo, which confirms that discarding false negative samples plays an important role in contrastive learning. Our method performs the best among all methods and achieves 1.6 % and 1.17 % higher accuracy than Ring and FNC respectively on average. CVC consistently underperforms across all datasets, which may be due to its original design for transformer architectures rather than CNNs. Its clustering-based objective relies on confident global embeddings, but CNNs tend to produce locally focused features, leading to noisy clustering and reduced learning efficiency. Moreover, our method is also the most stable, *i.e.,* its standard deviation values are always smaller than other methods. The results indicate that identifying false negative samples according to diverse feature spaces enhances the accuracy and stability of identification results.

4.2.2. Feature visualization

The performance of the proposed model is compared against MoCo, Ring, and FNC through feature visualization. Using the t-SNE method, samples from the test set of the CIFAR-10 dataset are mapped into a two-dimensional space. The visualization results are presented in Fig. 2, where points of different colors represent samples from different classes.

The feature spaces generated by Ring, FNC, and our proposed method exhibited greater class separability compared to the baseline MoCo, indicating the effectiveness of the modifications applied to the original framework. However, the feature spaces produced by Ring and FNC demonstrated less separability compared to our method. Specifically, the boundaries of most classes, except the classes in cyan, yellow and orange, are not clear in Ring and FNC. In contrast, all classes are more distinctly separated with clear boundaries in the feature space learnt by our method. This improved class separability corresponds to the results presented in Table 1 and provides an intuitive understanding of the comparative effectiveness of the methods. These results confirm the conclusion that the proposed method outperforms the baseline and SOTA methods in capturing more representative feature spaces.

4.2.3. Transfer evaluation

The generalization ability of the models trained by the different CL methods is discussed in this section. An FC layer concatenated with the models pre-trained by MoCo, Ring, FNC, and our method on ImageNet-100 is updated by other six datasets, including CIFAR10, CIFAR100, Caltech-101 [38] , DTD [39], Pets [40] and Flowers [41]. The results are shown in Table 2.

The results are consistently with the ones of the linear evaluation in the previous section. Our method performs the best consistently in all datasets and achieves 2.91 %, 2.01 % and 1.29 % higher accuracy than MoCo, Ring and FNC respectively in average. The necessary of effective false negative sample identification is also confirmed, *i.e.,* the performance of MoCo is the worst. The feature representation trained by our model has the greatest generalization ability comparing with other methods.

4.2.4. Performance on different baselines

This section investigates the generalization capability of our method across different baseline frameworks. Two widely-used baselines, SimCLR [15] and PIRL [42], are considered in this evaluation. For SimCLR, the official implementation is utilized, with a ResNet-18 network serving as the encoder. The model is trained for 300 epochs with a batch size of 256, and K is set to 40. For PIRL, the implementation follows the specifications outlined in the original paper, using parameter settings consistent with those applied in the SimCLR experiments.

The top-1 classification accuracy for these methods across three datasets is reported in Table 3. Our method achieves an average improvement of 8.56 % over the baseline for SimCLR. Similarly, for PIRL, an identical increase of 10.21 % is observed in classification accuracy. These consistent improvements across both baselines underscore the robustness and adaptability of our approach. Moreover, the results highlight the method's capability to reduce bias and enhance the performance of contrastive learning tasks.

4.2.5. Running time

The training times of the contrastive models for 300 epochs are investigated on CIFAR10 are shown in Table 4. It is expected that our model requires the longest training time mainly caused by the training of the auxiliary model. The training time of FNC is much longer than MoCo and Ring since FNC needs to compute similarity scores between each pair of the negative and support samples for an anchor. It can be noted that the training time of MoCo and Ring are almost the same since only additional time for sorting similarity scores of all negative samples in Ring. Although the training time of our method is longer than other methods, the running complexity of the inference phase is identical to others since the complexity of the trained main models is the same.

