sensors

MDPI

Article

Multi-Agent Multi-View Collaborative Perception Based on Semi-Supervised Online Evolutive Learning

Di Li [1] and Liang Song [2,*]

[1] College of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China
[2] Academy for Engineering & Technology, Fudan University, Shanghai 200433, China
* Correspondence: songl@fudan.edu.cn

Abstract: In the edge intelligence environment, multiple sensing devices perceive and recognize the current scene in real time to provide specific user services. However, the generalizability of the fixed recognition model will gradually weaken due to the time-varying perception scene. To ensure the stability of the perception and recognition service, each edge model/agent needs to continuously learn from the new perception data unassisted to adapt to the perception environment changes and jointly build the online evolutive learning (OEL) system. The generalization degradation problem can be addressed by deploying the semi-supervised learning (SSL) method on multi-view agents and continuously tuning each discriminative model by collaborative perception. This paper proposes a multi-view agent's collaborative perception (MACP) semi-supervised online evolutive learning method. First, each view model will be initialized based on self-supervised learning methods, and each initialized model can learn differentiated feature-extraction patterns with certain discriminative independence. Then, through the discriminative information fusion of multi-view model predictions on the unlabeled perceptual data, reliable pseudo-labels are obtained for the consistency regularization process of SSL. Moreover, we introduce additional critical parameter constraints to continuously improve the discriminative independence of each view model during training. We compare our method with multiple representative multi-model and single-model SSL methods on various benchmarks. Experimental results show the superiority of the MACP in terms of convergence efficiency and performance. Meanwhile, we construct an ideal multi-view experiment to demonstrate the application potential of MACP in practical perception scenarios.

Keywords: semi-supervised learning; online evolutive learning; collaborative perception; discriminative information fusion

Citation: Li, D.; Song, L. Multi-Agent Multi-View Collaborative Perception Based on Semi-Supervised Online Evolutive Learning. *Sensors* **2022**, *22*, 6893. https://doi.org/10.3390/s22186893

Academic Editors: Dawid Połap, Robertas Damasevicius and Hafiz Tayyab Rauf

Received: 26 July 2022
Accepted: 8 September 2022
Published: 13 September 2022

1. Introduction

In the edge intelligence [1–3] environment, many sensing devices recognize the local scene to provide corresponding smart services in real-time. However, most current intelligent sensing applications still rely on a fixed recognition model or a unified cloud model [4]. Since the perception scene changes over time, the feature distribution of the perception data will continue changing, and the generalizability of the fixed recognition model will be highly affected. The degradation of model generalizability will significantly impact the service quality of edge agents/models. Relying on regular manual annotations for model tuning will incur high ongoing costs and increase the deployment difficulty.

To reduce manual annotation and continuously improve the adaptability of each edge model to perception data changing, it is required that each edge model/agent is effectively able to use the newly added unlabeled perceptual data to conduct online training and model tuning unassisted [5], forming an online evolutive learning [6,7] (OEL) system. The semi-supervised learning [8,9] (SSL) method can utilize the knowledge learned from a small amount of labeled data and dig the adequate discriminative information from massive unlabeled amounts of data to achieve continuous model optimization, which effectively

fits the online training strategy. At the same time, due to the characteristics of multi-view edge sensing devices, multiple sensing models can obtain perception data from different viewpoints with the same semantic target, and these perception data have pieces of strong complementary information. Through the collaborative discrimination and information fusion of multi-model perception data from different views, the discriminative reliability of edge models for unlabeled data will be greatly improved, and the confirmation bias [10] problem in the SSL process will be significantly reduced. Such an SSL-based OEL method with multi-model collaborative perception and continuous self-training will enable each edge sensing model to enhance its adaptability to data distribution changes, continuously improve model generalization without relying on manual annotations, and reduce the deployment complexity of the perception tasks.

SSL has developed rapidly [11,12] in recent years and has gradually become the primary method for solving label-scarcity problems and adapting data distribution for practical applications. However, multi-model and multi-view SSL methods [13] are still relatively rare. The existing multi-view SSL methods [14] are generally only for fixed small multi-view datasets, and the multi-view data are typically obtained through different feature extractors. Such multi-view data present challenges to guaranteeing the view independence between features, which makes these methods unable to be generalized to practical recognition tasks. In terms of multi-model SSL, the method of constructing multiple views from single-view data is relatively simple, and there are rare methods for multi-view agents to perceive scenes and solve practical OEL tasks.

Maintaining the discriminative independence between multi-view models and improving the model's collaborative discrimination reliability is the key to semi-supervised online evolutive learning systems. In this paper, we propose a multi-view agent's cooperative perception (MACP) semi-supervised online evolutive learning method that can solve the OEL problem well in the multi-view perception environment. Specifically, we first use different self-supervised [15] model-initialization (SMI) methods for different edge models, so that they can learn differentiated feature-extraction patterns from various self-supervised tasks. Combined with different data-normalization methods, SMI ensures the discriminative independence of each model when facing view-specific data. We then propose a discriminative information fusion (DIF) algorithm that votes and integrates the multi-view model's predicted distributions. DIF obtains a more reliable discriminant representation from unlabeled data for model training based on the discriminant criteria differences between multi-view models. To maintain the differentiation of the discriminant standards during the training process of each model, we further propose a parameter constraint (PC) method between models. By orthogonalizing some critical parameters of different models, the discriminant standards of each model are effectively prevented from converging during the training process. The proposed MACP achieves better convergence efficiency and final performance than other representative single-model SSL and multi-model SSL methods on multiple datasets. At the same time, we find that the pseudo-labels obtained by multi-view models based on DIF maintain a high accuracy rate during the training process, which further proves the reliability of our method. Moreover, since our method mainly constructs multi-view data from a single-view dataset, to explore the performance of MACP under an ideal multi-view sensing environment, we configure a collaborative perception experiment where the perception data streams are different data with the same categories. In an ideal multi-view perception environment, MACP achieves a performance that surpasses fully supervised learning methods under the same configuration, demonstrating our method's application potential in practical multi-agent collaborative sensing.

The main contributions of this paper can be summarized as:

- We analyze the existing problems of multi-agent collaboration and data-distribution adaptation in the multi-view sensing environment. We propose the multi-view agent's collaborative perception (MACP) semi-supervised online evolutive learning method. MACP can reduce the task complexity of multi-model SSL when processing multi-view perception data and realize real-time tuning of the local perception system.

- In MACP, we enable each model to learn a differentiated feature-extraction mode through a self-supervised model-initialization method, which enhances the discriminative independence of each model. By applying the discriminative information-fusion approach to the predictions of each view model, the reliability of the discriminant results is improved, and continuous consistency regularization training is realized. Through further regularization constraints on the parameters of each model in the training process, the model can continue to maintain a relatively independent discriminative ability, and the stability of the entire OEL system is improved.
- The proposed MACP achieves a better performance than the comparison methods on multiple datasets. In an ideal multi-view agent collaborative perception experiment, MACP exceeds the performance of the fully supervised learning method, which proves the applicability of the proposed method in practical multi-view sensing scenarios.

