electronics

MDPI

Article

Cross-Domain Facial Expression Recognition through Reliable Global–Local Representation Learning and Dynamic Label Weighting

Yuefang Gao [1], Yiteng Cai [1], Xuanming Bi [1], Bizheng Li [1], Shunpeng Li [1] and Weiping Zheng [2,*]

[1] College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China; gaoyuefang@scau.edu.cn (Y.G.); 20213162015@stu.scau.edu.cn (Y.C.); bxm@stu.scau.edu.cn (X.B.); leebz9037@stu.scau.edu.cn (B.L.); lishunpeng@scau.edu.cn (S.L.)

[2] School of Computer Science, South China Normal University, Guangzhou 510631, China

* Correspondence: zhengweiping@scnu.edu.cn

Abstract: Cross-Domain Facial Expression Recognition (CD-FER) aims to develop a facial expression recognition model that can be trained in one domain and deliver consistent performance in another. CD-FER poses a significant challenges due to changes in marginal and class distributions between source and target domains. Existing methods primarily emphasize achieving domain-invariant features through global feature adaptation, often neglecting the potential benefits of transferable local features across different domains. To address this issue, we propose a novel framework for CD-FER that combines reliable global–local representation learning and dynamic label weighting. Our framework incorporates two key modules: the Pseudo-Complementary Label Generation (PCLG) module, which leverages pseudo-labels and complementary labels obtained using a credibility threshold to learn domain-invariant global and local features, and the Label Dynamic Weight Matching (LDWM) module, which assesses the learning difficulty of each category and adaptively assigns corresponding label weights, thereby enhancing the classification performance in the target domain. We evaluate our approach through extensive experiments and analyses on multiple public datasets, including RAF-DB, FER2013, CK+, JAFFE, SFW2.0, and ExpW. The experimental results demonstrate that our proposed model outperforms state-of-the-art methods, with an average accuracy improvement of 3.5% across the five datasets.

Keywords: facial expression recognition; pseudo-label learning; label dynamic weight matching; domain adaptation

Citation: Gao, Y.; Cai, Y.; Bi, X.; Li, B.; Li, S.; Zheng, W. Cross-Domain Facial Expression Recognition through Reliable Global–Local Representation Learning and Dynamic Label Weighting. *Electronics* **2023**, *12*, 4553. https://doi.org/10.3390/electronics12214553

Academic Editor: Daniele Riboni

Received: 4 October 2023
Revised: 27 October 2023
Accepted: 5 November 2023
Published: 6 November 2023

1. Introduction

CD-FER is the task of automatically recognizing and inferring human emotional states across different domains, playing a significant role in human–computer interaction [1], affective computing [2], and similar applications. Unlike traditional facial expression recognition methods that operate within a single dataset [3–6], CD-FER faces significant challenges due to subtle variations between different facial expression categories and substantial differences among facial expression recognition datasets. Over the past decade, several CD-FER methods have been proposed to address the performance degradation caused by data inconsistency. These methods have been extensively evaluated using popular FER datasets, including RAF-DB [7], FER2013 [8], CK+ [9], JAFFE [10], SFW2.0 [11], and ExpW [12]. Earlier research primarily addressed this problem using techniques such as transfer learning [13] and supervised kernel mean matching [14]. However, these methods require several annotated samples in the target domain, making them unsuitable for unsupervised CD-FER scenarios. Subsequently, other learning strategies have been introduced, such as dictionary learning [15], metric learning [16], and contrastive learning [17]. These strategies enable cross-domain learning even without labeled data in the target domain.

Researchers have recently explored domain adaptation models to tackle the CD-FER task. These models incorporate adversarial learning mechanisms [18–20] to acquire transferable domain-invariant features. However, most of these models focus on extracting global features for domain adaptation, overlooking advantages of local features such as their greater transferability across different domains and their ability to provide fine-grained feature representation. A number of studies [21,22] have modeled the correlations between global and local features within and across domains as a way to mitigate domain drift. Furthermore, a few recent works [23,24] have incorporated semantic information into multi-view feature learning in order to narrow the semantic gap in the domain adaptation process. Unfortunately, these methods have often neglected the impact of imbalanced class distribution across various domains, resulting in unsatisfactory recognition performance.