4.3. Ablation study

4.3.1. Influence of auxiliary model

The performance of our model with varying numbers of auxiliary models is analyzed in this section. The additional auxiliary models are trained following the same approach described in Section 3.2. Specifically, these models share the feature encoder of the main model and are updated simultaneously using the auxiliary loss, as shown in Eq. (7). The diversity loss for a single auxiliary model, as defined in Eq. (2), is extended to accommodate multiple auxiliary models. This generalized diversity loss is calculated as the average of the diversity between each

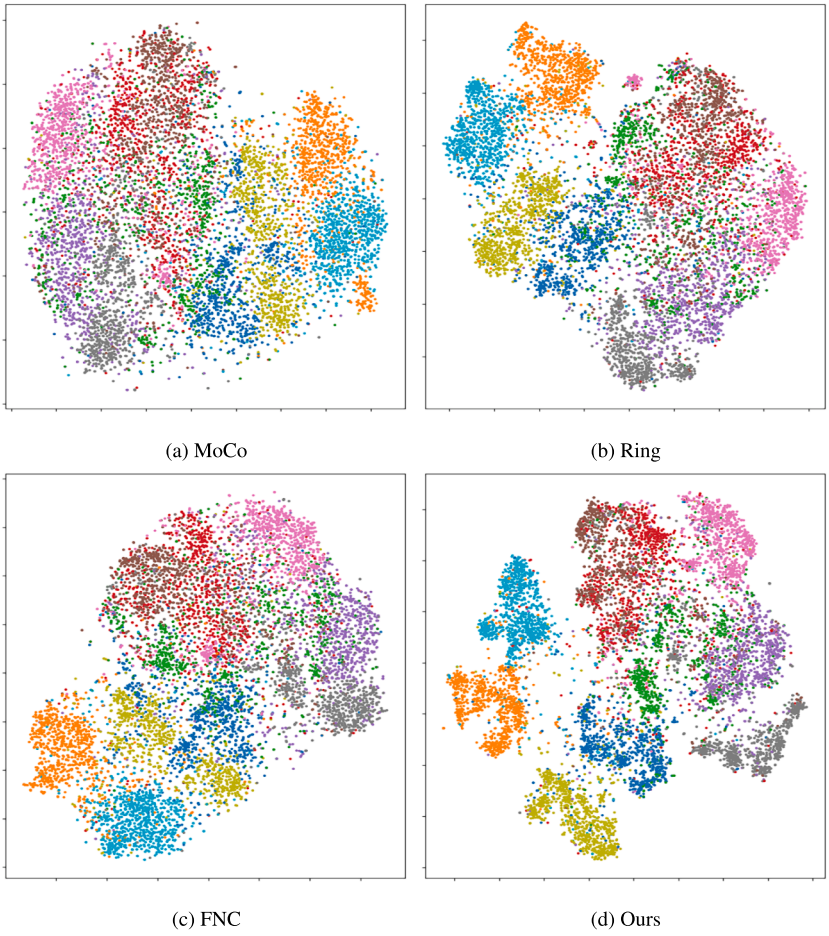


Fig. 2. Visualizations of MoCo, Ring, FNC and Ours on CIFAR10 by t-SNE.

Table 2
Average top-1 classification accuracy of MoCo, Ring, FNC and our method initialized by ImageNet-100 transferring to different datasets in transfer evaluation.

Methods	ImageNet100	CIFAR10	CIFAR100	Caltech-101	DTD	Pets	Flowers
MoCo	63.72	68.00	38.54	66.64	47.81	45.68	52.82
Ring	67.96	68.05	38.77	67.46	48.34	45.87	53.46
FNC	68.08	69.15	39.24	67.64	49.05	46.97	55.15
Ours	**69.62**	**69.33**	**40.87**	**68.15**	**49.65**	**47.56**	**58.42**

Table 3
The top-1 classification accuracy of SimCLR, PIRL and our method based on these two baselines respectively in different datasets in linear evaluation.

Methods	CIFAR10	CIFAR100	STL10
SimCLR	74.34	34.05	65.01
Ours(SimCLR)	**85.04**	**48.01**	**66.04**
PIRL	62.10	33.90	46.04
Ours(PIRL)	**81.52**	**35.06**	**56.08**

Table 4
The training time (in hour) on CIFAR10.

