Semi-Supervised RGB-D Hand Gesture Recognition via Mutual Learning of Self-Supervised Models

JIAN ZHANG, School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
KAIHAO HE, School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
TING YU, School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
JUN YU, School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
ZHENMING YUAN, School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China

Human hand gesture recognition is important to human–computer interaction. Gesture recognition based on RGB and Depth (RGB-D) data exploits both RGB and depth images to provide comprehensive results. However, the research under scenario with insufficient annotated data is not adequate. In view of the problem, our insight is to perform self-supervised learning with respect to each modality, transfer the learned information to modality-specific classifiers, and then fuse their results for final decision. To this end, we propose a semi-supervised hand gesture recognition method known as Mutual Learning of Rotation-Aware Gesture Predictors (MLRAGP), which exploits unlabeled training RGB and depth images via self-supervised learning and achieves multi-modal decision fusion through deep mutual learning. For each modality, we rotate both labeled and unlabeled images to fixed angles and train an angle predictor to predict the angles, then we use the feature extraction part of the angle predictor to construct the category predictor and train it through labeled data. We subsequently fuse the category predictors about both modalities by impelling each of them to simulate the probability estimation produced by the other, and making the prediction of labeled images to approach the ground truth annotation. During the training of category predictor and mutual learning, the parameters of feature extractors can be slighted fine-tuned to avoid under-fitting. Experimental results on NTU-Microsoft Kinect Hand Gesture dataset and Washington RGB-D dataset demonstrate the superiority of this framework to existing methods.

CCS Concepts: • **Computing methodologies** → **Machine learning**;

Additional Key Words and Phrases: RGB-D hand gesture recognition, Semi-supervised learning, Self-supervised learning, Mutual learning, Rotation angle prediction

This work is supported by the National Natural Science Foundation of China under Grant No. 61972361.
Authors' Contact Information: Jian Zhang (corresponding author), School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China; e-mail: jeyzhang@hznu.edu.cn; Kaihao He, School of Computer Science, Hangzhou Dianzi University, Hangzhou, China; e-mail: 2428957769@qq.com; Ting Yu, School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China; e-mail: yut@hznu.edu.cn; Jun Yu, School of Computer Science, Hangzhou Dianzi University, Hangzhou, China; e-mail: yujun@hdu.edu.cn; Zhenming Yuan (corresponding author), School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China; e-mail: zmyuan@hznu.edu.cn.

ACM 1551-6865/2025/3-ART104
https://doi.org/10.1145/3689644

ACM Reference format:
Jian Zhang, Kaihao He, Ting Yu, Jun Yu, and Zhenming Yuan. 2025. Semi-Supervised RGB-D Hand Gesture Recognition via Mutual Learning of Self-Supervised Models. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 4, Article 104 (March 2025), 20 pages.
https://doi.org/10.1145/3689644

1 Introduction

Hand gesture recognition [48] is an important research topic in human–computer interaction. Most relevant works focus on image and video-based gesture recognition [31, 34]. With the development of sensor technology, depth information can be captured and utilized in gesture recognition. Under this situation, joint usage of **RGB and Depth (RGB-D)** information plays important role in salient object detection [42], object recognition [39], scene classification [41], person re-identification [46], and also hand gesture recognition [49]. The performance of above methods heavily depends on the amount of well-annotated training samples. However, labeling RGB-D images manually remains labor-intensive, and sufficient annotated data are often unavailable in practical applications. For this reason, hand gesture recognition with insufficient annotated data shows great significance, and this is referred to as semi-supervised RGB-D hand gesture recognition. Although semi-supervised RGB-D objection recognition has been studied for years [10–12], little work has been done on semi-supervised gesture recognition based on RGB-D data.

The key to the above-mentioned problem is semi-supervised learning methods. Cheng et al. [10–12] depict typical semi-supervised RGB-D object classification methods. In feature extraction, the study by Cheng et al. [12] directly use **Convolutional Neural Network (CNN)-Spatial Pyramid Matching (SPM)-Recurrent Neural Network (RNN)** while the studies by Cheng et al. in [10] and [11] resort to pre-training scheme to make use of the unlabeled data. The work by Cheng et al. [11] adopts unsupervised convolutional-recursive neural networks for feature extraction with the support of a fixed tree structure derived from the RGB-D images. The tree is originally designed to describe the scene image, so it may not adapt to object image well. The work by Cheng et al. [10] uses all RGB-D pairs to train a reconstruction network and construct classifier based on the reconstruction network. However, it is not interpretable which sort of information can be learned and how the object categorization will benefit from this information. In decision phase, all the three works use co-training scheme. The work by Cheng et al. [10] adopts both category classifier and attribute classifier to select highly confident samples to generate pseudo labels and enlarge the labeled training dataset. However, co-training requires that two modality-specific classifiers select training samples to enlarge the training set for each other, and the classifier about one modality may not adapt well to the selected samples with the other modality. This often degrades the effect of co-training.

Recently, **Self-Supervised Learning (SSL)** [29, 57] has been proposed to solve semi-supervised learning task. SSL relies on some pretext tasks to learn certain information from unlabeled data samples in an unsupervised manner. The learned information is proved to be beneficial to various downstream tasks such as object segmentation or object categorization. Contrastive learning [7] is an effective pretext task, which augments the dataset through color distortion, geometric distortion, cropping and filtering of images, and minimizes the distance between similar samples (augmented images from a same image) while maximizes the distance between dissimilar samples (augmented images from different images). Gidaris et al. [17] disclosed that abundant structural information can be learned by simply rotating the images to pre-designated angles and training classifier to predict the angles. This is because the model has to know the shape, profile, and orientation of the object in the image to realize accurate angle prediction. We deem this structural information is crucial to object categorization. In addition, deep mutual learning [55] indicates that classification accuracy

of multiple classifiers can be further improved by making each classifier to simulate the prediction result of other classifiers. This is because the prediction result consists of probabilities that the object belongs to each category, which provide the classifiers with exceptional inter-category similarities that cannot be reflected by the one-hot vector of training label. Mutual learning could be an effective way to multi-modal decision fusion in RGB-D hand gesture recognition.

Based on the above concerns, this article proposes a semi-supervised framework for RGB-D hand gesture recognition, which is referred to as **Mutual Learning of Rotation-Aware Gesture Predictors (MLRAGP)**. First, we rotate both labeled and unlabeled RGB-D gesture image pairs for 90°, 180°, and 270°, and train a rotation angle predictor for RGB images and depth images, respectively. Then we fix the feature extraction part of each angle predictor and link a gesture category predictor to each extractor, which is trained via the labeled RGB or depth gesture images. At last, we adopt deep mutual learning technique to exploit the complementary information implied in both RGB and depth images to fuse the predicted results of both gesture predictors. In training each gesture predictor and fusing the two gesture predictors, we allow the parameters of the feature extractors to be updated slightly, and the reason is that the feature extraction part should adapt to RGB-D fused gesture recognition. The contribution of this article is three-fold:

(1) We propose a semi-supervised RGB-D hand gesture recognition framework which resorts to SSL to exploit unlabeled data in training modality-specific gesture predictors, and adopts mutual learning to fuse the two predictors for better performance.
(2) We apply rotation angle prediction as the pretext task to SSL from both RGB and depth gesture images. The self-supervised model perceives the shape, profile, and the orientation of the gestures that are beneficial to recognition. In addition, mutual learning exploits the complementary information between RGB and depth gestures to enhance the performance of gesture predictors about RGB and depth images.
(3) We conduct extensive experiments on NTU-Microsoft Kinect Hand Gesture dataset and Washington RGB-D object dataset. The results indicate that the proposed approach yields state-of-the-art classification results in case of insufficiently annotated RGB-D sample pairs.

The rest part of the article is organized as follows: Section 2 reviews the related works. Section 3 introduces the proposed approach. Experimental results are presented in Section 4 and Section 5 concludes this article.

2 Related Works

2.1 Hand Gesture Recognition

Most of the hand gesture recognition methods are image and video-based methods. Sarma et al. [34] incorporate semantic segmentation achieved by a UNet architecture with attention-module into classification to achieve improved results for both static and dynamic gesture recognition. Raj et al. [31] propose a hand sign and gesture recognition method by using CNN to extract features like centroid of the hand, finger state, and thumb position to identify the gestures from video frames. Wei and Chen [44] propose to recognize continuous signs from videos through a weakly supervised framework, which exploits the cross-lingual signs as auxiliary training data to improve the recognition results of another sign language. Ravali et al. [25] construct a hand gesture dataset and trains a YOLOv8 network to obtain gesture features and achieve classification of two crucial hand movements. Some method exploits other less common sensors to assist gesture recognition. Challa et al. [6] propose a method based on electromyography. The authors placed Electromyography sensors on different parts of the hand to capture signals about hand movements and fed the movement features to algorithms such as random forest and logistic regression to conduct classification.

Wang et al. [43] use millimeter wave radar to capture hand gesture signals to form temporal distance-time map and Doppler-time map and fed them to neural network for gesture classification.