To address these issues, we propose a novel framework for addressing the performance degradation caused by data inconsistency in cross-domain scenarios. Our framework focuses on global–local feature learning and dynamic label weighting based on consistency regularization. It introduces two main modules, PCLG and LDWM. In the PCLG module, we generate pseudo-labels and complementary labels for target domain samples, applying a fixed threshold to filter reliable pseudo-labels. Collaborative training of pseudo-labels and complementary labels helps to mitigate the performance degradation caused by noisy labels. The LDWM module adjusts class weights based on the number of generated pseudo-labels, reflecting the importance of these classes. This ensures that the model pays more attention to minority classes, thereby mitigating the class imbalance issue. Moreover, we incorporate local features and construct multiple classifiers to learn global and local features and select the best classification performance.

Our contributions can be summarized as follows: (i) we propose a global–local feature learning and dynamic label weighting framework to address domain shift; (ii) we develop a pseudo-complementary label generation method to help calibrate confirmation biases; (iii) we introduce a dynamic label weighting matching strategy to mitigate the impact of class imbalance. Codes and trained models are available at https://github.com/chttyte//GLRLDLW (accessed on 26 October 2023).

The structure of this paper is as follows: in Section 2, we provide a summary of the most relevant works related to our research; Section 3 introduces the details of our proposed method; experimental results and comprehensive analyses are presented in Section 4; finally, our conclusions are presented in Section 5.

2. Related Work

2.1. Cross-Domain Facial Expression Recognition

Due to variations in subjective annotation and data collection methods, different facial expression recognition datasets inherently exhibit distribution disparities. The primary challenge is addressing the divergence in data distributions between the source and target domains. Zheng et al. [25] combined labeled samples from the source domain with unlabeled auxiliary data from the target domain to jointly learn discriminative subspace unsupervised. Similarly, [26] introduced the DR framework, which focuses on learning a domain regenerator capable of regenerating micro-expression samples from the source and target databases. This ensures that the generated source and target domains exhibit similar feature distribution. In [27], Li et al. observed different conditional probability distributions between the source and target domains and developed a deep Emotion-Conditioned Adaptation Network (ECAN) to address the data inconsistency. In [28,29], an attention mechanism was employed to achieve fine-grained feature alignment. Similarly, Tsai et al. [30] used discriminators to enforce similar semantic outputs between the two domains, emphasizing alignment and the acquisition of shared knowledge. In contrast, Zhang et al. [31] generated decision boundaries for each category by maximizing the divergence of classifiers and then trained a feature generator to deceive these two classifiers, thereby aligning the domain distribution between the existing source subject and the new subject. Compared with the existing methods described above, our proposed approach

introduces a global–local representation learning and dynamic label weighting method to tackle the domain shift problem in CD-FER.

2.2. Pseudo-Label Learning

Pseudo-label learning involves incorporating unlabeled data into model training to enhance the performance of supervised processes by utilizing model-predicted transformed hard labels [32–35]. Occasionally, certain samples may exhibit uncertainty regarding their categorization across multiple classes. In response to this challenge, Rizve et al. [36] introduced uncertainty into the framework and selected pseudo-labels based on uncertainty, thereby incorporating unlabeled data to enhance model robustness and performance. Zheng et al. [35] employed a progressive soft pseudo-label refinement mechanism, starting from coarse labels and refining them progressively. This approach aims to generate more robust and refined soft pseudo-labels, allowing the model to learn more discriminative features tailored for challenging samples. The "Noisy Student" approach [37] has been utilized when generating pseudo-labels to prevent the model from overfitting to known data as well as to enhance its generalization. Recently, the confidence-based threshold strategy has been considered to select appropriate pseudo-label data for the target domain. In [38], FixMatch was introduced; this approach utilizes high-confidence predictions from the weak augmentation branch to generate pseudo-labels, then trains the model by aligning predictions from the strong augmentation branch with the pseudo-labels using the standard cross-entropy loss. Unlike the previous work, Xie et al. [39] guided model training using the consistency loss in order to ensure similar data representation from different perspectives. Another recent study [40] applied adversarial training to enhance the confidence of pseudo-labels and improve the adversarial training process using these pseudo-labels. These two processes complement each other, strengthening the transfer effectiveness from the source domain to the target domain. However, reliance on a large amount of unlabeled data in generating pseudo-labels can lead to models classifying unlabeled data with excessive confidence, introducing erroneous pseudo-labels and misleading information. Furthermore, previous models have not considered the difference in class distributions between the source and target domains, a key factor influencing the ultimate transfer performance. In this work, we propose a novel approach that combines pseudo-complementary label generation with a dynamic label weight matching strategy to mitigate the impact of class imbalance.