2. Related Work

Consistency-based [16,17], semi-supervised learning methods have achieved great success in recent years. The main theoretical basis is that the model should maintain consistent predictions for different input variants with the same semantics. Based on such a natural constraint, using data-augmentation methods to transform the input features and train the model to mine the consistency information of different input variants can effectively improve the generalization of the SSL model. MixMatch [18] first performs multiple augmentation operations on the same unlabeled input, then uses sharpening [19] on the average of all augmented data predictions, and finally guides the SSL model training through the prediction targets generated by mixing up [20] data and labels. FixMatch [21] simplifies the complex process of the previous work and directly inputs weakly augmented and strongly augmented versions of the same unlabeled data into model training. Better performance is achieved by converting the higher confidence part of the weakly augmented predictions into pseudo-labels to guide the model training on the strongly augmented data. AWLDA [22] proposes a strategy to count the class-wise learning progress in the training process and improves the contribution of hard-to-learn classes to training, which reduces the class imbalance problem in the SSL process. Meanwhile, this method makes better use of the consistent relationship between low-confidence predictions and significantly improves the convergence speed of SSL. Since this paper mainly studies the application of multi-view and multi-model semi-supervised learning methods in OEL scenarios, we will focus on analyzing research related to these goals.

For multi-view, semi-supervised learning methods, Co-training [23] first trains two different classifiers with labeled data of different views, and then exchanges the high-confidence predicted results for unlabeled data between each classifier for SSL training, realizing the discriminative information-sharing of each view model. DCT [24] proposes a differential constraint method based on adversarial samples for the co-training models, which makes the discriminative criteria of each model's approach to the adversarial samples of each other while providing mutual annotations. This method can continuously improve the discriminative difference of different view models but adds certain extra computation. Tri-training [25] first proposes a multi-view training method that does not rely on view differences. It uses bootstrap to sample three different subsets of data from the labeled dataset to train three different initial classifiers and then performs a voting process on the prediction results of the unlabeled data. The predictions agreed by the majority classifier will be used as the training target of the minority classifier. Tri-net [26] introduces the output-smearing [27] process for the three models with shared parameters to maximize the prediction difference of each model and then uses the voting results of the two models' predictions as the training target for the training of another model. This approach requires the periodic fine-tuning of the discriminative variability of the models during training, thus potentially reducing the coherence of implementation. Ref. [28] proposes a multi-view, semi-supervised feature-representation learning method that utilizes orthogonalization and adversarial constraints to improve the consistency between models and the ability to

extract complementary information. Other graph-based multi-view SSL methods [29,30] achieve better results on various specific multi-view databases by learning the relationship between feature representations from different views.

For multi-model, semi-supervised learning, such methods mainly utilize knowledge transfer between models in different learning states to achieve common performance improvements. The Π-model [31] achieves steady performance gains by using two models to predict different variants of the same unlabeled data and attaching consistency constraints to the predicted distributions. Meanteacher [32] adopts the mechanism of a teacher–student model for comparative learning and uses the exponential moving average of the updated parameters of the student model in the past training process as the teacher model. During training, the similarity between the discriminative results of the teacher model and the student model on unlabeled data is continuously strengthened. Dualstudent [33] believes that the teacher model in the aforementioned work is the historical average of the parameters of the student model, so there may be a performance bottleneck when guiding the training of the student model. It proposes a dual-student model structure, which further improves the model performance by evaluating the uncertainty of the prediction results of the two models and adding mutual stability constraints to the high-reliability predictions. MPL [34] proposed a teacher–student model structure based on the idea of meta-learning. The prediction of the teacher model on unlabeled data is used as the training target of the student model, and the classification loss of the student model on the labeled data is used as the training feedback for the teacher model. Such an information-interaction method enables the teacher model to optimize the pseudo-label discrimination criteria continuously.

The main work of existing multi-view or multi-model SSL algorithms [35] is to improve the discriminative difference between models and then mine complementary information between models to enhance the reliability of unlabeled data prediction. However, these methods generally have problems, such as the complicated design of differential constraint strategy and the insufficient universality of the method. For multi-model collaboration in OEL scenarios, more comprehensive research is needed.

3. Proposed Methods

In this section, we first provide the problem definition and state the perception environment and main goals of a multi-view agent learning system. Then, the realization method of each part-module of MACP is introduced. We represent view-specific agents with multiple different SSL models and utilize a continuous unlabeled data stream to simulate a multi-agent sensing environment.

3.1. Problem Definition

As shown in Figure 1, in a multi-agent perception environment, we can let $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_v$ be a set of multi-view edge models, where v represents different views. The models from different viewpoints will perceive the same scene in real-time.

For a semi-supervised multi-view classification task, let $\mathcal{X}^v = (x_i^v, y_i), i \in (1, \ldots, B)$ represent a batch of B-labeled training data x_i^v of the SSL model with view v, where y_i is the unified label of all view data. Let $\mathcal{U}^v = u_i^v, i \in (1, \ldots, \mu B)$ represent a batch of μB unlabeled perception data u_i^v of the SSL model with view v, where μ is a hyperparameter that controls the proportion of labeled and unlabeled data. Note that, due to the fixed view of the training dataset, the $\mathcal{X}^v$ and $\mathcal{U}^v$ of models with different views are generated by random image-augmentation methods $\mathcal{A}_w(\mathcal{X})$ and $\mathcal{A}_w(\mathcal{U})$, respectively, where $\mathcal{A}_w(\cdot)$ represents applying a random weak augmentation transformation to the input. The total amount of labeled data $\mathcal{X}^v$ for each view model will be much less than the unlabeled data $\mathcal{U}^v$. In each training iteration, the models $\mathcal{M}_v$ from different views will use real-time generated data streams $\mathcal{X}^v$ and $\mathcal{U}^v$ for SSL training, and the high-confidence pseudo-label of $\mathcal{U}^v$ will be determined by the collaborative discrimination of each view model.

Let $P_{\mathcal{M}_v}(y|x)$ represent the prediction result of the model $\mathcal{M}_v$ for the input data x. The main goal of the multi-view agent's collaborative perception (MACP) method is to fuse the discriminative results of each view model $\mathcal{M}_v$ on the unlabeled perception data $\mathcal{U}^v$ to obtain more reliable pseudo-labels for model training on their respective view data. As a result, the generalizability of each view model and its adaptability to the data distribution changes will continue to improve. Maintaining the discriminative independence of each view model during the training process will be the key to improving the reliability of the collaborative discriminant results.