MoCo	Ring	FNC	Our
6.4	6.5	14.7	14.75

auxiliary model and the main model.

$$L_{aux}(x_a) = \sum_{i=1}^{m} \left(L_{CL}\left(x_a, x_p, S_n(x_a); \theta^i_{aux}\right) - \lambda L_{div}\left(x_a; \theta^i_{aux}\right) \right) \quad (7)$$

where θ^i_{aux} represents the i^{th} auxiliary model and m denotes the total number of auxiliary models are used. In this experiment, m is set to 0 (no auxiliary model is used), 1, 3 and 5. For consistency, the framework of our proposed model is retained in all settings, with the diversity loss applied only between each auxiliary model and the main model. To enhance the distinction among multiple auxiliary models, the final fully connected layers are configured with varying dimensions ranging from 64 to 256. This architectural variation encourages the learning of diverse feature spaces and reduces redundancy among representations.

The results are shown in Fig. 3. Adding the first auxiliary model to the main model contributes the most significantly. The benefit of auxiliary models to our method decreases with the increase of the number of auxiliary models. The accuracy of the model using five auxiliary models is even lower than the one with three auxiliary models. This relatively limited improvement is likely due to the increased difficulty of effectively extracting complementary and informative representations from multiple models. Under the limited supervision typical in contrastive learning, additional auxiliary models may fail to learn sufficiently diverse or meaningful features, thereby constraining their overall contribution.

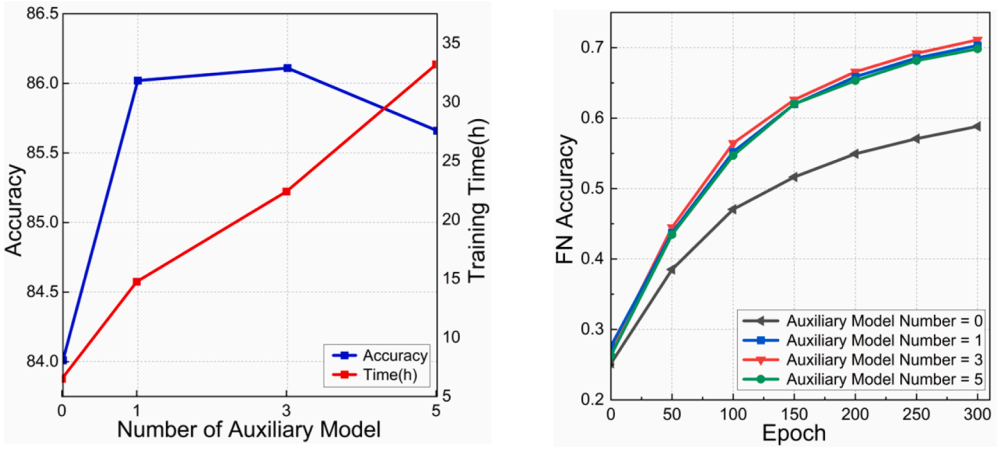
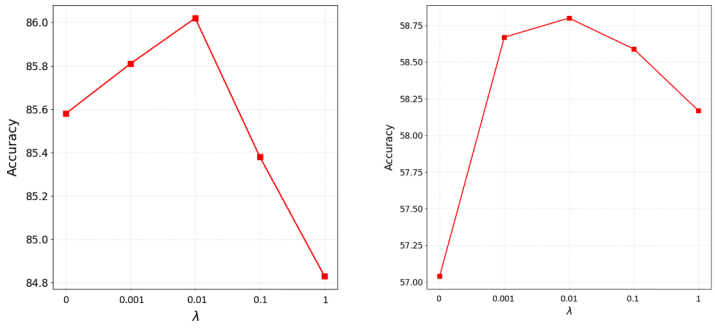


Fig. 3. Top-1 accuracy and training time (left), and false negative sample identification accuracy (right) of our method with different numbers of the auxiliary models on CIFAR10.



(a) CIFAR10 (b) CIFAR100

Fig. 4. Accuracy of our model with different λ values on CIFAR10 and CIFAR100.

The training times of different models for 300 epochs are also reported in Fig. 3 left. The training time of our method increases nonlinearly with the number of auxiliary models. The cost of adding an auxiliary model is less than training a complete main model since the feature encoder of the main model is shared as a backbone with the auxiliary models. In order to balance the performance and training complexity, one auxiliary model is suggested by our method.