3. Methodology

3.1. Overview

Our objective is to tackle the CD-FER task for a given source domain dataset $D_s = \{x_i^S, y_i^S\}_{i=1}^{N_S}$ and target domain dataset $D_t = \{x_i^t\}_{i=1}^{N_t}$. Each sample x_i^s is associated with a corresponding label y_i^s in the source domain, while the target domain dataset D_t consists of unlabeled samples. As illustrated in Figure 1, during the training phase, the feature extractor is responsible for extracting seven local and global feature vectors from each input image in the source domain. These feature vectors are then input to their respective classifiers for learning. The global and global–local feature vectors are extracted for the unlabeled target domain data. Subsequently, the Pseudo-Complementary Label Generation (PCLG) module is utilized to generate pseudo-labels and complementary labels. The PCLG module aims to provide label-like information for the unlabeled target data. Next, we employ the Label Dynamic Weight Matching (LDWM) module to assign appropriate weights to these generated labels. The LDWM module assesses the learning difficulty associated with each category and adapts label weights accordingly. Specifically, suppose that a category is predicted by the model with a lower frequency, indicating a higher level of learning difficulty. In this case, a higher label weight is assigned to encourage the model to prioritize learning that category. Conversely, the label weight is reduced if the category is predicted more frequently. Additionally, the LDWM module updates the number of pseudo-labels generated for each class. Finally, the source and target domains

are aligned using the loss function $\mathcal{L}$, facilitating knowledge transfer from the source to the target domain.

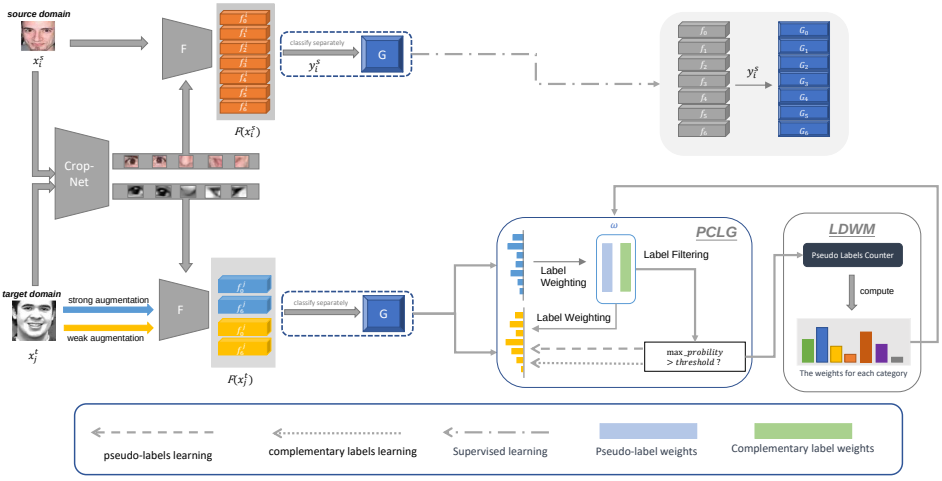

Figure 1. An overall depiction of our proposed method. Supervised learning is applied to train the labeled source data in the source domain, establishing a robust foundation for subsequent training on the target data. In the target domain, both strong and weak augmentations are performed to enhance the quality of the training data. The PCLG module generates corresponding pseudo-complementary label pairs and filters reliable pseudo-labels using a specified threshold. Simultaneously, the LDWM module calculates weights for each class and assigns the appropriate label weighting based on the model's prediction frequencies. This dynamic weighting scheme ensures that the model focuses on challenging categories and optimizes its learning process.