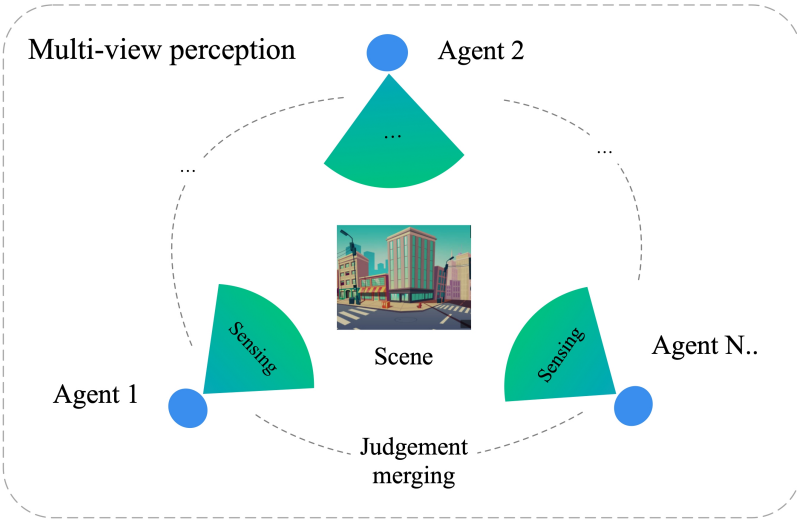


Figure 1. Diagram of multi-view agents collaborative perception and discrimination process.

3.2. Overall Framework

The main task of MACP is to design an effective collaborative discriminant process, which uses the multi-view model to predict perceptual data and achieve high-reliability pseudo-label extraction from the view-specific predictions. Each view model uses the collaborative discrimination results for continuous perception and training, so that the overall learning system can continuously adapt to changes in data distribution.

The overall framework of the proposed method is shown in Figure 2. Note that, in addition to a small amount of labeled data, a large amount of unlabeled perceptual data will continue to feed into different view models, constituting a continuous online learning process. MACP mainly includes three steps. First, the models from different views are initialized based on different self-supervised learning methods to ensure that each model has a differentiated feature-extraction pattern. Then, each model performs discriminative information-fusion processing on the predictions of view-specific perceptual data to obtain high-confidence pseudo-labels, which are used in the consistency-regularization training process of SSL. Finally, additional parameter constraints are introduced into the model-training process to maintain the discriminative independence of each model during the training, thereby preserving the stable operation of the entire learning system.

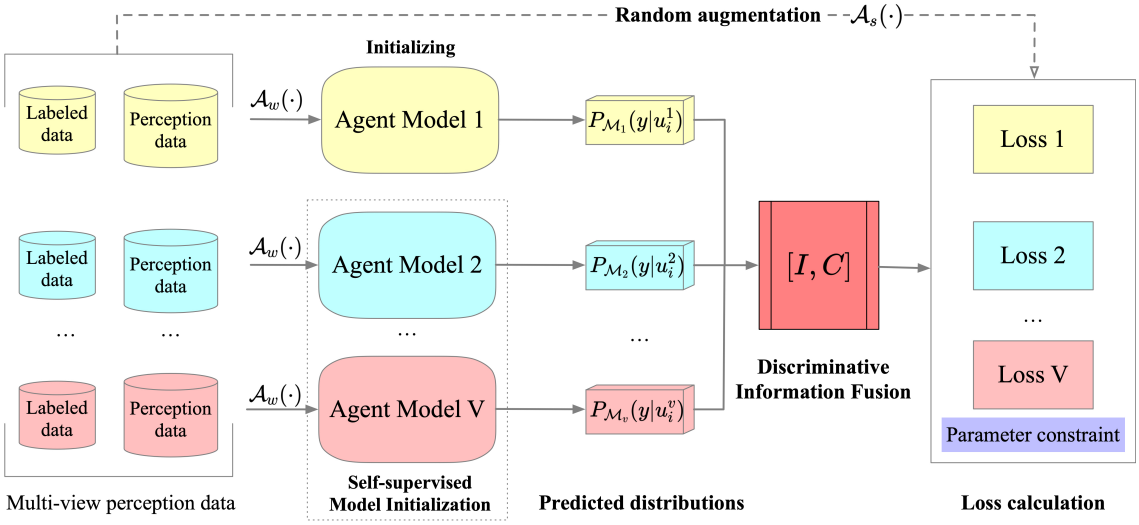


Figure 2. The overall framework of MACP.

3.3. Self-Supervised Model Initialization

Since models from different views will predict different perceptual data with the same semantics, each model's discriminative independence will significantly impact the final discriminative information-fusion results. The more independent discriminant ability will make each model make mistakes in different places, so the obtained fusion discriminant results will have higher reliability. We use various initialization methods for models from different views to obtain each model's view-specific feature-extraction pattern in the model initialization stage.

Specifically, taking the three-view perception models set as an example, for the first-view model, we only perform default parameter initialization processing for it. For the models from the other two views, we pre-train them on the self-supervised learning-based jigsaw puzzle solving [36,37] task and the generative adversarial network [38] task, respectively, to increase the differences in the feature-extraction patterns among the models.

For the self-supervised jigsaw-puzzle-solving task, let $\mathcal{U}^v$ be the self-supervised training data. For each u_i^v in $\mathcal{U}^v$, we slice it into N image patches of equal size and assign labels y^n to each patch in order. Then, we randomly shuffle the image patches and stitch them into a new image $\hat{u}_i^v$. Through an N-way classifier, the model $\mathcal{M}_v$ will predict the position of each image patch in the spliced image, and the label y^n will be used to guide the model to generate the correct image patch order prediction for the scrambled image. The loss function of the jigsaw puzzle solving task is as follows:

$$\mathcal{L}_{jigsaw} = -\frac{1}{\mu BN}\sum_{b=1}^{\mu B}\sum_{n=1}^{N} y^n log(P_{\mathcal{M}_v}(y|\hat{u}_b^v)^n), \tag{1}$$

where $P_{\mathcal{M}_v}(y|\hat{u}_b^v)^n$ represents the category prediction of model $\mathcal{M}_v$ for the n-th image block in $\hat{u}_b^v$, and $\mathcal{L}_{jigsaw}$ is the loss function of N-way categorical cross-entropy for all images in the current batch B. By solving the jigsaw puzzle task, the view-specific model $\mathcal{M}_v$ can learn a good representation of the spatial positional relationship of the image, thereby focusing on extracting differentiated features that are different from other view models.

For the self-supervised generative adversarial network (GAN) task, while maintaining the basic GAN composed of the generator and the discriminator, we train the generator to continuously learn the ability to generate images from the current database. At the same time, we modify the discriminator so that it is not only responsible for predicting the authenticity of the generated images, but also has image-classification capabilities. Specifically, for the original classification model $\mathcal{M}_v$, we keep its basic classifier unchanged and perform additional activation processing on the output logits z_c of the model:

$$q_{binary}(z_c) = \frac{\sum_{c=0}^{C} exp(z_c)}{\sum_{c=0}^{C} exp(z_c) + 1} \quad , \tag{2}$$

where C is the number of categories of the original image classifier, and, through exponential normalization, the multi-dimensional output z_c of the model for a certain input is converted into a one-dimensional binary prediction $q_{binary}(z_c)$, which is used for the training of the discriminator.