An important group of gesture recognition approaches takes advantage of depth information captured by depth sensors. The early work by Xie et al. [49] combines the RGB and depth image features to fine-tune an Inception V3 network for static gesture recognition. Elboushaki et al. proposed a multi-dimensional feature learning approach that combines convolutional residual network and long short-term memory network to learn high-level gesture features for gesture recognition in RGB-D videos [16]. Xu et al. [51] exploited the angle between fingertips and the distances between contour to palm center as features to conduct hand gesture recognition. El Ouariachi et al. [15] apply multi-channel cartesian Jacobi moments invariants method to depth, RGB, and RGB-depth data for hand gesture recognition. Wu and Ding [47] propose a depth-assisted hand gesture recognition approach to categorizing six different gestures. The authors segmented latent hand regions based on depth value of interest and tracked hand regions through multiple features, which were further used to achieve gesture recognition. Sun et al. [38] extracted features from RGB and depth images, respectively, and performed multi-level feature fusion with respect to related and independent features of the two modalities for gesture recognition. Li et al. [24] leveraged sufficiency and compactness rules to build novel loss function that generated optimized feature representation and reduced the impact of gesture-irrelevant information. Ma et al. [26] used a **Three-Dimensional (3D)** central difference convolution stem module to capture spatio-temporal features, which were passed to multiple factorized spatio-temporal stages to construct dimension-independent semantic primitives for gesture recognition. These works heavily rely on the amount of training data, and hand gesture recognition with insufficient annotated samples is not well studied.

2.2 RGB-D Object Categorization

Most RGB-D object categorization methods are fully supervised, among which deep learning [52–54] methods hold dominant position. Zhu et al. [56] use pre-trained CNN to extract features and use **Support Vector Machine (SVM)** to conduct classification. Cheng et al. [10], Chiu et al. [13], and Tan et al. [39] train the CNN-based feature extractors in an end-to-end way. Song et al. [37] extract the multi-modal semantic features based on object detections through faster RCNN. In multi-modal fusion strategy, these methods adopt either feature fusion or semantic fusion. Commonly seen feature fusion strategy is to concatenate the RGB and depth features directly [39] and send them to the classifier. Semantic fusion may happen in both intermediate part [56] and the decision part [13] of the model. Intermediate-level fusion appends semantic mapping layer to map the intermediate representations of RGB and depth features into a common semantic space. The semantic mapping may follow discriminative constraints [56] or correlation constrains. Decision-level semantic fusion could be the algebraic combination of the probability estimations of modality- specific classifiers or realized by cross-modal focal loss, without operating on the low-level or intermediate-level features.

Semi-supervised RGB-D categorization is proposed to exploit insufficiently annotated sample data [10–12]. The work by Cheng et al. [11] combines k-means clustering, CNN, and RNN with fixed tree structure for RGB and depth feature extraction and uses SVM for categorization. The work by Cheng et al. [12] adopts cascaded CNN, spatial pyramid pooling, and RNN for feature extraction. The studies by Cheng et al. in [11, 12] adopt co-training to exploit unlabeled training data. The classifier with respect to one modality predicts the category labels for unlabeled data, and then those samples with high confidence probabilities are selected as the most confident samples, whose counterpart samples in the other modality are added to the training set of the other classifier to boost the training. The study by Cheng et al. [10] improves co-training by introducing the attribute

classifier. This method uses convex clustering to construct attribute set for each category of each modality, and trains attribute classifier for RGB images and depth images, respectively, to predict the attribute of a given sample. In the process of co-training, the category classifiers and attribute classifiers are jointly used to select samples for enriching the training set such that the selected samples not only have high confidence scores but also cover all the attributes as even as possible.

2.3 Semi-Supervised Learning

Existing SSL algorithms consist of four categories. Pseudo label methods train an initial classifier with labeled data and use the classifier to predict pseudo labels for those unlabeled data, then append the data with pseudo labels to the training set to update the initial classifier [21]. The original pseudo label method does not consider the confidence score of unlabeled data, which may introduce noise into the classifier. Co-training solves this problem by initializing two classifiers and letting each classifier to select highly confident samples to update the other classifier [27, 30]. The second category refers to graph-based label propagation, which builds an affinity graph depicting the relationships between all samples and constrains the feature learning such that the learned features satisfy the graph constraints while the predictions of the labeled samples approach their labels. In deep learning version of graph-based method [45], the graph constraints can be applied to either the intermediate representations or predictions of the network. The difficulty is that the graph should be maintained during the learning process, which occupies extra storage space. This problem will be more serious in case of large data amount. The third category refers to the consistency regularization methods. Consistency regularization means that the classifier should make same prediction for perturbed versions of a labeled or unlabeled sample. The perturbation can be exerted either on model parameters [1] or directly on data samples [33, 50]. Consistency regularization has found its usage in many recent works. Zhu et al. proposed an **Unsupervised Semantic Consistent Learner (USCL)** [58]. They trained a supervised semantic learner using labeled and unlabeled data to obtain a series of semantic-aware proxies, which were then used to guide an unsupervised semantic learner by forcing the output of both learners about the unlabeled data to be consistent. Li et al. proposed to exploit **Consistency Between Student and Teacher (CBST)** [23] that integrates consistency learning into a student–teacher framework. They used labeled data to train a student network, whose moving average parameters were assigned to a series of teacher networks for unsupervised data learning at different training phases. The consistency regularization guaranteed that the teacher networks yield approximately the same prediction as the student network at each phase. The fourth category refers to self-supervised methods, which use both labeled and unlabeled data to pre-train a model according to certain pretext tasks and train classifier based on the model using the labeled data. RoPAWS [28] augments each unlabeled sample to form two views and learns a non-parametric similarity classifier with both labeled and unlabeled data to predict the class distributions from the network features of the two views. The pretext task is forcing the class distributions of the two views to be identical. Contrastive learning [7] exerts stochastic data augmentation to all the data and requires the predictions of similar samples (augmented from the same datum) to approach each other while those of the dissimilar samples (augmented from different data) stay far from one another. **Meta Contrastive Learning (MetaCL)** [22] makes predictions from multiple perturbed versions of a same sample and combines contrastive loss with component redundancy minimization and variability maximization to achieve the pretext training. Gidaris et al. [17] first rotate the images to fixed angles and train a model to predict the rotation angles, then construct classifiers based on the angle predictor. Though it is simple, the classifier still benefits a lot from angle prediction because the model has to be aware of the object shape and orientation to predict right angles and this information is also useful for categorization.

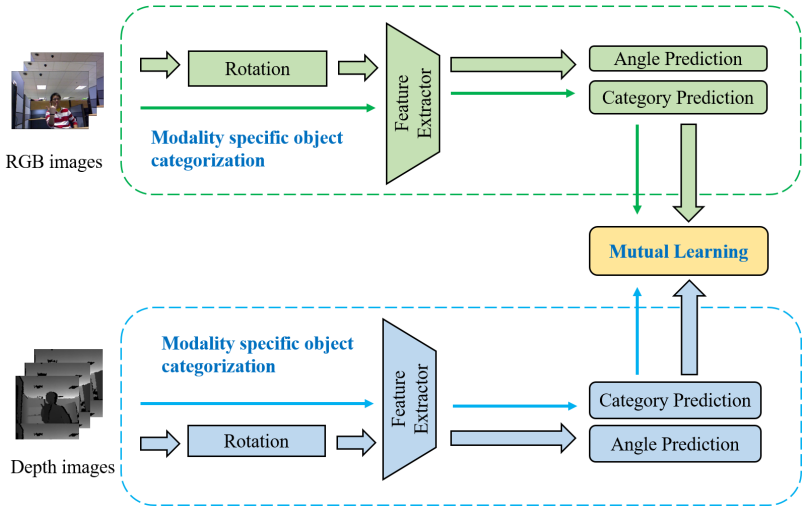

Fig. 1. The framework of the proposed semi-supervised hand gesture recognition approach. We use rotation angle prediction as pretext task to train a self-supervised feature extractor for RGB and depth modality, respectively, and train modality-specific gesture predictors using labeled data. Then we exploit mutual learning to fusion the two predictors. The model about RGB modality is rendered in green while the counterparts about depth modality are rendered in blue; orange module stands for the mutual learning. The wide hollow arrows represent unlabeled data and the thin solid arrows represent labeled data.

3 The Proposed Approach: MLRAGP

3.1 The Framework

We propose a semi-supervised framework for RGB-D hand gesture recognition named MLRAGP. For each of RGB and depth modalities, we adopt SSL to pre-train a feature extractor using both labeled and unlabeled data, based on which we construct the gesture category predictor and optimize the parameters using labeled data. We fuse the two modality-specific gesture predictors through deep mutual learning to enhance the performance. The pretext task of the SSL is rotation angle prediction. Specifically, we rotate the RGB images and depth images to fixed angles to enlarge the sample set and train an angle predictor for each modality to recognize the rotation angles. The feature extraction part of the angle predictor is regarded as the feature extractor. This framework can be illustrated by Figure 1 where the angle predictor and gesture predictor with respect to RGB images are rendered in green and the counterparts about depth images are rendered in blue; orange module stands for the mutual learning. The wide hollow arrows represent large amount of unlabeled data and the narrow solid arrows represent a few labeled data.