3.2. Pseudo-Complementary Label Generation

3.2.1. Pseudo-Label Generation

When generating pseudo-labels for the target domain, we consider samples with label confidence above a certain threshold as reliable samples. Setting an appropriate threshold is essential; thus, for our experiments with ResNet50 as the backbone network and RAF-DB as the source domain, we set the fixed threshold to 0.99. In the PCLG module (see Figure 1), samples were selected only if the class with the highest predicted probability scored above this threshold, with the rest being filtered out. During the second stage of training, we fine-tuned the model on both the source and target domains using the following pseudo-label training objective:

$$\mathcal{L}_{\text{pos}} = \frac{1}{N_t^{\text{trust}}} \sum_{i=1}^{N_t^{\text{trust}}} w_{y_i^t}^t \cdot \mathcal{L}_{CE}\left(G\left(f_i^t\right), y_i^t\right) \tag{1}$$

$$f_i^t = F\left(x_i^t + \xi_2\right) \tag{2}$$

$$y_i^t = G\left(F\left(x_i^t + \xi_1\right)\right) \tag{3}$$

where $w_{y_i^t}^t$ represents the value of the label re-weighting (explained in Section 3.3), f_i^t is the feature vector obtained using the weak augmentation strategy ξ_2, and y_i^t is the pseudo-label generated using ξ_1. It is important to note that pseudo-labels $D_t^{\text{trust}} = \{x_i\}_{i=1}^{N_t^{\text{trust}}}$, $x_i \in D_t$, $\max G(F(x_i)) >$ threshold are only generated for samples with confidence above a certain threshold. Here, F represents the feature extractor, ξ_1 represents the weak augmentation strategy, and ξ_2 represents the strong augmentation strategy.

3.2.2. Complementary Label Learning

In complementary label learning, unlabeled data are treated as both positive and negative. Positive instances are those labels the classifier correctly assigns, while negative instances are those that the classifier fails to assign correctly. Complementary labels are generated for negative instances to indicate which class the unlabeled data least belong to, aiming to compensate for the performance degradation caused by erroneous pseudo-labels. Complementary labels corresponding to low-confidence predictions are obtained as follows:

$$\bar{y}_i = \operatorname{argmin}\left(G\left(f_i^t\right)\right). \tag{4}$$

Introducing complementary label learning, the learning objective is as follows:

$$\mathcal{L}_{neg} = -\frac{1}{N_t}\sum_{i=1}^{N_t} w_{y_i^t}^t \cdot \left(\sum_{j=1}^{c} \bar{y}_{ij} \log\left(1 - G\left(f_i^t\right)_j\right)\right) \tag{5}$$

where $y_i^t or \bar{y}_i$ is a given pseudo-label or complementary label. Adaptive sample weights $w_{y_i^t}^t$ are assigned to the artificial labels, with $\bar{y}_{ij}$ denoting the j-th element of the corresponding one-hot vector.

3.3. Label Dynamic Weight Matching

We introduce a learnable class weighting parameter to explore the predicted class distribution in the target domain. Specifically, following the approach described in [41], the learning difficulty of a class is determined by the number of samples predicted to belong to that class and surpass the threshold. The formula for the learning difficulty of a class is as follows:

$$\lambda_j = \frac{\sigma_j}{\max_c \sigma_t} \tag{6}$$

where σ_j denotes the number of pseudo-labels generated by classifier G for a particular class j, c is the total number of classes, and $\max_c \sigma_t$ is the maximum number of pseudo-labels generated by G across all classes. The values of λ_j range from 0 to 1, with the best-learned class having λ_j equal to 1. To introduce a warm-up process, we modify the denominator in the equation as follows:

$$\lambda_j = \frac{\sigma_j}{\max\left(\max_c \sigma_t, N - \Sigma_{c=1}^{C}\sigma_t\right)} \tag{7}$$

where $N - \sum_{c=1}^{C} \sigma_t$ can be considered the number of unlabeled samples not used in the target domain. This ensures that at the beginning of training, all estimated learning effects start from 0 and gradually increase until the unlabeled target domain samples no longer dominate. The goal is to increase the weight of difficult samples while decreasing the weight of easy samples. To achieve this, we assign adaptive sample weights to pseudo-labels as follows:

$$w_j^t = 1 - \frac{\lambda_j^2}{\tau} \tag{8}$$

where w_j^t represents the adaptive weight for class j in the target domain. In the later stages of training, the best-learned class has a weight of $1 - 1/\tau$, while τ is a hyperparameter that determines the strength of the weight-matching mechanism. As training progresses, the weights for those classes with better learning effects decrease; however, it is important to note that the weights may not always decrease. Suppose that unlabeled target domain data are classified into different classes in subsequent iterations; in this case, a deterioration in the learning effect is indicated for that class, increasing its weight.