In order to further understand the effectiveness of our proposed false negative sample identification method with different auxiliary models, their accuracy calculated by using the ground-true labels during the training is shown in Fig. 3 right. It should be noted that accuracy refers to the overall classification success rate on all test samples, while false negative sample identification accuracy measures the probability of correctly identifying false negative samples (i.e., samples belonging to the same class as the anchor) from the candidate negative set within a small training batch. Generally, the accuracy of our identification methods gradually raises during training. The performance of our methods with 1, 3 and 5 auxiliary models are similar, which confirms the previous finding, *i.e.,* additional auxiliary models are not helpful in training. Our methods significantly outperform the one without an auxiliary model (*i.e.,* num = 0). This explains why the performance of our contrastive learning method achieves excellent performance.

4.3.2. *Influence of the tradeoff parameter*

The parameter λ of our model controls the importance of the diversity between the auxiliary model from the main model and the standard contrastive loss in training. The accuracy of our models with different λ values on CIFAR10 and CIFAR100 are shown in Fig. 4.

From both results, a consistent trend is observed in how classification accuracy changes with respect to λ. For example, on the CIFAR10 dataset, the performance of our model rises from 85.58 % when λ increases. It reaches its peak, *i.e.,* 86.02 %, when $\lambda = 0.01$. The accuracy drops after λ is larger than 0.01. When $\lambda = 0$, the auxiliary and main models are optimized independently. As a result, they may learn some similar feature information, and the feature space may be biased. On the other hand, the diversity loss dominates the learning when $\lambda = 1$. The auxiliary model will be different from the main model but may not learn the data well. These explain the models with two extreme settings perform badly.

4.3.3. *Influence of the number of subtracted FN samples*

Samples with the top K average similarity scores are identified as false negatives and removed from the negative sets in our proposed method. To analyze the impact of the K parameter, our models trained with varying K values are evaluated on the CIFAR10 and CIFAR100 datasets in this section. The results are shown in Fig. 5.

The results indicate that the accuracy improves significantly as K increases from 0 to 40. Specifically, the accuracy rises from 77.66 % and 50.16 % at $K = 0$ on CIFAR10 and CIFAR100 respectively, where no false negative identification is performed. In this case, some false negative samples may be misidentified as true negatives, negatively impacting the learning. Accuracy reaches its peak at 83.67 % and 58.04 % respectively when $K = 40$, suggesting that increasing K enhances the ability of the model to learn more accurate feature representations. However, when K is further increased to 70, a slight decline in accuracy is observed. This decrease implies that excessively large K values

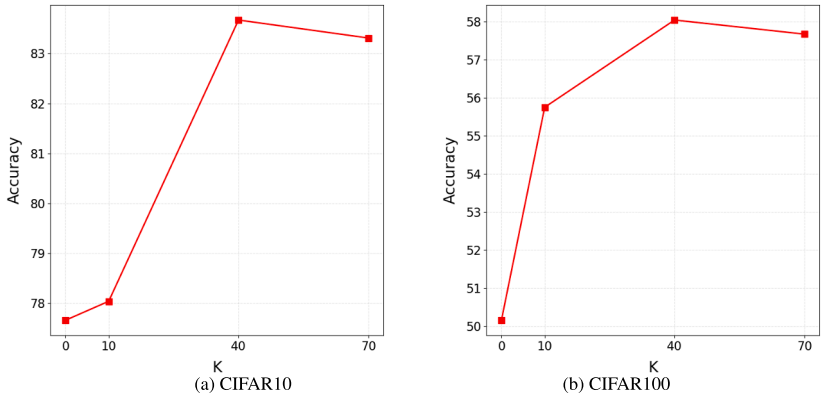


Fig. 5. Accuracy of our model with varying K values on CIFAR10 and CIFAR100.

may unintentionally remove true negative samples, thereby reducing the amount of useful information available for learning. These findings demonstrate the existence of an optimal range for K, beyond which the performance is negatively affected.

5. Conclusion

This study aims to improve the false negative sample identification in CL by reducing the bias of the identification under a unique embedding space. An auxiliary model shared with the backbone of the main CL method is considered additionally in our model. To provide reliable and different feature space to the main model, the auxiliary model is trained by minimizing the standard CL loss and maximizing the diversity. The false negative samples are finally determined by the similarity measures in the feature spaces of the main and auxiliary models. The experimental results suggest that our CL method achieves better performance than the SOTA methods in terms of accuracy and stability. Moreover, the accuracy of false negative sample identification by using an auxiliary model is significantly higher than the base model. The results also indicate that using more auxiliary models cannot improve the accuracy of the main CL model. It may be because it is harder to train accurate and diverse auxiliary models when more auxiliary models are used. The proposed false negative samples identification method in this study improves the quality of the negative set and enhances the performance of contrastive learning.