3.2 Modality-Specific Semi-Supervised Hand Gesture Recognition Based on Angle Prediction

Suppose the RGB training gesture images are $\mathbf{X} = \{\mathbf{X}^l, \mathbf{X}^u\}$, where $\mathbf{X}^l = \{\mathbf{x}_1^l, \ldots, \mathbf{x}_{N_l}^l\}$ denote the labeled images and $\mathbf{X}^u = \{\mathbf{x}_1^u, \ldots, \mathbf{x}_{N_u}^u\}$ denote the unlabeled images. The depth gesture images are similarly represented as $\mathbf{Y} = \{\mathbf{Y}^l, \mathbf{Y}^u\}$ where $\mathbf{Y}^l = \{\mathbf{y}_1^l, \ldots, \mathbf{y}_{N_l}^l\}$ and $\mathbf{Y}^u = \{\mathbf{y}_1^u, \ldots, \mathbf{y}_{N_u}^u\}$. The annotations of $\mathbf{X}^l$ and $\mathbf{Y}^l$ are $\{\mathbf{v}_1, \ldots, \mathbf{v}_{N_l}\}$ where $\mathbf{v}_i$ is a C-dimensional one-hot vector whose element $\mathbf{v}_i^j = 1$ if $\mathbf{x}_i^l$ or $\mathbf{y}_i^l$ belongs to the jth gesture category and $\mathbf{v}_i^j = 0$ otherwise.

We rotate $\mathbf{X}$ to 90°, 180°, and 270° to form the augmented image set, and merge the original images (the rotation angle is 0) into this set to obtain a training set covering four rotation angles. For simplicity, we re-use $\mathbf{X}$ to denote the training set. Let the angle label of $\mathbf{x}_i$ be a one-hot vector $\mathbf{z}_i = [\mathbf{z}_i(1), \mathbf{z}_i(2), \mathbf{z}_i(3), \mathbf{z}_i(4)]$ where $\mathbf{z}_i(j) = 1$ means j represents the true rotation angle and $\mathbf{z}_i(j) = 0$ otherwise, we can train a rotation angle predictor using this training set through cross-entropy loss. Usually, we use a CNN consisting of some **Convolutional (ConV)** blocks and **Fully Connected (FC)** layers with optional batch normalizations and **Rectified Linear Unit (ReLU)** activations. Suppose $\mathbf{p}_i = [\mathbf{p}_i(1), \mathbf{p}_i(2), \mathbf{p}_i(3), \mathbf{p}_i(4)]$ is the probability estimation produced by the softmax function, the cross-entropy loss can be denoted as

$$\text{Loss}_{\text{angle}} = -\sum_{i=1}^{4(N_l+N_u)} \sum_{j=1}^{4} \mathbf{z}_i(j) \log\left(\mathbf{p}_i(j)\right). \tag{1}$$

Once the training of the angle predictor converges, we fix the feature extraction part of the angle predictor (the last convolutional layer and the network layers before), and link several FC layers with optional batch normalizations and ReLU activations followed by a softmax function to build a gesture category predictor. Subsequently, we can optimize the parameters of the FC layers using the labeled images through cross-entropy loss. Assume the output of the last convolutional layer about $\mathbf{x}_i^l$ is $\tilde{\mathbf{x}}_i^l$, the gesture category predictor with one FC layer and ReLU activation can be described as

$$\begin{cases} \mathbf{g}_i^l = \text{BN}\left(\mathbf{W}^T \tilde{\mathbf{x}}_i^l + \mathbf{b}\right) \\ \mathbf{q}_i^l = \text{softmax}\left(\text{ReLU}(\mathbf{g}_i^l)\right), \end{cases} \tag{2}$$

where BN refers to batch normalization, $\mathbf{q}_i^l = [\mathbf{q}_i^l(1), \ldots, \mathbf{q}_i^l(C)]$ represent probability of category, and $\mathbf{W}$ and $\mathbf{b}$ are parameters. The cross-entropy loss function can be denoted as

$$\text{Loss}_{\text{class}} = -\sum_{i=1}^{N_l} \sum_{j=1}^{C} \mathbf{v}_i(j) \log\left(\mathbf{q}_i^l(j)\right). \tag{3}$$

Gesture category predictor with more FC layers can easily be constructed based on Equation (2).

The rotation angle predictor and gesture category predictor with respect to depth images can be constructed in the same way, so we omit the detailed descriptions for simplicity. The modality-specific semi-supervised hand gesture recognition can be illustrated by Figure 2 where ConV is the feature extraction part of the angle predictor, FC may contain batch normalization, and the brackets means ReLU is optional. The black arrows represent both unlabeled and labeled data while the red arrows represent only the labeled data. The dashed arrow denotes the back-propagation of the gradient, which means that the parameter update of the feature extractor is allowed.

3.3 Mutual Learning of Modality-Specific Gesture Predictors

We adopt deep mutual learning to discover the complement information of two modality-specific gesture predictors and fuse them for performance improvement. The estimated vector of the gesture predictor is not a one-hot vector. Each element of the estimated vector represents the probability that current sample belongs to a specific category. Therefore, the output of the category predictor contains abundant information denoting the similarities between the gesture categories, which are missing in the ground truth one-hot category annotations. The key to mutual learning is to let each of the two modality-specific gesture predictors to simulate the probabilities estimated by the other predictor, such that the correct similarities between the categories can be discovered by the complementary effort of the two predictors. This inter-category similarity should be useful in gesture recognition.

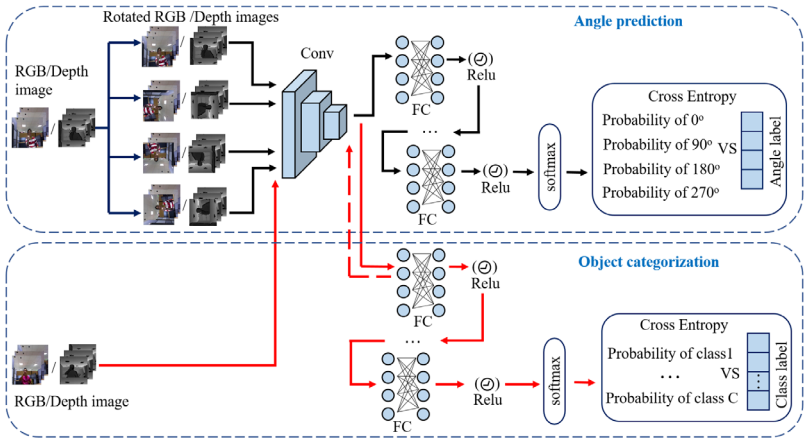

Fig. 2. The modality-specific semi-supervised hand gesture recognition. ConV is the feature extractor, FC may contain batch normalization, and the brackets mean the ReLU is optional. The black arrows represent all the data while the red arrows represent the labeled data. The dashed arrow denotes the back-propagation of the gradient.

Specifically, we fix the feature extraction parts of both modality-specific gesture predictors and link respective feature extractor to multiple FC layers with optional batch normalization and ReLU activation followed by a softmax function. The mutual learning is realized by minimizing the bi-directional **Kullback-Leibler (KL)**-Divergence between the probabilities produced by the softmax functions. In order to guarantee the results of mutual learning do not deviate from the ground truth, we simultaneously expect the output of each modality-specific gesture predictor to approach the annotations for those labeled data.

Let $\tilde{\mathbf{x}}_i$ be the features of RGB images wherein the labeled images are $\tilde{\mathbf{x}}_i^l$, $\tilde{\mathbf{y}}_i$ be the features of depth images wherein the labeled images are $\tilde{\mathbf{y}}_i^l$, one layer category probability estimation with batch normalization and ReLU activation can be described by

$$\begin{cases} \mathbf{g}_i^{RGB} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathbf{W}_{RGB}^T \tilde{\mathbf{x}}_i + \mathbf{b}_{RGB}\right)\right) \\ \mathbf{q}_i^{RGB}(j) = \dfrac{\exp(\mathbf{g}_i^{RGB}(j)/\tau)}{\sum_{j=1}^{C} \exp(\mathbf{g}_i^{RGB}(j)/\tau)} \\ \mathbf{g}_i^{depth} = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathbf{W}_{depth}^T \tilde{\mathbf{y}}_i + \mathbf{b}_{depth}\right)\right) \\ \mathbf{q}_i^{depth}(j) = \dfrac{\exp(\mathbf{g}_i^{depth}(j)/\tau)}{\sum_{j=1}^{C} \exp(\mathbf{g}_i^{depth}(j)/\tau)} \end{cases}, \tag{4}$$

where $\mathbf{W}$ and $\mathbf{b}$ are parameters, $\mathbf{q}_i^{RGB}(c)$ and $\mathbf{q}_i^{depth}(c)$ represent the probability that the RGB image and depth image belong to the cth category, and τ is a hyper-parameter manipulating the difference between the elements of $\mathbf{q}_i^{RGB}$ or $\mathbf{q}_i^{depth}$. $\mathbf{q}_i^{RGB}(j)$ and $\mathbf{q}_i^{depth}(j)$ will be used in the subsequent mutual learning as soft category labels to guide the prediction result from the other modality. Therefore, the mutual learning can be regarded as mutual distillation of two neural networks. According to the principle of network distillation, soft labels reflect the similarity between difference categories that is useful to classification, and should be smooth to suppress the potential over-fitting and decrease the influence of the noise contained in the labels. However, the Sofmaxsoftmax operation

in (4) enlarges the differences between items in the soft labels to make $\mathbf{q}_i^{RGB}(j)$ and $\mathbf{q}_i^{depth}(j)$ sharper. In order to obtain smoothed category probabilities, we append the hyper-parameter τ in the distillation procedure (4). In training stage, we assign τ a bigger value between 20 and 35, and let $\tau = 1$ in prediction stage to enable robust mutual learning.