3.4. Loss Function

In the first stage of training, we trained seven classifiers by minimizing the cross-entropy between the predicted labels $G\left(f_i^s\right)$ and the true labels. This strategy established an initial condition for learning the target domain task. The loss function for this stage is defined as

$$\mathcal{L}_{\text{sup}} = \frac{1}{N_S} \sum_{i=1}^{N_s} w^s \cdot \mathcal{L}_{CE}\left(G\left(f_i^S\right), y_i^S\right) \tag{9}$$

where w is the weight array assigned to the seven sets of features. We set the weights for global and global–local features to 7, while the weights for the five sets of local features were set to 1. The overall training objective combines the supervised loss L_{sup} from the first stage with the positive learning objective L_{pos} and negative learning objective L_{neg} from the second stage:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}} \tag{10}$$

The above loss function guides the training process, effectively ensuring that the model learns from the source and target domain. For a clearer understanding of the main flow of the proposed framework, please refer to the pseudocode provided in Algorithm 1.

Algorithm 1 Global–Local Representation Learning and Dynamic Label Weighting framework

Input: X^s: source domain dataset; X^t: target domain dataset; C: total number of categories τ: fixed threshold; w_j^t: the label weight for category j; σ_j: the number of generated pseudo-labels for category j; ζ_1: weak augmentation strategy;

while not reach the maximun iteration **do**
  for $c = 1$ to C **do**
    Calculate λ_j using Equation (7)
    Calculate w_j^t using Equation (8)
  end for
  Sample mini-batch of size B from $X^s : X_B^s$
  Sample mini-batch of size B from $X^t : X_B^t$
  for $b = 1$ to B **do**
    if $\max\left(G\left(F\left(x_i^t + \zeta_1\right)\right)\right) > \tau$ **then**
      Pseudo labels $y_i^t \leftarrow$ Equation (3) on x_i^t
      Complementary labels $\leftarrow$ Equation (4) on x_i^t
      Update σ_j;
    end if
  end for
  Compute the loss via Equations (9), (1), (5) and (10)
end while

Output: Model parameters.

3.5. Evaluation Metrics

In evaluating the performance of our model for CD-FER, we employed a set of essential metrics to gauge its effectiveness in classifying facial expressions across different domains.

Accuracy is a fundamental classification model evaluation metric that quantifies a model's correctness in predicting labels. It represents the proportion of samples correctly classified by the model, typically expressed as a percentage. This metric is suitable for datasets with a relatively balanced distribution of classes.

Recall is a crucial metric used to assess a model's ability to correctly identify positive samples, especially in applications where it is essential to ensure that positive instances are not missed. The recall calculates the proportion of true positive samples among all actual

positive samples, and is particularly valuable in scenarios where false negatives are costly. The recall formula is as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

where TP represents the number of correctly predicted positive samples and FN represents the number of incorrectly predicted negative samples.

Precision is another critical metric for evaluating classification performance. It measures the proportion of true positive samples among those predicted as positive by the model. The precision formula is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

where TP represents the number of correctly predicted positive samples and FP represents the number of incorrectly predicted positive samples.

F1 score is a comprehensive performance metric for classification models. It provides a balanced assessment of a model's precision and recall, making it particularly useful in cases involving class imbalance. The F1 Score is calculated as follows:

$$\text{Fl score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{13}$$

3.6. Implementation Details

3.6.1. Network Architecture

In our implementation, we built upon the work of Chen et al. [22] and adopted the ResNet50 variant and MobileNet-v2 as the backbone networks for feature extraction. Both of these networks consist of four block layers. Starting with an input image size of 112×112, a feature map is obtained from the second layer of the network, which is denoted as m_1 and has dimensions of $28 \times 28 \times 128$. Similarly, another feature map is acquired from the fourth layer, which is denoted m_2 and has a size of $7 \times 7 \times 512$. A convolutional operation is initiated to transform m_2 into a $7 \times 7 \times 64$ feature map to capture the global features. Subsequently, an average pooling layer is applied to process the feature map, resulting in a 64-dimensional feature vector. We leveraged the MT-CNN to crop five regions of size $7 \times 7 \times 128$ centered around the corresponding facial landmarks in m_1 to extract local features. These regions undergo similar convolutional and average pooling operations, yielding five 64-dimensional local vectors. To integrate global and local information, the global feature vector is concatenated with the five local feature vectors, producing a combined 384-dimensional feature vector. The feature vector is then formed into a set, denoted as f, comprising seven feature vectors. We constructed seven classifiers using fully-connected layers for classification.