We know that in Equation (2), when the value of z_c is relatively large, $q_{binary}(z_c)$ will be close to 1, and when the value of z_c is relatively small, $q_{binary}(z_c)$ will be close to 0. This additional activation can train the discriminator to predict larger logits for real images and smaller logits for generated fake images, enabling the co-training of the discriminator for both multi-classification and generative adversarial tasks. The loss function of the self-supervised GAN task is as follows:

$$\mathcal{L}_{gan} = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} ylog(P_{\mathcal{M}_v}(y|x)) + \frac{1}{|D|} \sum_{x \in D} BCE(y_b, q_{binary}(P_{\mathcal{M}_v}(y|x))). \tag{3}$$

In Equation (3), the first term is the categorical cross-entropy loss for the labeled dataset $\mathcal{X}$, and the second term is the discriminator's binary cross-entropy loss for the real images and generated images in the entire dataset D. After the training of the self-supervised GAN, the model $\mathcal{M}_v$ will focus on mining the essential feature representation related to image generation, so as to gain a differentiated feature-extraction ability.

Through the designed self-supervised learning task, the view-specific initialization of each model is realized, and models from different views will use different feature-extraction patterns to predict the perceptual data. This paper takes the three-view model as an example to illustrate the self-supervised model-initialization (SMI) process. The SMI of more views can be implemented by using increasingly different self-supervised tasks such as image colorization [39], image super-resolution [40], contrastive learning [41], etc.

3.4. Discriminative Information-Fusion

The discriminative results of relatively independent multi-view perceptual data contain the consensus and complementary information of the predicted target. The effective fusion of these predicted distributions can obtain a more accurate class representation for the consistency regularization of SSL training.

The discriminative information-fusion process of the multi-view agent is shown in Figure 3. First, for each view v, we perform high-confidence filtering on the predictions of model $\mathcal{M}_v$ on the current batch of perceptual data $\mathcal{U}^v = u_i^v, i \in (1, \ldots, \mu B)$ to obtain predicted class labels and their corresponding indices for samples that satisfy the threshold condition:

$$[I^v, C^v]_{v \in V} = \sum_{i=1}^{\mu B} \mathbb{1}(max(P_{\mathcal{M}_v}(y|u_i^v)) > \tau) \cdot (argmax(P_{\mathcal{M}_v}(y|u_i^v))), \tag{4}$$

where $\mathbb{1}(max(P_{\mathcal{M}_v}(y|u_i^v)) > \tau)$ represents fetching the predicted class distributions greater than the threshold τ from the predictions of $\mathcal{U}^v$, and $argmax(P_{\mathcal{M}_v}(y|u_i^v))$ means obtaining the category label with the maximum probability in the corresponding prediction result. The predicted class labels and indices of high confidence predictions in the current batch are obtained through the above processing. I^v and C^v are two vectors that store the indices and class labels of valid samples obtained from the perceptual data of the current view v, and V is the total number of views.

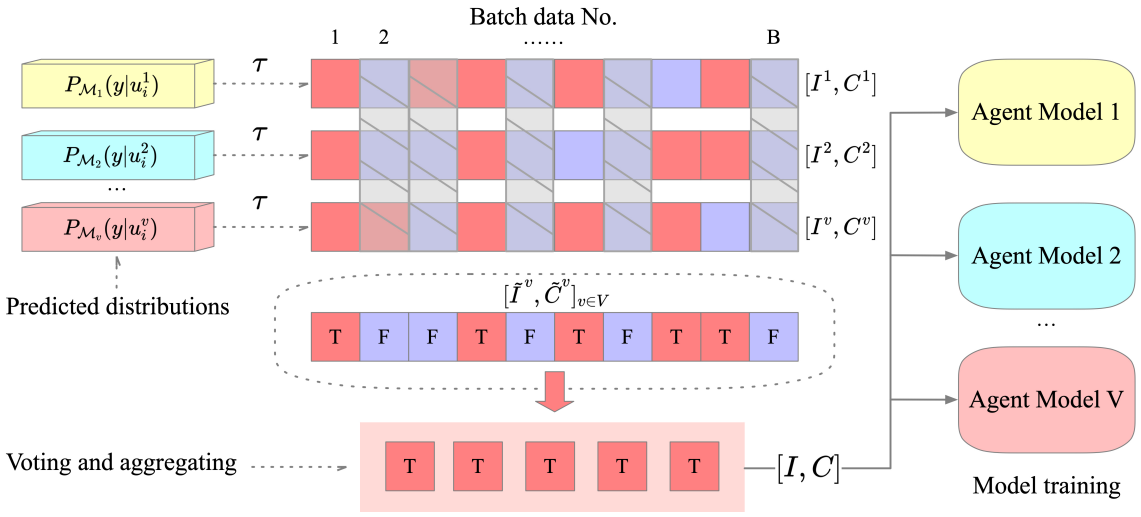


Figure 3. Multi-view agents discriminative information-fusion process.

Using Equation (4), we extract the indices of high-confidence samples and their corresponding class labels $[I^v, C^v]_{v\in V}$ in the prediction results of each view model. These results will be used for voting and aggregating to perform discriminative information fusion. The voting and aggregating process can be expressed as:

$$[\tilde{I}^v, \tilde{C}^v]_{v\in V} = [I^v, C^v] \cap [I^{\neg v}, C^{\neg v}], v \in (1, \ldots, V), \tag{5}$$

$$[I, C] = [\tilde{I}^v, \tilde{C}^v] \cup [\tilde{I}^{\neg v}, \tilde{C}^{\neg v}], v \in V. \tag{6}$$

In Equation (5), for the index- and class-label vectors $[I^v, C^v]$ of each view model, we, respectively, intersect them with the results of other view models to obtain the prediction consensus. Where $[I^{\neg v}, C^{\neg v}]$ represents the sample indices and class labels of other view models, and $[\tilde{I}^v, \tilde{C}^v]_{v\in V}$ are each view model's voted results. Then, according to Equation (6), we take the union of all the compatible parts of $[\tilde{I}^v, \tilde{C}^v]_{v\in V}$ to obtain the final discriminative fusion result $[I, C]$.

Then, the discriminative fusion results will be used by each view model for SSL training. For the unlabeled data $\mathcal{U}^v$ of each view, we first obtain the corresponding samples according to the index I and perform strong data-augmentation processing on them:

$$\widetilde{\mathcal{U}^v} = \mathcal{A}_s(u^v_{i\in I}), \tag{7}$$

where $\mathcal{A}_s(\cdot)$ represents the random strong data-augmentation function, and $u^v_{i\in I}$ represents the samples extracted from $\mathcal{U}^v$ according to the index I. At this time, the unlabeled perception data of each view in the current batch that meet the conditions will be assigned a pseudo-label C_i from C, and the unlabeled data batch becomes $\widetilde{\mathcal{U}^v} = (\widetilde{u}^v_i, C_i), i \in (1, \ldots, |C|)$.