The KL-Divergence that matches the probability estimation of the gesture predictor about RGB images to that of the gesture predictor about depth images is

$$\mathrm{D_{KL}}(\mathbf{q}_i^{depth}|\mathbf{q}_i^{RGB}) = \sum_{i=1}^{N_l+N_u}\sum_{j=1}^{C}\mathbf{q}_i^{depth}(j)\log\left(\frac{\mathbf{q}_i^{depth}(j)}{\mathbf{q}_i^{RGB}(j)}\right). \tag{5}$$

Similarly, the KL-Divergence that matches the probability estimation of the gesture predictor about depth images to that of the gesture predictor about RGB images is

$$\mathrm{D_{KL}}(\mathbf{q}_i^{RGB}|\mathbf{q}_i^{depth}) = \sum_{i=1}^{N_l+N_u}\sum_{j=1}^{C}\mathbf{q}_i^{RGB}(j)\log\left(\frac{\mathbf{q}_i^{RGB}(j)}{\mathbf{q}_i^{depth}(j)}\right). \tag{6}$$

Hence, the unsupervised loss of the mutual learning can be represented as

$$\mathrm{Loss_{unsup}} = \beta_1\mathrm{D_{KL}}(\mathbf{q}_i^{depth}|\mathbf{q}_i^{RGB}) + \beta_2\mathrm{D_{KL}}(\mathbf{q}_i^{RGB}|\mathbf{q}_i^{depth}), \tag{7}$$

where β_1 and β_2 are hyper-parameters.

The supervised loss of the mutual learning can be denoted as

$$\mathrm{Loss_{sup}} = -\left(\sum_{i=1}^{N_l}\sum_{j=1}^{C}\mathbf{v}_i(j)\log(\mathbf{q}_i^{l_{RGB}}(j)) \tag{8}$$

$$+\sum_{i=1}^{N_l}\sum_{j=1}^{C}\mathbf{v}_i(j)\log(\mathbf{q}_i^{l_{depth}}(j))\Bigg), \tag{9}$$

where $\mathbf{q}_i^{l_{RGB}}$ and $\mathbf{q}_i^{l_{depth}}$ are the estimated gesture category probabilities about labeled data.

Therefore, the total loss about mutual learning of modality-specific gesture predictors is

$$\mathrm{Loss_{mutual}} = \alpha\mathrm{Loss_{sup}} + (1-\alpha)\mathrm{Loss_{unsup}}, \tag{10}$$

where α is a hyper-parameter manipulating the ratio of supervised loss in the mutual learning loss. Mutual learning with multiple FC layers can be similarly constructed.

The mutual learning can be illustrated by Figure 3 where green color corresponds to RGB modality, blue color corresponds to depth modality, and orange color represents constraints with KL-Divergence. The wide solid arrows stand for both labeled and unlabeled data while the thin solid arrows stand for only the labeled data. Similarly, the dashed arrows mean that the parameter update of the feature extractors is allowed. It is noteworthy that the mutual learning consists of two tasks. The unsupervised task is to prompt the two modality-specific predictors to yield consistent predictions with each other via KL-Divergence loss. Since this process does not need the labels of samples, both labeled and unlabeled training data are used to compute the KL-Divergence, and the unlabeled data are the same as those used in training the modality-specific gesture predictors that are shown in Figure 2. The supervised task is to prompt the prediction results of the labeled data to approach the true labels.

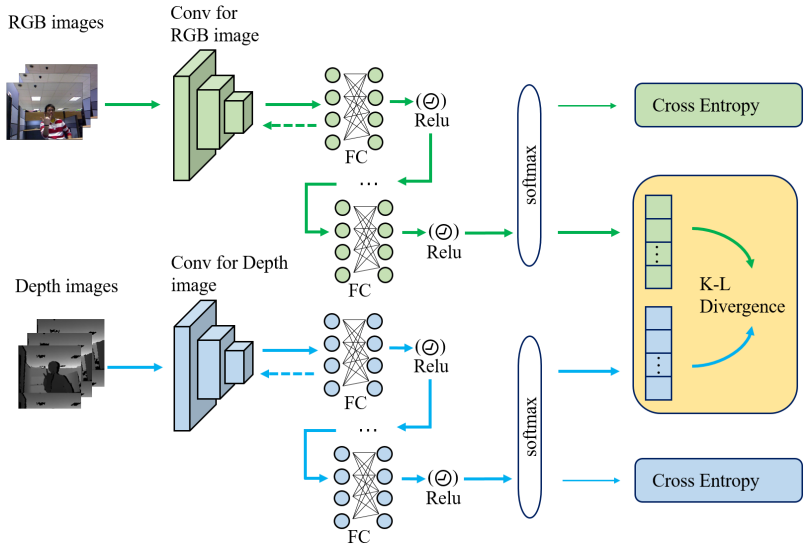

Fig. 3. The mutual learning of modality-specific gesture predictors. Green color corresponds to RGB modality, blue color corresponds to depth modality, and orange color depicts constraints with KL-Divergence. The wide arrows represent both labeled and unlabeled data while the thin arrows represent the labeled data. The dashed arrows mean that the parameter update of the feature extractors is allowed.

3.4 Implementation Details

In our implementation, we use ResNet-18 and modified VGG-11 as the backbone feature extractors. For VGG-11, we alter the kernel numbers of eight convolutional layers to [16, 32, 64, 64, 128, 128, 128, 128], and the kernel size equals 3 with padding size 1. Between the convolutional layers, we also insert ReLU activations and max poolings. An average pooling generates the features of the VGG-11. In angle prediction, the VGG-based feature extractor is connected to a 6272×100 linear mapping with batch normalization and ReLU, then ends up with a 100×4 linear mapping to generate the angle probability. In gesture category predictors, the feature extractor is linked to three FC blocks with batch normalizations and ReLU activations, and the linear mapping of FCs is of 6272×2048, 2048×512, and 512×51, respectively.

For ResNet-18, we directly use the basic network structure with an average pooling for feature extraction. In angle prediction, a simple 25088×4 linear mapping is adopted to generate the angle probability. In gesture predictors, we use the same FC blocks as used by the VGG-based category predictor.

Equation (7) defines the objective of the mutual learning, which aims to simulate the predicted category probability from one modality using that from the other modality. Specifically, the first term of Equation (7) matches the prediction from RGB images to that from the depth images while the second term matches the prediction from depth images to that from the RGB images. Obviously, Equation (7) has a symmetric structure and the two terms contribute equally to the mutual learning process. For this reason, we set β_1 and β_2, the hyper-parameters adjusting the importance of the two terms, as 0.5, respectively.

4 Experiments

This article adopts NTU-Microsoft Kinect Hand Gesture dataset and Washington RGB-D object dataset [19] to evaluate the proposed framework. This section firstly introduces the datasets and

Table 1. The Dataset Configurations Adopted by the Experiments

Dataset	Object No.	Image No. per object	Experiments	Ratio of training data	Ratio of testing data	Ratio of labeled training data
NTU-Microsoft Hand Gesture	10	100	Ablation study and comparative study with existing semi-supervised methods	9/10	1/10	1/5 1/10 1/20
Washington RGB-D	300	600–800	Ablation study and comparative study with existing semi-supervised methods	9/10	1/10	1/20 1/50 1/100 1/200
			Comparative study with existing RGB-D categorization methods	1/6	1/30	1/120

data configurations adopted by the experiments, and then presents the ablation study results aiming to justify the effect of each part of the framework. At last, we present the comparisons between the MLRAGP algorithm and typically adopted semi-supervised learning algorithms to show the efficacy of proposed method. The evaluation metrics used in the experiments include accuracy (Acc), Precision, and Recall with respective to each gesture category in recognition.

4.1 The Dataset

NTU-Microsoft Kinect Hand Gesture dataset covers 10 different hand gestures, and each gesture is performed by 10 persons with 10 variations. The hand gestures are captured through Kinect sensors, yielding a color image and the corresponding depth map for each of the poses. Therefore, this dataset consists of $10 \times 10 \times 10 = 1{,}000$ RGB-D image pairs.