3.6.2. Training Details

In the two-stage training process, we initialized our backbone network with models pretrained on the MS-Celeb-1M dataset [42]. The parameters of the additional layers were initialized using the Xavier algorithm. The training specifics for each stage were as follows: first, we used the Stochastic Gradient Descent (SGD) optimizer to train the feature extractor and classifier on the source domain. We conducted training for fifteen epochs, for which we employed the cross-entropy loss. The learning rate was set to 0.0001, the momentum to 0.9, and the weight decay to 0.0005. In the second stage, we fine-tuned the feature extractor and classifier by optimizing the target loss function in Equation (10). We maintained the same momentum and weight decay values as in the first stage; however, we adjusted the learning rate to 0.00001. After thirty epochs, the learning rate was halved to facilitate convergence and improve training stability.

4. Experiments and Analyses

4.1. Datasets

In our experiments, we utilized several diverse datasets for CD-FER. Each dataset presents unique challenges and characteristics, allowing us to evaluate the performance of our proposed model comprehensively. Below, we provide an overview of the datasets used:

RAF-DB [7] comprises 29,672 images of human faces featuring multiple human species. The dataset is gender-balanced, has a large age span, and has various poses. In all, 15,339 images are labeled with seven basic expressions, divided into 12,271 training samples and 3068 test samples for evaluation.

FER2013 [8] is a large uncontrolled dataset collected using the Google image search engine. It contains 35,887 images, including 4953 angry samples, 547 disgusted samples, 5121 worried samples, 8989 happy samples, 6077 sad samples, 4002 surprised samples, and 6198 neutral samples. Among these images, 28,709 samples were used for training, 3589 for verification, and 3589 for testing.

CK+ [9] is a lab-controlled dataset often used to measure the performance of algorithms on facial expression recognition tasks. It comprises 593 video data sequences from 123 subjects, including 309 sequences labeled using the Facial Action Coding System (FACS). We constructed a dataset of 1236 images using the same method for extracting expression data.

JAFFE [10] is a lab-controlled facial expression dataset featuring 213 images collected from ten Japanese women. Due to its Asian origin, this dataset is particularly suitable for evaluating cross-domain facial expression recognition tasks.

SFEW2.0 [11] is a static in-the-wild expression dataset encompassing unconstrained facial expressions. It includes various head poses, a wide age range, occlusions, varied focus, and different face resolutions. The dataset is divided into training, validation, and test datasets that contain 958, 436, and 372 samples, respectively.

ExpW [12] is another in-the-wild dataset, featuring 91,793 face images collected from the Google Image Search API. Each image is manually labeled with one of the seven basic emotion categories.

The class quantity distribution across these datasets is illustrated in Figure 2. These diverse datasets enable us to assess our proposed CD-FER model's robustness and generalization capabilities across multiple domains and real-world scenarios.

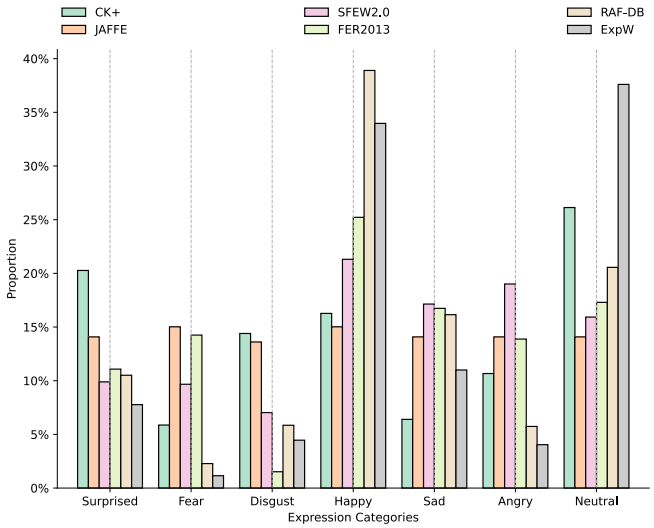

Figure 2. Bar plot illustrating the category distribution in the training set across six datasets: CK+, JAFFE, SFEW2.0, FER2013, RAF-DB, and ExpW.