One future work is to adjust of the number of false negative samples for each anchor dynamically in training. At the beginning of training, the main and auxiliary models may not learn well, and their decisions on false negative samples may not be reliable. The number of false negative samples increases with the training epoch. Another future work can address the limitations of the top-K strategy, which may include true negatives in low-class-count settings. While fixed-threshold methods may offer better adaptability, they require careful tuning and are unstable early in training. A promising direction is to explore hybrid or adaptive strategies that combine the strengths of both approaches.

CRediT authorship contribution statement

Linyi Xu: Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization; **Patrick P.K. Chan:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization; **Jingwen Deng:** Writing – review & editing, Visualization, Data curation; **Xuanming Liang:** Writing – review & editing, Validation, Software; **Daniel S. Yeung:** Supervision, Resources, Project administration, Investigation.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Yu, Q. Dai, Self-supervised multi-task learning for medical image analysis, Pattern Recogn. 150 (2024) 110327.
[2] X. Yu, S. Sun, Y. Tian, Self-distillation and self-supervision for partial label learning, Pattern Recogn. 146 (2024) 110016.
[3] C. Tang, X. Zeng, L. Zhou, Q. Zhou, P. Wang, X. Wu, H. Ren, J. Zhou, Y. Wang, Semi-supervised medical image segmentation via hard positives oriented contrastive learning, Pattern Recogn. 146 (2024) 110020.
[4] F. Zheng, J. Cao, W. Yu, Z. Chen, N. Xiao, Y. Lu, Exploring low-resource medical image classification with weakly supervised prompt learning, Pattern Recogn. 149 (2024) 110250.
[5] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, Y. Shi, Positional contrastive learning for volumetric medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 221–230.
[6] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[7] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.
[8] J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.
[9] X. Shu, B. Xu, L. Zhang, J. Tang, Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2022) 7559–7576.
[10] Y. Wu, D. Zeng, Z. Wang, Y. Shi, J. Hu, Distributed contrastive learning for medical image segmentation, Med. Image Anal. 81 (2022) 102564.
[11] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, J. Zhang, B. Ding, B. Cui, Contrastive learning for sequential recommendation, in: 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, 2022, pp. 1259–1273.
[12] S. Lin, C. Liu, P. Zhou, Z.Y. Hu, S. Wang, R. Zhao, Y. Zheng, L. Lin, E. Xing, X. Liang, Prototypical graph contrastive learning, IEEE Trans. Neural Netw. Learn. Syst. 35 (2022) 2747–2758.
[13] E. Xing, M. Jordan, S.J. Russell, A. Ng, Distance metric learning with application to clustering with side-information, Adv. Neural Inf. Process. Syst. 15 (2002) 521–528.
[14] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.
[15] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[16] J. Xie, X. Zhan, Z. Liu, Y.S. Ong, C.C. Loy, Delving into inter-image invariance for unsupervised visual representations, Int. J. Comput. Vis. 130 (12) (2022) 2994–3013.