We conduct the ablation study and comparative experiments with existed semi-supervised algorithms on the NTU-Microsoft Kinect Hand Gesture dataset. In these experiments, we use $\frac{9}{10}$ of all the data as the training set and use the rest $\frac{1}{10}$ data as testing set. In order to explore the performance of the algorithms, we set the ratio of labeled data in the training set to $\frac{1}{5}$, $\frac{1}{10}$, and $\frac{1}{20}$, respectively, and pretend not to know the labels of rest data in the training set. The dataset configuration is shown in the upper part of Table 1.

Washington RGB-D object dataset contains pairwise video (RGB) and depth (D) image sequences captured from 300 household objects. For each object, there are three RGB-D sequences captured with three different tilt angles in vertical direction. Each RGB or depth image sequence consists of over 200 frames captured with various pan angles in horizontal direction. The RGB-D sequences are recorded and synchronized by a Kinect style 3D camera. The 300 objects belong to 51 categories. Therefore, the whole dataset includes over 200,000 RGB-D image pairs, with 600–800 pairs for each object.

In the ablation study and comparative experiments with semi-supervised algorithms, we use $\frac{9}{10}$ of all the data as the training set and use the rest $\frac{1}{10}$ data as testing set. In order to explore the performance of the algorithms, we set the ratio of labeled data in the training set to $\frac{1}{20}$, $\frac{1}{50}$, $\frac{1}{100}$, $\frac{1}{200}$, and $\frac{1}{300}$, respectively, and pretend not to know the labels of rest data in the training set. In the comparative experiments with existing RGB-D object categorization algorithms, we sub-sample all the data at five frame intervals to obtain a reduced sample set, and use $\frac{5}{6}$ of the reduced sample

Table 2. Ablation Study with Respect to Different Ratio of Labeled Training Data and Different Mutual Learning Hyper-Parameters on NTU-Microsoft Kinect Hand Gesture Dataset

Backbone	Ratio of labeled training data	Modality	W/O mutual learning	Mutual learning		
				$\tau = 20, \alpha = 0.95$	$\tau = 30, \alpha = 0.95$	$\tau = 35, \alpha = 0.9$
VGG-11	1/5	depth	0.59	0.59	0.57	0.61
		RGB	0.90	0.93	0.94	0.91
	1/10	depth	0.50	0.36	0.38	0.39
		RGB	0.72	0.74	0.73	0.71
ResNet-18	1/5	depth	0.60	0.66	0.62	0.63
		RGB	0.87	0.89	0.91	0.89
	1/10	depth	0.35	0.38	0.37	0.38
		RGB	0.71	0.69	0.70	0.73
	1/20	depth	0.22	0.29	0.26	0.22
		RGB	0.39	0.42	0.37	0.39

set as the training set and use the rest $\frac{1}{6}$ of the reduced sample set as the testing set. Within the training set, we set the ratio of labeled data to $\frac{1}{20}$ and pretend not to know the label of the rest data in the training set. This is to comply with the experimental setup of existing algorithms for comparison. The dataset configuration is shown in the lower part of Table 1.

4.2 The Ablation Study

The ablation experiments are conducted on both datasets. First, we compute the ACCs of the MLRAGP approach with different ratios of labeled training data and different mutual learning hyper-parameters on NTU-Microsoft Kinect Hand Gesture dataset. The backbone of MLRAGP exploits VGG-11 and ResNet-18, the ratio of labeled training data is set to 1/5, 1/10, and 1/20, respectively, and the hyper-parameters refer to τ in Equation (4) and α in Equation (10). The results are shown in Table 2, where both the ACCs of modality-specific gesture recognition and those of gesture recognition after mutual learning are demonstrated.

It can be observed that with the increase of labeled training data, the recognition results improve accordingly and this is consistent with commonsense. Also, the recognition results based on RGB images are much better than those based on depth images, and this is owing to the fact that RGB images contain much more discriminative information than depth images. More importantly, Table 2 shows that mutual learning enhances the modality-specific gesture recognition. When τ lies between 20 and 35, α lies between 0.9 and 0.95, mutual learning obtains obvious performance improvement. Specifically, the recognition accuracy based on ResNet-18 with 1/5 and 1/20 labeled depth samples rises from 0.6 and 0.22 to 0.66 and 0.29 after mutual learning, the ACC based on VGG-11 with 1/5 labeled RGB samples rises from 0.9 to 0.94, the ACC based on ResNet-18 with 1/5 labeled RGB samples rises from 0.87 to 0.91. The reason is that mutual learning well exploits the complement information of the RGB and depth modality. Surprisingly, the result of VGG-11 with 1/10 labeled depth samples decreases after mutual learning. We think the reason is that the structure of VGG-11 is simpler than ResNet-18 with much less parameters, and this degrades its ability to discover high-quality features from small quantity of high-resolution (640×640) depth images. This introduces semantic inaccuracy into the subsequent mutual learning procedure and causes the decrease of accuracy.

Also, we test the performance of different model configurations under different ratios of labeled training data on Washington RGB-D object dataset. This dataset contains much more samples than NTU-Microsoft Kinect Hand Gesture dataset. The involved model configurations include

Table 3. Ablation Study with Respect to Different Ratios of Labeled Training Data on Washington RGB-D Dataset

Backbone	Methods	Ratio of labeled training data			
		1/20	1/50	1/100	1/200
VGG-11	Supervised (depth)	0.8528	0.5828	0.5343	0.4475
	Supervised (RGB)	0.9789	0.9117	0.8151	0.7434
	MLRAGP (depth)	0.8879	0.6959	0.6309	0.5686
	MLRAGP (depth) after fusion	0.8070	0.7027	0.6365	0.5744
	MLRAGP (RGB)	0.9818	0.9421	0.8701	0.8151
	MLRAGP (RGB) after fusion	0.9845	0.9474	0.8773	0.8203
ResNet-18	Supervised (depth)	0.8150	0.5083	0.4591	0.4031
	Supervised (RGB)	0.9525	0.8538	0.7444	0.6578
	MLRAGP (depth)	0.9089	0.7481	0.6736	0.6201
	MLRAGP (depth) after fusion	0.8886	0.7630	0.6909	0.6305
	MLRAGP (RGB)	0.9859	0.9454	0.8708	0.8017
	MLRAGP (RGB) after fusion	0.9880	0.9550	0.8864	0.8259

"MLRAGP (depth)" (depth image-based recognition), "MLRAGP (depth) after fusion" (recognition by mutual learning), "MLRAGP (RGB)" (RGB image-based recognition), and "MLRAGP (depth) after fusion" (recognition by mutual learning). We also apply fully supervised recognition to either RGB images ("Supervised (RGB)") or depth images ("Supervised (depth)") to provide baseline results.

The recognition ACCs based on VGG-11 and ResNet-18 backbone are shown in Table 3. It is obvious that "MLRAGP (depth)" and "MLRAGP (RGB)" yield better results than the baselines (supervised methods) on the two modalities regardless of either VGG or ResNet is used. The improvement is even more significant when fewer labeled training data are used. This is owing to the contribution of the unlabeled training data. It can be generally observed that the mutual learning can improve "MLRAGP (depth)" and "MLRAGP (RGB)" further. Exception appears in the learning of the depth images. The ACC of "MLRAGP (depth)" with VGG backbone drops for over 8% points after mutual learning while the ACC of "MLRAGP (depth)" with ResNet backbone drops for over 2% points, when $\frac{1}{20}$ training data have labels. The reason is that $\frac{1}{20}$ labeled data are enough to train a good classifier, and the mutual simulation of the probability estimation may interfere with the trained classifier.

Another discovery is that the less labeled training data are used, the more significant performance gain of rotation-aware pre-training and mutual learning can be observed. For example, in case of VGG backbone, the ACC of "MLRAGP (depth)" exceeds the ACC of "Supervised (depth)" by 3% points using $\frac{1}{20}$ labeled training data while the counterpart improvement using $\frac{1}{50}$ labeled training data surpasses 11% points. In case of ResNet backbone, the ACC of "MLRAGP (depth)" exceeds the ACC of "Supervised (depth)" by 9% points using $\frac{1}{20}$ labeled training data while the ACC of of "MLRAGP (depth)" exceeds that of "Supervised (depth)" by nearly 25% points using $\frac{1}{50}$ labeled training data. The ablation experimental results justify the effect of the fine-tuning of the feature extractors and the multi-modal fusion of modality-specific gesture predictors through mutual learning.