Based on the discriminative information fusion (DIF) of the multi-view models, the final SSL loss of each view model can be expressed as:

$$\mathcal{L}^v_{ssl} = \frac{1}{B}\sum_{i=1}^{B} H(y_i, P_{\mathcal{M}_v}(y|x^v_i)) + \frac{1}{|C|}\sum_{i=1}^{|C|} H(C_i, P_{\mathcal{M}_v}(y|\widetilde{u}^v_i)), \tag{8}$$

where $H(y, x)$ represents the categorical cross-entropy loss, the first term of Equation (8) is the supervised loss of the labeled data x^v_i under the current batch, and the second term is the unsupervised loss for the augmented unlabeled data $\widetilde{u}^v_i$ with the pseudo-label C_i as the target. In each iteration of the different view models, reliable pseudo-labels are obtained by collaborative DIF of perceptual data for their respective unsupervised loss

calculations. Multi-view agents can fully use their different discriminant criteria to better mine generalization information from multi-view perception data.

3.5. Parameter Constraint

Since models from different views will use the pseudo-labels provided by DIF to train on their own perception data, as the model reaches higher iterations, the discrimination independence provided by SMI will gradually weaken. Therefore, the discriminative criteria of each view model may have the convergence risk in the later training stages. To prevent the increase of confirmation bias during model training, we further introduce a parameter-regularization constraint for different view models, making each model maintain discriminative independence as much as possible during the training process.

Specifically, we sequentially perform orthogonalization constraints on the critical parameters of each view model. During the training process, the parameters of the output layer and the critical feature-extraction layer of each model are kept irrelevant, thereby reducing the possibility of model-discriminating pattern convergence.

Let the critical parameters of the model $\mathcal{M}_v$ be Θ_v. The additional parameter-constraint loss can be expressed as:

$$\mathcal{L}_{reg}^{v} = |\sum(\Theta_v \cdot \Theta_{v-1})|_{v\neq 1}, \tag{9}$$

where Θ_{v-1} is the critical parameter of the previous view model, and the inner product loss of the two sets of critical parameter vectors will ensure that the parameters of different view models are orthogonalized. Through parameter constraints, the model of each view will gradually increase the differentiated discriminative ability during the training process.

Combining the SSL loss and parameter constraint loss, the total loss function of each view model in MACP is:

$$\mathcal{L}^{v} = \mathcal{L}_{ssl}^{v} + \mathcal{L}_{reg}^{v}. \tag{10}$$

4. Experiments

This section will first introduce the implementation details and hyperparameter configuration of the proposed method and report the performance comparison and efficiency analysis of MACP with other representative multi-model and single-model SSL algorithms. Then, we configure an experiment in an ideal multi-view perception environment to illustrate the effectiveness of MACP in practical OEL applications. In the ablation study, the effects of different modules of MACP on the training performance are analyzed, and the variants in discriminative information-fusion methods are discussed.

4.1. Implementation Details

Since several representative SSL algorithms are configured in different experimental environments, to ensure a fair comparison, we re-implement several methods used for performance analysis in the same environmental configuration, while ensuring that the training hyperparameters are as similar as possible. Our main programming environment is the Keras deep learning library with Tensorflow as the backend.

Experiment Settings We configure two experiments to demonstrate the effectiveness of the proposed method. When comparing with general multi-model or single-model SSL methods, we train the model multiple times with different labeled-data splits and compare the test-error rate with other methods. Since the existing multi-view datasets are generally small and unrepresentative, in order to better reflect the performance of the proposed method in a realistic perception environment, we then use the existing dataset to simulate an ideal multi-view perception experiment. Specifically, we configure the perceptual data stream composed of different data with the same category for each view model to perform collaborative perceptual learning. The relatively independent perceptual data environment we construct may bring more supervision information to each view model

than SSL, so we compare the model's performance with the fully-supervised learning method in this experiment.

Datasets We configure multiple sets of experiments with different amounts of labeled data on the CIFAR-10/100 [42] and SVHN [43] datasets, which are widely used for SSL methods. Both CIFAR-10 and SVHN are 10-class datasets, where CIFAR-10 contains 50,000 training images and 10,000 test images with class balance, and SVHN contains 73,357 training images and 26,032 testing images with imbalanced classes. CIFAR-100 is a relatively complex 100-class dataset containing 50,000 training images and 10,000 testing images, with only 500 training images for each class. In each group of experimental configurations, we extract a small amount of class-balanced data from the training set to construct the labeled dataset and remove the labels of all training set data to form the unlabeled dataset. Note that the unlabeled datasets of SVHN are imbalanced. The test set of each dataset is used to evaluate the performance of different methods.

Data Normalization We employ different data-normalization techniques for models from different views. The main methods include normalizing the pixels of each image to conform to the standard normal distribution by calculating the channel-wise mean and standard deviation of the training set images, normalizing the image pixel values to be between 0 and 1, between -1 and 1, etc.

Data Augmentation We employ two different data-augmentation methods, weak augmentation $\mathcal{A}_w(\cdot)$ and strong augmentation $\mathcal{A}_s(\cdot)$. Weak augmentation methods will randomly flip and crop images to generate different views' perceptual data. The strong augmentation method adopts Randaugment [44], which will be used for consistency regularization in SSL training on the collaborative discriminant results.

Base Model We use Wide-ResNet [45,46] as the base model, and, for simpler CIFAR-10 and SVHN tasks, we use WRN-28-2. For the CIFAR-100 classification task with more categories, we use the wider WRN-28-8. Models from different views will use different parameter-initialization methods [47] to further increase the discriminant difference.

Optimizer Settings We adopt the unified SGD optimizer for all models, with momentum $\beta = 0.9$ and Nesterov, and use weight decay with the coefficient of 0.0005. We uniformly set models to the initial learning rate $\eta = 0.05$ and use the cosine learning rate decay [48] strategies.

We set a uniform number of iterations $K = 2^{18}$ for each experiment. For each experimental configuration with different amounts of labeled data, we use different random seeds to sample three sets of labeled data for model training to ensure the reliability of the experimental results. The total number of different view models is V. All hyperparameters used in the experiments are reported in Table 1.

Table 1. List of hyperparameters for all datasets.

Dataset	CIFAR-10	SVHN	CIFAR-100
τ		0.95	
V		[2, 3]	
μ		4	
B		64	
K		2^{18}	
η		0.05	
β		0.9	
Weight decay		0.0005	

4.2. Main Results

We report the performance comparison of our method with the multi-model SSL methods Mean Teacher [32], Dual Student [33], Deep CT [24], Tri-net [25], and the representative single-model SSL methods UPS [49], MixMatch [18], and FixMatch [21]. We configure 2-view and 3-view MACP experiments to compare the difference in the number of views;

each experiment was run three times with a different labeled data split, and the mean and standard deviation of the final test error rates are reported.