[17] T.S. Chen, W.C. Hung, H.Y. Tseng, S.Y. Chien, M.H. Yang, Incremental false negative detection for contrastive learning, in: International Conference on Learning Representations, 2022.
[18] M. Wu, M. Mosse, C. Zhuang, D. Yamins, N. Goodman, Conditional negative sampling for contrastive learning of visual representations, in: International Conference on Learning Representations, 2021.
[19] T. Huynh, S. Kornblith, M.R. Walter, M. Maire, M. Khademi, Boosting contrastive self-supervised learning with false negative cancellation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2785–2795.
[20] H.F. Langroudi, C. Merkel, H. Syed, D. Kudithipudi, Exploiting randomness in deep learning algorithms, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.
[21] F. Huang, G. Xie, R. Xiao, Research on ensemble learning, in: 2009 International Conference on Artificial Intelligence and Computational Intelligence, 3, IEEE, 2009, pp. 249–252.
[22] Q. Huang, Y. Zhou, L. Tao, W. Yu, Y. Zhang, L. Huo, Z. He, A Chan-Vese model based on the Markov chain for unsupervised medical image segmentation, Tsinghua Sci. Technol. 26 (6) (2021) 833–844.
[23] X. Peng, K. Wang, Z. Zhu, M. Wang, Y. You, Crafting better contrastive views for siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16031–16040.
[24] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, P. Isola, What makes for good views for contrastive learning? Adv. Neural Inf. Process. Syst. 33 (2020) 6827–6839.
[25] Z. Shen, Z. Liu, Z. Liu, M. Savvides, T. Darrell, E. Xing, Un-mix: rethinking image mixtures for unsupervised visual representation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 36, 2022, pp. 2216–2224.
[26] H. Xu, X. Zhang, H. Li, L. Xie, W. Dai, H. Xiong, Q. Tian, Seed the views: hierarchical semantic alignment for contrastive representation learning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 3753–3767.
[27] C.H. Ho, N. Nvasconcelos, Contrastive learning with adversarial examples, Adv. Neural Inf. Process. Syst. 33 (2020) 17081–17093.
[28] C.Y. Chuang, J. Robinson, Y.C. Lin, A. Torralba, S. Jegelka, Debiased contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 8765–8775.
[29] Y. Liu, X. Zan, X. Li, W. Liu, G. Fang, Contrastive visual clustering for improving instance-level contrastive learning as a plugin, Pattern Recogn. 154 (2024) 110631.
[30] O.A. van den, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, (2018). arXiv preprint arXiv:1807.03748.
[31] Y. Liu, X. Yao, Ensemble learning via negative correlation, Neural Netw. 12 (10) (1999) 1399–1404.
[32] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).
[33] A. Coates, A. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.
[34] Y. Le, X. Yang, Tiny imagenet visual recognition challenge, CS 231N 7 (7) (2015) 3.
[35] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: European Conference on Computer Vision, Springer, 2020, pp. 776–794.
[36] C.H. Yeh, C.Y. Hong, Y.C. Hsu, T.L. Liu, Y. Chen, Y. LeCun, Decoupled contrastive learning, in: European Conference on Computer Vision, Springer, 2022, pp. 668–684.
[37] C. Zhang, K. Zhang, T.X. Pham, A. Niu, Z. Qiao, C.D. Yoo, I.S. Kweon, Dual temperature helps contrastive learning without many negative samples: towards understanding and simplifying moco, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14441–14450.
[38] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, in: 2004 Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2004, pp. 178.
[39] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3606–3613.
[40] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and dogs, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3498–3505.
[41] M.E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.
[42] I. Misra, M.L. van der, Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6707–6717.

Linyi Xu received B.Sc. degree in Computer Science and Technology from South China University of Technology, Guangzhou, China in 2020. She is currently a master student of School of Computer Science and Engineering, and a member of Machine Learning and Cybernetics Research Laboratory in South China University of Technology, Guangzhou, China. Her current research interests include deep learning, images recognition and computer vision.

Patrick P. K. Chan received his Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2009. He is currently Deputy Dean and Associate Professor at the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, China. His research interests include machine learning security, adversarial learning, and deep learning. Dr. Chan serves as Associate Editor-in-Chief of the IEEE Transactions on Cybernetics and Information Sciences.

Jingwen Deng received the bachelor's degree from South China University of Technology in 2023. She is currently studying for the M.S. degree in the School of Future Technology, South China University of Technology. Her research interests include deep learning, image generation and image editing.

Xuanming Liang obtained his bachelor's degree from Nanchang University in 2024. He is currently pursuing a master's degree at the School of Future Technology, South China University of Technology. His research interests are focused on deep learning and large language models.

Daniel S. Yeung received the Ph.D. degree from Case Western Reserve University, Cleveland, OH, USA. He was a Faculty Member with the Rochester Institute of Technology, Rochester, NY, USA, from 1974 to 1980. In the next ten years, he held industrial and business positions in the USA. In 1989, he joined the Department of Computer Science, City Polytechnic of Hong Kong, Hong Kong, as an Associate Head/Principal Lecturer. He served as the Founding Head and Chair Professor with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, until his retirement in 2006. His research interests include neural network sensitivity analysis, large-scale data retrieval problem, and cyber security.