4.3 The Comparative Experiments

So far as we know, there is very few work about semi-supervised RGB-D hand gesture recognition. Therefore, this subsection compares the MLRAGP framework with several semi-supervised learning

Table 4. Comparison with Popular Semi-Supervised Methods with Different Ratios of Labeled Training Data on NTU-Microsoft Kinect Hand Gesture Dataset

Methods	Ratio of labeled training data			
	1/5		1/10	
	Depth	RGB	Depth	RGB
Deep co-train (VGG) [30]	0.31	0.61	0.30	0.35
Pseudo label (VGG) [21]	0.32	0.74	0.27	0.58
CBST (VGG) [23]	0.32	0.66	0.24	0.61
MetaCL (VGG) [22]	0.45	0.79	0.26	0.63
USCL (VGG) [58]	0.50	0.72	0.31	0.55
RoPAWS (VGG) [28]	0.55	0.77	0.34	0.63
MLRAGP(VGG)	0.57	0.94	0.36	0.74
Deep co-train (ResNet) [30]	0.11	0.11	0.12	0.11
Pseudo label (ResNet) [21]	0.39	0.87	0.36	0.62
CBST (ResNet) [23]	0.38	0.82	0.38	0.71
MetaCL (ResNet) [22]	0.59	0.87	0.34	0.72
USCL (ResNet) [58]	0.55	0.91	0.37	0.68
RoPAWS (ResNet) [28]	0.56	0.91	0.36	0.65
MLRAGP (ResNet)	0.62	0.91	0.38	0.73

methods including two commonly used algorithms (pseudo label [21], deep co-training [30]) and four state-of-the-art algorithms (USCL [58], CBST [23], MetaCL [22], and RoPAWS [28]). According to pseudo label method [21], an initial classifier is trained based on the labeled training data and is subsequently used to predict pseudo labels for the unlabeled data. Then the data with pseudo labels are progressively appended to the labeled dataset for model update. Since the pseudo label method is not designed for multi-modal learning, we apply it to RGB and depth modality, respectively, and fuse the two modality-specific classifiers through mutual learning. Deep co-training method [30] trains two classifiers with two independent labeled training subsets and generates adversarial samples from unlabeled training data for each classifier. The classification of adversarial samples is difficult to the current classifier but easier to the other one. The co-training loss expects the prediction of the original unlabeled data by each classifier to approach the prediction of the adversarial samples by the other classifier. In the meanwhile, cross-entropy loss is applied to those labeled data. If applied to RGB-D recognition, each of the two classifiers is supposed to address RGB and depth modality, respectively. Therefore, each classifier needs to accept adversarial samples of the other classifier (modality). However, this actually degenerates the performance of the classifier owing to the modality mismatch. For this reason, we also apply deep co-training to RGB and depth modality, respectively, and fuse the two predictors through mutual learning. Without using image augmentation, USCL [58] lacks the ability to discover semantic features that are invariant across different views of a same RGB or depth image. Although CBST [23], MetaCL [22], and RoPAWS [28] perturb original images for consistency or contrast-based training, the perturbation is random augmentation that does not intentionally focus on fine-grained variation on object shape, contour, and orientation. We adopt both VGG and ResNet as the backbone network to test the methods.

The comparison between the ACCs of MLRAGP framework and existed semi-supervised methods with different ratios of labeled training data on NTU-Microsoft Kinect Hand Gesture dataset is shown in Table 4. With 1/5 and 1/10 labeled training data, the MLRAGP with either ResNet or VGG backbone yields the best results for RGB and depth image recognition. For instance, in RGB

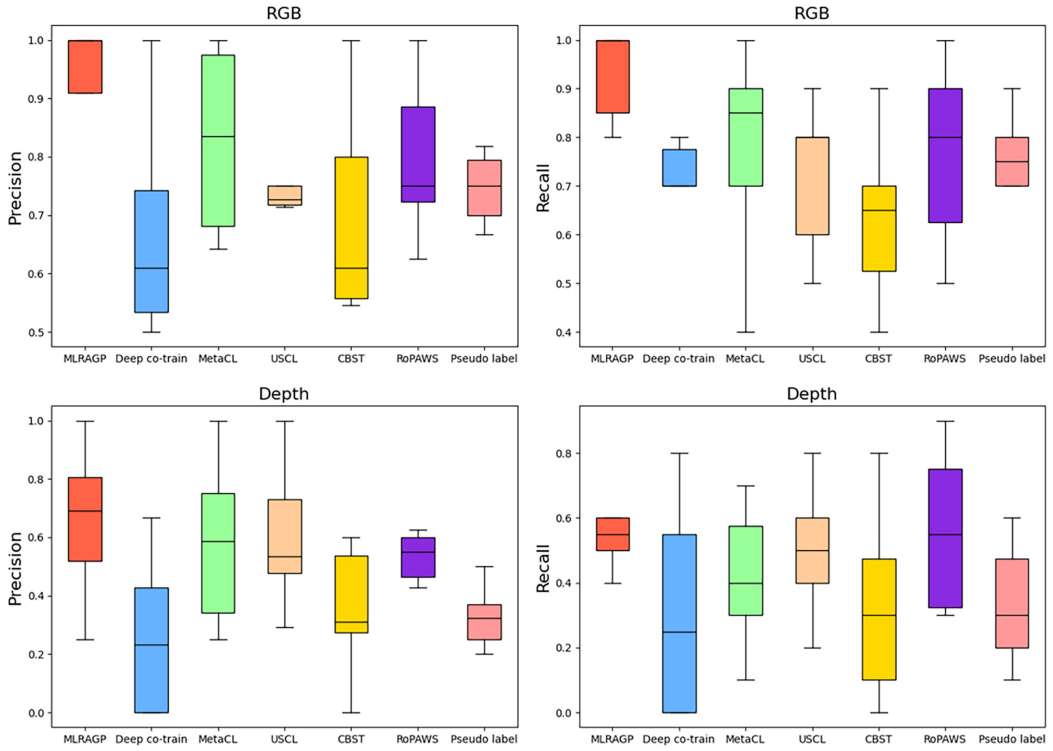

Fig. 4. The statistic about Precisions and Recalls of existed semi-supervised methods and MLRAGP based on VGG backbone with $\frac{1}{5}$ labeled training data.

image recognition with VGG backbone, the MLRAGP outperforms the state-of-the-art methods RoPAWS and MetaCL by 11% points with 1/10 labeled samples, and by over 15% points with 1/5 labeled samples. This can be attributed to the rotation-aware pre-training under mutual learning framework. Specifically, the rotation-aware pre-training enables the gesture predictors to perceive the shape and orientation of gestures based on large volume of unlabeled data and the mutual learning enables the gesture predictors to learn from each other to disclose the category similarity hiding in the unlabeled data. It can be observed that the deep co-training method yields obviously worse result than the pseudo label method. We believe the reason may be two-fold. First, the adversarial samples are generated through image augmentation technique that cannot be applied to depth images well. Second, the gesture dataset has great intra-category similarity, which means the adversarial sample is prone to resemble another original sample of the same category and this will weaken the complementarity of the two gesture predictors.

We also calculate the Precision and Recall of MLRAGP and aforementioned semi-supervised methods with respect to each of the gesture category, and visualize the statistics (box-plot) about the Precisions and Recalls in Figures 4 and 5. Figure 4 shows the Precision and Recall based on VGG backbone while Figure 5 shows the results based on ResNet backbone; both models are trained using 1/5 labeled training data. It is obvious that VGG-based MLRAGP yields the best average Precision and Recall in gesture prediction using either RGB image or Depth image. This can be attributed to the ability of MLRAGP to perceive fine-grained shape, contour, and orientation of hand that is critical to gesture recognition. ResNet-based MLRAGP also obtains the best average

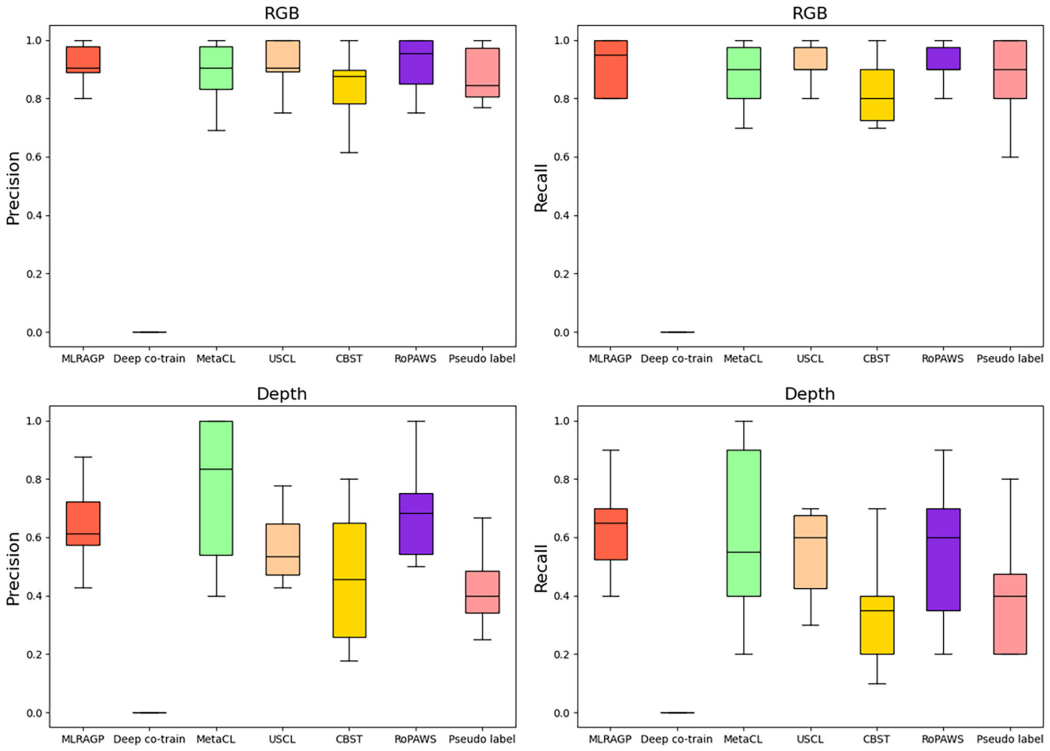

Fig. 5. The statistic about Precisions and Recalls of existed semi-supervised methods and MLRAGP based on ResNet backbone with $\frac{1}{5}$ labeled training data.