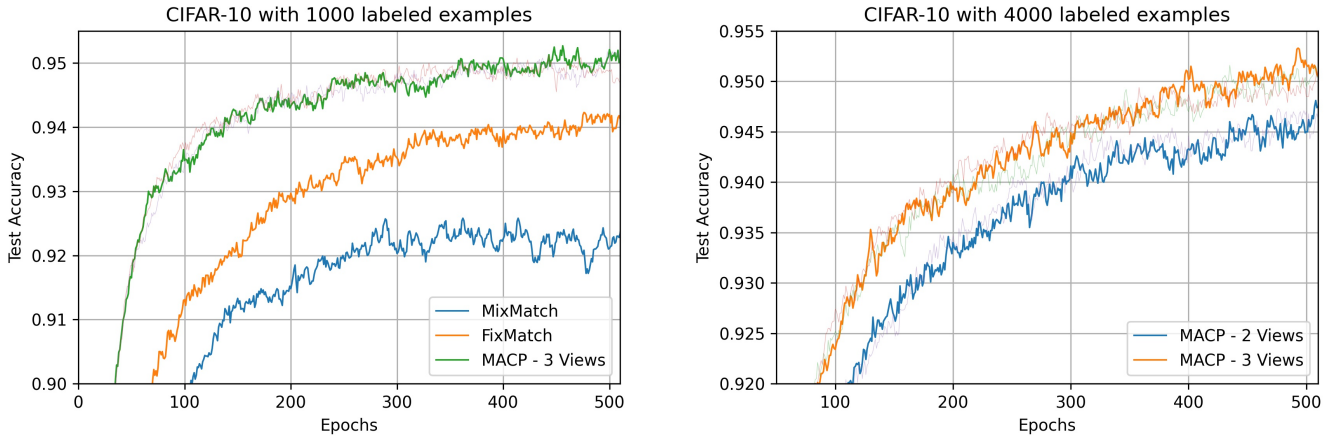
As Table 2 shows, as the past multi-model SSL methods only focus on the prediction consistency between different models or have shortcomings in constructing differentiated multi-view data, the final performance is weaker than other methods. The proposed self-supervised model-initialization method can make models from different perspectives have more independent feature-discrimination criteria, thus increasing the reliability of the fused discriminant results. Some state-of-the-art single-model SSL methods strengthen the consistency-regularization constraint, adopt more post-processing algorithms for the predicted distribution of unlabeled data to obtain more reliable pseudo-labels, and achieve better performance. However, these methods cannot directly realize multi-view model interaction in the OEL environment. Under the same hyperparameter configuration, our proposed MACP method outperforms other algorithms in both two-view and three-view experiments. In the class-imbalanced SVHN experiment, MACP also achieved better performance than other methods, indicating that the multi-view discriminative information fusion performs a more reliable class judgment on unlabeled data. Moreover, in the CIFAR-10-1000-label and CIFAR-100-4000-label experiments with relatively few labeled data, MACP achieves 5.29% and 31.67% test error rates, respectively, which are significantly better than other methods. These results show the superiority of MACP in the face of perceptual environments where labeled data are lacking.

Table 2. Comparison of error rate (%) for CIFAR-10/100 and SVHN on three different labeled data folds, the comparison methods are tested under the same codebase.

	CIFAR-10			**SVHN**		**CIFAR-100**	
Method	**1000 Labels**	**2000 Labels**	**4000 Labels**	**250 Labels**	**1000 Labels**	**4000 Labels**	**10,000 Labels**
Mean Teacher	21.55 ± 1.48	15.73 ± 0.31	12.31 ± 0.28	4.35 ± 0.50	3.95 ± 0.19	-	-
Dual Student	14.17 ± 0.38	10.72 ± 0.19	8.89 ± 0.09	4.24 ± 0.10	-	-	33.08 ± 0.27
Deep CT	-	-	8.54 ± 0.12	-	3.38 ± 0.05	-	34.63 ± 0.14
Tri-net	-	-	8.30 ± 0.15	-	3.45 ± 0.10	-	-
UPS	8.18 ± 0.15	-	6.39 ± 0.02	-	-	40.77 ± 0.10	32.00 ± 0.49
MixMatch	7.72 ± 0.37	6.89 ± 0.39	5.21 ± 0.09	4.06 ± 0.18	3.49 ± 0.32	36.12 ± 0.62	29.12 ± 0.34
FixMatch	6.18 ± 0.56	5.92 ± 0.32	4.99 ± 0.11	3.83 ± 0.45	3.08 ± 0.63	33.78 ± 0.31	25.69 ± 0.61
MACP (2 views)	6.02 ± 0.39	5.69 ± 0.40	4.91 ± 0.08	3.57 ± 0.34	2.99 ± 0.26	33.52 ± 0.45	25.77 ± 0.83
MACP (3 views)	5.29 ± 0.37	5.12 ± 0.31	4.75 ± 0.20	3.32 ± 0.51	2.72 ± 0.15	31.67 ± 0.29	24.72 ± 0.11

We further analyze the training efficiency and stability of MACP, see Figure 4. As Figure 4a shows, in the CIFAR-10-1000-labels experiment, the convergence efficiency and test accuracy of MACP are significantly higher than MixMatch and FixMatch, thus achieving a better final performance. Since the methods used for comparison are single-model SSL, to ensure the fairness of the comparison, we did not use the integrated prediction results of multi-view MACP for the performance evaluation but reported the independent evaluation results of each view model separately. The training curve of one view model in MACP is displayed normally in Figure 4, and the training curves of the other view models are represented by thin transparent curves. We also evaluate the performance of multi-view discriminative information fusion during MACP training. We find that, in the later training stage, MACP can obtain more than 90% unlabeled perception data in each batch for SSL training, and the pseudo-label accuracy of these data is higher than 97.5%, which shows that the DIF process of MACP generates more reliable pseudo-labels, thus achieving a more stable learning performance. If we follow the ensemble learning strategy and fuse the prediction results of different view models on the test set, the test accuracy will be further improved. However, to ensure a fair comparison, we still use the results of the single-view model to compare with existing methods. Furthermore, MACP's performance under different numbers of views is evaluated. As can be seen from Figure 4b, the 3-view MACP can achieve higher convergence efficiency and test accuracy than the 2-view version

during the whole training process. We find that, with the increase of views, due to the addition of more independent discriminative information, the reliability of the DIF of the models will be continuously improved, and the generalizability of models from different views will be jointly enhanced.



(**a**) (**b**)

Figure 4. Performance comparison of MACP with different methods and different number of views during training. (**a**) Performance comparison for MixMatch, FixMatch and MACP on CIFAR-10-1000-label experiment; (**b**) Performance comparison for 2-view and 3-view MACP on CIFAR-10-4000-label experiment.

4.3. Ideal Multi-View Perception Experiments

The MACP method assumes that each view model continuously obtains perception data from different viewpoints. These view-specific perception data can be used for model training through the DIF process. The multi-view perception data in real scenes are naturally quite different. However, in the experiments above, we perform random data-augmentation methods on the same data to generate simulated multi-perspective data, which may still contain more related information, thus limiting the model's performance.

More differentiated multi-view perception data will help each view model learn more independent feature-extraction patterns. To strengthen the difference between perception data from different views, we design a new sensing environment to test the online evolutive learning of multi-view agents under ideal conditions. Specifically, for each iteration, we assign a batch of different data with the same category to each view model. The joint discrimination results of each view model on these perceptual data will be used for SSL training.