Recall in either RGB or Depth modality prediction, but fails to outperform MetaCL and RoPAWS in average Precision. The reason is that ResNet-18 has much less parameters compared with VGG-11, which impairs the ability of MLRAGP to perceive the fine-grained object shape feature well. This enlightens us to choose models with appropriate complexity in MLRAGP according to the data volume.

The comparison of ACCs with different ratios of labeled training data on Washington RGB-D object dataset is shown in Table 5. Similarly, the MLRAGP with ResNet or VGG backbone yields the best results. The reason has been discussed in the analysis about the results of Table 4. It is noteworthy that MLRAGP with VGG backbone is slightly inferior to pseudo labeling in predicting the RGB images based on $\frac{1}{50}$ labeled training data, and we think this is caused by normal fluctuation which depends on the training data.

Finally, we compare MLRAGP with a group of supervised and semi-supervised object classification algorithms. To be consistent with the data configuration adopted by the comparative algorithms, we use a reduced training set to evaluation MLRAGP, and the detailed data configuration can be found in Table 1. For those fully supervised methods, all the data in the reduced training set have been manually annotated. For MLRAGP, about 1/20 of the images in the reduced training set have labels. The ACCs of category predictors about RGB image, depth image, and the fused predictors are shown in Table 6. For MLRAGP, we only demonstrate the ACC of the best category predictor after mutual learning. It can be observed that the ACC of the MLRAGP algorithm on RGB image classification outperforms other methods by a large margin. It even succeeds the fully supervised methods DCNN [59] (the second best method) by at least 6% points. This demonstrates the

Table 5. Comparison with Popular Semi-Supervised Methods with Different Ratios of Labeled Training Data on Washington RGB-D Object Dataset

Methods	Ratio of labeled training data					
	1/50		1/100		1/200	
	Depth	RGB	Depth	RGB	Depth	RGB
Deep co-train (VGG) [30]	0.5456	0.8229	0.4088	0.7541	0.3310	0.6159
Pseudo label (VGG) [21]	0.6477	0.9584	0.5681	0.8674	0.4240	0.7243
CBST (VGG) [23]	0.4040	0.6591	0.3623	0.4685	0.3169	0.3879
MetaCL (VGG) [22]	0.6098	0.9213	0.5427	0.8154	0.4459	0.7438
USCL (VGG) [58]	0.5816	0.9069	0.4933	0.7810	0.4009	0.7011
RoPAWS (VGG) [28]	0.5906	0.9081	0.5312	0.8315	0.4629	0.7359
MLRAGP (VGG)	0.7027	0.9474	0.6365	0.8773	0.5744	0.8203
Deep cotrain (ResNet) [30]	0.4586	0.6484	0.2879	0.5095	0.1958	0.2832
Pseudo label (ResNet) [21]	0.6422	0.9256	0.4896	0.7949	0.3375	0.7035
CBST (ResNet) [23]	0.2650	0.7215	0.2357	0.5990	0.1723	0.3246
MetaCL (ResNet) [22]	0.5117	0.8730	0.4608	0.7633	0.3757	0.6636
USCL (ResNet) [58]	0.3224	0.6209	0.2928	0.4848	0.2075	0.3318
RoPAWS (ResNet) [28]	0.5123	0.8480	0.4696	0.7584	0.3889	0.6572
MLRAGP (ResNet)	0.7630	0.9550	0.6909	0.8864	0.6305	0.8259

Table 6. Comparison with Existing RGB-D Object Categorization Methods on Washington RGB-D Object Dataset

Supervised method	**Depth**	**RGB**	**Fused**
Linear SVM [19]	0.531	0.743	0.819
Kernel SVM [19]	0.647	0.745	0.838
Random forest [19]	0.668	0.747	0.796
IDL [20]	0.702	0.786	0.854
3D SPMK [32]	0.678	\	\
KDES [3]	0.788	0.777	0.862
CKM [2]	\	\	0.864
HMP [4]	0.703	0.747	0.821
SP-HMP [5]	0.812	0.824	0.875
CNNRNN [36]	0.789	0.808	0.868
CNN [35]	\	0.831	0.894
R^2ICA [18]	0.839	0.857	0.896
FusCNN (HHA) [14]	0.830	0.841	0.910
FusCNN (jet) [14]	0.838	0.841	0.913
Warping [8]	\	\	0.92.7
NMSS [40]	0.756	0.746	0.885
CFK [9]	0.858	0.868	0.912
DCNN [59]	0.7843	0.8896	0.918
Semi-supervised method	**Depth**	**RGB**	**Fused**
CT + SVM1 [11]	0.718	0.771	0.816
CT + SVM2 [12]	0.754	0.787	0.837
DP co-training [10]	0.826	0.855	0.892
MLRAGP (VGG)	0.7190	0.9586	0.9648
MLRAGP (ResNet)	0.7810	0.9642	0.9733

significant performance gain yielded by the rotation-aware pre-training. For depth images, MLRAGP does not yield competitive results, the reason might be the choice of backbone network. Once a network more suitable for depth images is adopted, we believe the classification performance of depth images will be improved. Even though the classification about depth images is not satisfactory, the fused MLRAGP still has the best classification result. Specifically, MLRAGP (ResNet) outperforms the second best semi-supervised classification method, i.e., diversity preserving co-training, by over 8% points, and it even succeeds the best supervised method [8] in Table 6, for over 4% points. This reveals that the mutual learning improves the performance further.

5 Conclusion

Current RGB-D hand gesture recognition algorithms may suffer from insufficient labeled training data. We propose a semi-supervised framework named MLRAGP to address this problem. We use rotation angle prediction to pre-train the feature extractors of RGB and depth gesture images with unlabeled data, and then train gesture predictors based on the feature extractors with labeled data. For multi-modal fusion, we adopt mutual learning to make the two gesture predictors to learn from each other for performance boost. We testify the significance of both the rotation-aware pre-training and the mutual learning process through comprehensive experiments. The proposed framework can involve various backbone networks, and the performance of MLRAGP could be further improved by using more powerful backbones.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles. In *27th International Conference on Neural Information Processing Systems - Volume 2*, 3365–3373.

[2] Manuel Blum, Jost Tobias Springenberg, Jan Wülfing, and Martin Riedmiller. 2012. A learned feature descriptor for object recognition in RGB-D data. In 2012 IEEE International Conference on Robotics and Automation, 1298–1303.

[3] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2011a. Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 821–826.

[4] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2011b. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *24th International Conference on Neural Information Processing Systems*, 2115–2123.

[5] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2013. Unsupervised feature learning for RGB-D based object recognition. In *Experimental Robotics. Springer Tracts in Advanced Robotics,* Vol. 88. J. Desai, G. Dudek, O. Khatib, and V. Kumar (Eds.), Springer, Heidelberg.

[6] Koundinya Challa, Issa W. AlHmoud, A. K. M. Kamrul Islam, and Balakrishna Gokaraju. 2023. EMG-based hand gesture recognition using individual sensors on different muscle groups. In *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–4.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *2008 International Conference on Machine Learning (ICML)*, 1597–1607.

[8] Yanhua Cheng, Rui Cai, Chi Zhang, Zhiwei Li, Xin Zhao, Kaiqi Huang, and Yong Rui. 2015a. Query adaptive similarity measure for RGB-D object recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 145–153.

[9] Yanhua Cheng, Rui Cai, Xin Zhao, and Kaiqi Huang. 2015b. Convolutional Fisher kernels for RGB-D object recognition. In *2015 International Conference on 3D Vision*, 135–143.

[10] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui. 2016. Semi-supervised multimodal deep learning for RGB-D object recognition. In *25th International Joint Conference on Artificial Intelligence*. AAAI Press, 3345–3351.

[11] Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan. 2014. Semi-supervised learning for RGB-D object recognition. In *2014 22nd International Conference on Pattern Recognition*, 2377–2382.

[12] Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan. 2015c. Semi-supervised learning and feature evaluation for RGB-D object recognition. *Computer Vision and Image Understanding* 139 (2015), 149–160.

[13] Meng-Tzu Chiu, Hsun-Ying Cheng, Chien-Yi Wang, and Shang-Hong Lai. 2021. High-accuracy RGB-D face recognition via segmentation-aware face depth estimation and mask-guided attention network. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8.

[14] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. 2015. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 681–687.

[15] Ilham El Ouariachi, Rachid Benouini, Khalid Zenkouar, Arsalane Zarghili, and Hakim El Fadili. 2022. RGB-D feature extraction method for hand gesture recognition based on a new fast and accurate multi-channel cartesian Jacobi moment invariants. *Multimedia Tools and Applications* 81, 9 (2022), 12725–12757.

[16] Abdessamad Elboushaki, Rachida Hannane, Karim Afdel, and Lahcen Koutti. 2020. MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications* 139 (2020), 112829.

[17] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. arXiv:1803.07728. Retrieved from http://arxiv.org/abs/1803.07728

[18] I-Hong Jhuo, Shenghua Gao, Liansheng Zhuang, D. T. Lee, and Yi Ma. 2015. Unsupervised feature learning for RGB-D image classification. In *Computer Vision – ACCV 2014*. Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang (Eds.), 276–289.

[19] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2011a. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE International Conference on Robotics and Automation*, 1817–1824.

[20] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. 2011b. Sparse distance learning for object recognition combining RGB and depth information. In *2011 IEEE International Conference on Robotics and Automation*, 4007–4013.

[21] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *2019 International Conference on Machine Learning (ICML)*. AAAI Press, 1–6.

[22] Chengyang Li, Yongqiang Xie, Zhongbo Li, and Liping Zhu. 2024b. MetaCL: A semi-supervised meta learning architecture via contrastive learning. *International Journal of Machine Learning and Cybernetics* 15, 2 (2024), 227–236.

[23] Sai Li, Peng Kou, Miao Ma, Haoyu Yang, Shuo Huang, and Zhengyi Yang. 2024a. Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data. *IEEE Access* 12 (2024), 27331–27343.

[24] Yunan Li, Huizhou Chen, Guanwen Feng, and Qiguang Miao. 2023. Learning robust representations with information bottleneck and memory network for RGB-D-based gesture recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 20911–20921.

[25] Srilakshmi Ravali M., S. Mohamed Mansoor Roomi, B. Sathyabama, and M. Senthilarasi. 2023. Hand gesture recognition system using transfer learning. In *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)*, 1–5.

[26] Yujun Ma, Benjia Zhou, Ruili Wang, and Pichao Wang. 2023. Multi-stage factorized spatio-temporal representation for RGB-D action and gesture recognition. In *31st ACM International Conference on Multimedia*, 3149–3160.

[27] Shaobo Min, Xuejin Chen, Hongtao Xie, Zheng-Jun Zha, and Yongdong Zhang. 2021. A mutually attentive co-training framework for semi-supervised recognition. *IEEE Transactions on Multimedia* 23 (2021), 899–910.

[28] Sangwoo Mo, Jong-Chyi Su, Chih-Yao Ma, Mido Assran, Ishan Misra, Licheng Yu, and Sean Bell. 2023. RoPAWS: Robust semi-supervised representation learning from uncurated data. In *11th International Conference on Learning Representations*, 1–26.

[29] Luntian Mou, Chao Zhou, Pengtao Xie, Pengfei Zhao, Ramesh Jain, Wen Gao, and Baocai Yin. 2023. Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion. *IEEE Transactions on Multimedia* 25 (2023), 529–542.

[30] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *2018 European Conference on Computer Vision (ECCV)*. Springer, 135–152.

[31] Priyal Raj, Sakshi Pandey, Yuvraj Singh, and Monu Singh. 2023. Hand sign & gesture recognition system. In *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Vol. 6, 639–642.

[32] Carolina Redondo-Cabrera, Roberto J. López-Sastre, Javier Acevedo-Rodriguez, and Saturnino Maldonado-Bascón. 2012. SURFing the point clouds: Selective 3D spatial pyramids for category-level object recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3458–3465.

[33] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, Vol. 29, 1171–1179.

[34] Debajit Sarma, H. Pallab Jyoti Dutta, Kuldeep Singh Yadav, M. K. Bhuyan, and Rabul Hussain Laskar. 2024. Attention-based hand semantic segmentation and gesture recognition using deep networks. *Evolving Systems* 15, 1 (2024), 185–201.

[35] Max Schwarz, Hannes Schulz, and Sven Behnke. 2015. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1329–1335.

[36] Richard Socher, Brody Huval, Bharath Bhat, Christopher D. Manning, and Andrew Y. Ng. 2012. Convolutional-recursive deep learning for 3D object classification. In *25th International Conference on Neural Information Processing Systems - Volume 1*, 656–664.

[37] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen. 2020. Image representations with spatial object-to-object relations for RGB-D scene recognition. *IEEE Transactions on Image Processing* 29 (2020), 525–537.

[38] Ying Sun, Yaoqing Weng, Bowen Luo, Gongfa Li, Bo Tao, Du Jiang, and Disi Chen. 2023. Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. *IET Image Processing* 17, 4 (2023), 1280–1290.

[39] Yanhao Tan, Mohammad Muntasir Rahman, Yanfu Yan, Jian Xue, Ling Shao, and Ke Lu. 2022. Fine-grained categorization from RGB-D images. *IEEE Transactions on Multimedia* 24 (2022), 917–928.

[40] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. 2015. MMSS: Multi-modal sharable and specific feature learning for RGB-D object recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1125–1133.

[41] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. 2018. Structure-aware multimodal feature fusion for RGB-D scene classification and beyond. *ACM Transactions on Multimedia Computing, Communications and Applications* 14, 2s, Article 39 (2018), 1–22.

[42] Ruimin Wang, Fasheng Wang, Yiming Su, Jing Sun, Fuming Sun, and Haojie Li. 2023a. Attention-guided multi-modality interaction network for RGB-D salient object detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 3, Article 68 (Oct 2023), 1–22.

[43] Yu Wang, Jun Zhang, and Xinchun Zhao. 2023b. Research on hand gesture recognition based on millimeter wave radar. In *2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 205–209.

[44] Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 23555–23564.

[45] J. Weston, F. Ratle, H. Mobahi, and R. Collobert. 2008. Deep learning via semi-supervised embedding. In *2008 International Conference on Machine Learning (ICML)*. AAAI Press, 1168–1175.

[46] Jingjing Wu, Jianguo Jiang, Meibin Qi, Cuiqun Chen, and Jingjing Zhang. 2022. An end-to-end heterogeneous restraint network for RGB-D cross-modal person re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4, Article 109 (2022), 1–22.

[47] You-Jia Wu and Jian-Jiun Ding. 2023. Real-time hand gesture recognition using depth information and path analysis. In *2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 727–731.

[48] Yi Xiao, Tong Liu, Yu Han, Yue Liu, and Yongtian Wang. 2023. Realtime recognition of dynamic hand gestures in practical applications. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2, Article 50 (Sep 2023), 1–17.

[49] Bin Xie, Xiaoyu He, and Yi Li. [n. d.]. RGB-D static gesture recognition based on convolutional neural network. *The Journal of Engineering* 2018, 16 ([n. d.]), 1515–1520.

[50] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 33, 6256–6268.

[51] Jun Xu, Hanchen Wang, Jianrong Zhang, and Linqin Cai. 2022. Robust hand gesture recognition based on RGB-D data for natural human–computer interaction. *IEEE Access* 10 (2022), 54549–54562.

[52] Jun Yu, Yong Rui, and Dacheng Tao. 2014. Click prediction for web image reranking using multimodal sparse coding. *IEEE Transactions on Image Processing* 23, 5 (2014), 2019–2032.

[53] Jun Yu, Min Tan, Hongyuan Zhang, Yong Rui, and Dacheng Tao. 2022. Hierarchical deep click feature prediction for fine-grained image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 2 (2022), 563–578.

[54] Jun Yu, Dacheng Tao, Meng Wang, and Yong Rui. 2015. Learning to rank using user clicks and visual features for image retrieval. *IEEE Transactions on Cybernetics* 45, 4 (2015), 767–779.

[55] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4320–4328.

[56] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu. 2016. Discriminative multi-modal feature fusion for RGBD indoor scene recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2969–2976.

[57] Lei Zhu, Xiaoqiang Wang, Ping Li, Xin Yang, Qing Zhang, Weiming Wang, Carola-Bibiane Schönlieb, and C. L. Philip Chen. 2023. S^3 Net: Self-supervised self-ensembling network for semi-supervised RGB-D salient object detection. *IEEE Transactions on Multimedia* 25 (2023), 676–689.

[58] Ye Zhu, Jie Yang, Siqi Liu, and Ruimao Zhang. 2024. Inherent consistent learning for accurate semi-supervised medical image segmentation. In *Medical Imaging with Deep Learning (Proceedings of Machine Learning Research, Vol. 227)*, 1581–1601.

[59] Saman Zia, Buket Yüksel, Deniz Yüret, and Yücel Yemez. 2017. RGB-D object recognition using deep convolutional neural networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 887–894.

Received 11 March 2024; revised 9 June 2024; accepted 15 August 2024