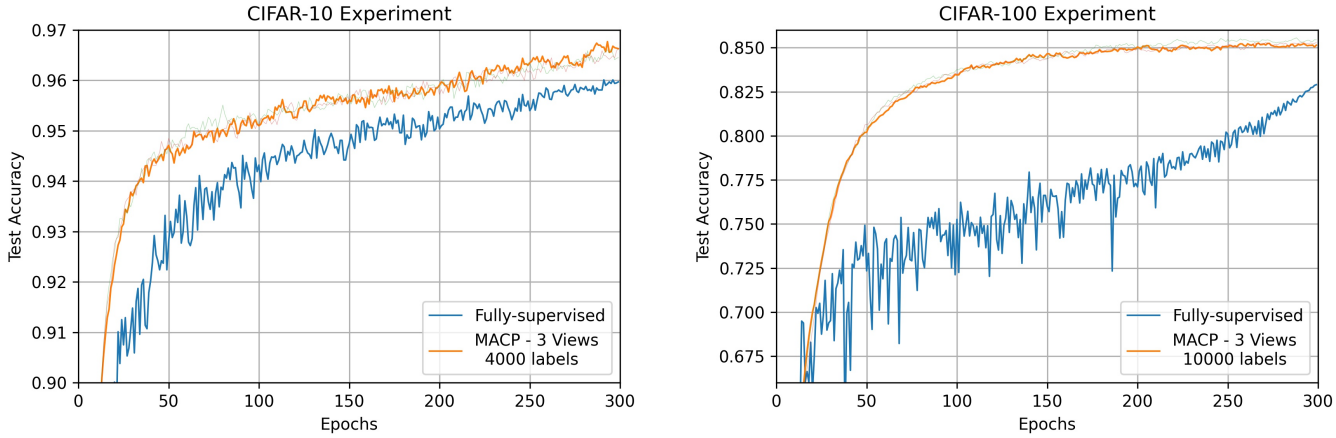
In Table 3, we report the performance comparison of MACP and the fully-supervised learning method with various configurations under the ideal perceptual environment. In a more independent perception environment, since models from different views can provide more differentiated discriminant information, the reliability of the final DIF result will be significantly enhanced, providing more accurate classification supervision for each model. At the same time, through the information exchange between multi-view models, each model can obtain more generalized knowledge, which significantly reduces the model variance and enables MACP to achieve a better performance than fully-supervised learning. These results show that MACP has great application potential in practical multi-view sensing environments, which can solve the labeled data scarcity problem for edge models and enable each agent to adapt to the sensing-environment changes continuously.

In Figure 5a, we compare the training curves of MACP and the fully-supervised method under the CIFAR-10 databset. It can be seen that MACP has a faster convergence speed than fully-supervised learning, and the training fluctuation is smaller, which indicates that the multi-view models' collaborative perception makes the training process more stable.

As shown in Figure 5b, the stability advantage of MACP is more pronounced in the more complex CIFAR-100 experiments. In the early training stage of each view model, a small number of discriminative fusion results can be extracted from the easy-to-judge perceptual data. As the training progresses, the discriminative ability of each view model is gradually enhanced, and more valuable information can be obtained from more perceptual data. Such a step-by-step training process prevents the models from converging to local minima, resulting in better performance.

Table 3. Comparison of test accuracy (%) between MACP and fully-supervised method in an ideal multi-view perception environment.

	CIFAR-10		SVHN		CIFAR-100
Method	**1000 Labels**	**4000 Labels**	**250 Labels**	**1000 Labels**	**10,000 Labels**
Fully-supervised	95.98		97.72		82.82
MACP (2 views)	96.23 ± 0.12	96.45 ± 0.07	97.82 ± 0.31	98.16 ± 0.51	83.06 ± 0.18
MACP (3 views)	96.41 ± 0.21	96.75 ± 0.03	98.21 ± 0.17	98.38 ± 0.34	85.39 ± 0.11



(**a**) (**b**)

Figure 5. Performance comparison between MACP and fully-supervised method. (**a**) Performance comparison of 3-view MACP and fully-supervised method on CIFAR-10 experiment; (**b**) Performance comparison of 3-view MACP and fully-supervised method on CIFAR-100 experiment.

4.4. Ablation Study

Since the proposed MACP method consists of three modules, self-supervised model initialization (SMI), discriminative information fusion (DIF), and parameter constraints (PC), we will further analyze the impact of different module combinations on model performance.

We report the performance of 3-view MACP in CIFAR-10-4000-labels, SVHN-1000-labels, and CIFAR-100-10,000-label experiments under various module combinations. The DIF method is the key to the multi-view collaborative perception system. SMI and PC will provide differential regularization for each view model in the initial training stage and the subsequent training process, respectively, to ensure each view model's relatively independent discriminative ability.

As shown in Table 4, when the three modules are not applied, it is equivalent to training each model separately without any information exchange, and the performance of each view model is poor at this time. When only DIF is used, due to the lack of independence constraints between models, the discriminative pattern of each model will gradually converge during the training process, resulting in performance bottlenecks in the later training stage. When using SMI combined with DIF, due to the lack of continuous independence constraints of models in the later training stage, it faces the risk of falling into the plateau

as the model iterations, although it has high convergence performance in the early training stage. When DIF and PC are combined, although the initial independence constraint is lacking, each view model can gradually improve the difference in feature-discrimination patterns during the training process, thus achieving better performance. When the three modules are applied together, each view model can continuously exchange information in a relatively independent discriminative environment and achieve the optimal final performance.

Table 4. Ablation study on MACP with different module combinations, test accuracy (%) of each setting on CIFAR-10/100 and SVHN are reported.

Module Combination			Dataset		
SMI	**DIF**	**PC**	**CIFAR-10**	**SVHN**	**CIFAR-100**
			92.10 ± 0.72	95.15 ± 0.34	72.17 ± 0.53
	✓		93.17 ± 0.19	95.98 ± 0.12	73.97 ± 0.43
✓	✓		94.23 ± 0.09	96.53 ± 0.31	74.92 ± 0.19
	✓	✓	94.93 ± 0.36	97.03 ± 0.22	74.62 ± 0.25
✓	✓	✓	95.25 ± 0.20	97.28 ± 0.15	75.28 ± 0.11

In Section 3.4, our discriminative information-fusion method is implemented in a synchronous manner, and the DIF results of different view models on unlabeled perceptual data will be directly used for their respective training. Here we further explore the impact of the asynchronous DIF approach on model performance. In the asynchronous DIF setting, the parameter updates of each view model will be performed sequentially, that is, after the current model is updated according to the current DIF results, the model of the following view will be trained using the updated DIF results.

Through extensive experiments, we found that the asynchronous DIF approach is not conducive to optimizing MACP. Although asynchronous DIF enables a faster transfer of discriminative information between models, there is an increased risk of mis-discrimination. In the early training stage, when the discriminative ability of each perspective model is insufficient, there may be more misjudgments in the collaborative discrimination results. At this time, the alternate training of models will lead to the continuous accumulation of training errors, weakening the stability of DIF results. In asynchronous processing, the current model needs to wait for the update of other view models, which also affects the training efficiency. Moreover, the asynchronous DIF method will increase the possibility of the discriminative pattern convergence between models during the training process, affecting the discriminative independence of the models from different views. Overall, the asynchronous DIF approach will result in an up-to-5% performance degradation for each view model.

5. Conclusions

We propose MACP, which realizes online evolutive learning for efficient adaptation to the continuously changing sensing-data distribution through multi-view models' independence constraints and collaborative discrimination. MACP consists of three main modules. Through the self-supervised model-initialization method, each view model learns different feature-extraction patterns. Through the discriminative information-fusion process, more reliable pseudo-labeled predictions are mined from multi-view unlabeled perceptual data for SSL training. Combined with the multi-model parameter constraint during the training, MACP achieves excellent performance over multiple representative multi-model and single-model SSL methods. In experiments on simulated ideal multi-view perception environment, MACP achieves performance that surpasses the fully-supervised learning methods, proving the practical application value of the proposed method.

With the increase of various edge intelligent sensing devices, the online evolutive learning method that improves the continuous adaptability of edge models to environmental changes through multi-agent collaboration will have significant developmental

prospects. In future work, we will investigate more multi-agent interactive learning methods and discriminant independence-constraint methods to improve the adaptability and generalization of edge models to the perception environment. We will also explore the application of MACP in the real-world perception environment.

Author Contributions: Conceptualization, L.S. and D.L.; methodology, D.L. and L.S.; software, D.L.; validation, D.L. and L.S.; formal analysis, D.L.; investigation, D.L.; data curation, D.L.; writing—original draft preparation, D.L.; writing—review and editing, L.S.; visualization, D.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Shanghai Key Research Laboratory of NSAI and the China Mobile Research Fund of Chinese Ministry of Education (Grant No. KEH2310029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CIFAR-10/100 and SVHN datasets used for training and testing are available at https://www.cs.toronto.edu/~kriz/cifar.html and http://ufldl.stanford.edu/housenumbers/ (accessed on 24 July 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deng, S.; Zhao, H.; Fang, W.; Yin, J.; Dustdar, S.; Zomaya, A.Y. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet Things J.* **2020**, *7*, 7457–7469. [CrossRef]
2. Zhou, Z.; Chen, X.; Li, E.; Zeng, L.; Luo, K.; Zhang, J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proc. IEEE* **2019**, *107*, 1738–1762. [CrossRef]
3. Li, E.; Zhou, Z.; Chen, X. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In Proceedings of the 2018 Workshop on Mobile Edge Communications, Budapest, Hungary, 20 August 2018; pp. 31–36.
4. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
5. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
6. Song, L.; Hu, X.; Zhang, G.; Spachos, P.; Plataniotis, K.; Wu, H. Networking Systems of AI: On the Convergence of Computing and Communications. *IEEE Internet Things J.* **2022**. [CrossRef]
7. Li, D.; Zhu, X.; Song, L. Mutual match for semi-supervised online evolutive learning. *Appl. Intell.* **2022**, 1–15. [CrossRef]
8. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
9. Zhu, X.; Goldberg, A.B. *Introduction to Semi-Supervised Learning*; Synthesis Lectures on Artificial Intelligence and Machine Learning; Morgan & Claypool Publishers: San Rafael, CA, USA, 2009.
10. Klayman, J. Varieties of confirmation bias. *Psychol. Learn. Motiv.* **1995**, *32*, 385–418.
11. Nassar, I.; Herath, S.; Abbasnejad, E.; Buntine, W.L.; Haffari, G. All Labels Are Not Created Equal: Enhancing Semi-Supervision via Label Grouping and Co-Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7241–7250.
12. Sellars, P.; Aviles-Rivero, A.I.; Schönlieb, C.B. LaplaceNet: A Hybrid Energy-Neural Model for Deep Semi-Supervised Classification. *arXiv* **2021**, arXiv:2106.04527.
13. Yang, X.; Song, Z.; King, I.; Xu, Z. A survey on deep semi-supervised learning. *arXiv* **2021**, arXiv:2103.00550.
14. Shi, C.; Lv, Z.; Yang, X.; Xu, P.; Bibi, I. Hierarchical multi-view semi-supervised learning for very high-resolution remote sensing image classification. *Remote Sens.* **2020**, *12*, 1012. [CrossRef]
15. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [CrossRef] [PubMed]
16. Bachman, P.; Alsharif, O.; Precup, D. Learning with Pseudo-Ensembles. *Adv. Neural Inf. Process. Syst.* **2014**, *2*, 3365–3373.
17. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1163–1171.
18. Berthelot, D.; Carlini, N.; Goodfellow, I.J.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5050–5060.
19. Goodfellow, I.J.; Bengio, Y.; Courville, A.C. *Deep Learning*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2016.
20. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.

21. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
22. Li, D.; Liu, Y.; Song, L. Adaptive Weighted Losses with Distribution Approximation for Efficient Consistency-based Semi-supervised Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**. [CrossRef]
23. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.
24. Qiao, S.; Shen, W.; Zhang, Z.; Wang, B.; Yuille, A. Deep co-training for semi-supervised image recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–152.
25. Zhou, Z.H.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1529–1541. [CrossRef]
26. Dong-DongChen, W.; WeiGao, Z.H. Tri-net for semi-supervised deep learning. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2014–2020.
27. Breiman, L. Randomizing outputs to increase prediction accuracy. *Mach. Learn.* **2000**, *40*, 229–242. [CrossRef]
28. Jia, X.; Jing, X.Y.; Zhu, X.; Chen, S.; Du, B.; Cai, Z.; He, Z.; Yue, D. Semi-supervised multi-view deep discriminant representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2496–2509. [CrossRef]
29. Zhang, B.; Qiang, Q.; Wang, F.; Nie, F. Fast multi-view semi-supervised learning with learned graph. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 286–299. [CrossRef]
30. Nie, F.; Tian, L.; Wang, R.; Li, X. Multiview semi-supervised learning model for image classification. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 2389–2400. [CrossRef]
31. Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; Raiko, T. Semi-supervised Learning with Ladder Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 3546–3554.
32. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1195–1204.
33. Ke, Z.; Wang, D.; Yan, Q.; Ren, J.; Lau, R.W. Dual student: Breaking the limits of the teacher in semi-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6728–6736.
34. Pham, H.; Dai, Z.; Xie, Q.; Le, Q.V. Meta Pseudo Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11557–11568.
35. Wei, H.; Feng, L.; Chen, X.; An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13726–13735.
36. Wei, C.; Xie, L.; Ren, X.; Xia, Y.; Su, C.; Liu, J.; Tian, Q.; Yuille, A.L. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1910–1919.
37. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 69–84.
38. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
39. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 649–666.
40. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
41. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
42. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Master's Thesis, University of Tront, Toronto, ON, Canada, 2009.
43. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning. 2011. Available online: http://ufldl.stanford.edu/housenumbers/(accessed on 24 July 2022).
44. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 3008–3017.
45. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.

48. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
49. Rizve, M.N.; Duarte, K.; Rawat, Y.S.; Shah, M. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.