4.2. Comparisons with State-of-the-Art Methods

Our experimental evaluation compared our proposed model with state-of-the-art CD-FER methods. We chose ResNet50 and MobileNet-v2 as our backbone network, each having distinct hierarchical structures and feature extraction capabilities. This choice was taken for two main reasons: first, these networks can leverage pretrained weights from large-scale datasets, accelerating convergence and enhancing performance; second, we aimed to assess the robustness of our algorithm across different backbones. Taking data from Table 1(i) as an example for further data analysis, when using RAF-DB as the source domain and ResNet50 as the backbone network, our algorithm demonstrates significant performance improvements on the Ck+, JAFFE, SFEW2.0, FER2013, and ExpW datasets, with respective accuracy increases of 3.1%, 7.04%, 0%, 1.8%, and 2.23% compared to the state-of-the-art algorithms. These results indicate that our method is positioned competitively compared to existing approaches. However, when replacing the source domain with FER2013 while keeping ResNet50 as the backbone, the performance of almost all algorithms exhibited varying degrees of decline on all datasets. This underscores that the similarity between the source and target domains has an important effect on accuracy. It can be seen from Table 1 that when we maintained the same source domain and replaced ResNet50 with MobileNet-v2 as the backbone, the performance of most algorithms showed varying degrees of decline on all datasets. This could be attributed to MobileNet-v2's adoption of deeply separable convolution in its architecture, resulting in relatively weaker feature extraction capabilities than ResNet50.

Table 1. Accuracy of our proposed method (Ours) and the existing leading methods on CK+, JAFFE, SFEW2.0, FER2013, RAF-DB, and ExpW when using different source datasets and different backbone networks. The best results are highlighted in bold.

Method	**(i) Source=RAF-DB, Backbone=ResNet50**					
	CK+	JAFEE	SFEW2.0	FER2013	ExpW	Mean
CADA [43]	72.09	52.11	53.44	57.61	63.15	59.68
SAFN [44]	75.97	61.03	52.98	55.64	64.91	62.11
SWD [45]	75.19	54.93	52.06	55.84	68.35	61.27
ECAN [27]	79.77	57.28	52.29	56.46	47.37	58.63
AGRA [22]	85.27	61.50	56.43	58.95	68.50	66.13
CGLRL [24]	82.95	59.62	56.88	59.30	70.02	65.75
Ours	**88.37**	**68.54**	**56.88**	**61.10**	**73.25**	**69.63**
Method	**(ii) Source=FER2013, Backbone=ResNet50**					
	CK+	JAFEE	SFEW2.0	RAF-DB	ExpW	Mean
CADA [43]	81.40	45.07	46.33	65.96	54.84	58.72
SAFN [44]	68.99	45.07	38.07	62.80	53.91	53.77
SWD [45]	65.89	49.30	45.64	65.28	56.05	56.43
ECAN [27]	60.47	41.76	46.01	53.41	48.88	50.11
AGRA [22]	**85.69**	52.74	49.31	67.62	60.23	63.12
CGLRL [24]	79.84	53.52	**52.29**	71.84	**61.94**	63.87
Ours	82.95	**60.09**	49.08	**75.68**	54.68	**64.50**
Method	**(iii) Source=RAF-DB, Backbone=MobileNet-v2**					
	CK+	JAFEE	SFEW2.0	FER2013	ExpW	Mean
CADA [43]	62.79	53.05	43.12	49.34	59.40	53.54
SAFN [44]	66.67	45.07	40.14	49.90	61.40	52.64
SWD [45]	68.22	55.40	43.58	50.30	60.04	55.51
ECAN [27]	53.49	43.08	35.09	45.77	45.09	44.50
AGRA [22]	72.87	**55.40**	45.64	51.05	63.94	57.78
CGLRL [24]	69.77	52.58	**49.77**	52.46	64.87	57.89
Ours	**75.97**	49.77	47.25	**52.97**	**64.89**	**58.17**

Table 1. *Cont.*

Method	(iv) Source=FER2013, Backbone=MobileNet-v2					
	CK+	JAFEE	SFEW2.0	RAF-DB	ExpW	Mean
CADA [43]	66.67	**50.23**	41.28	53.15	51.84	52.63
SAFN [44]	66.67	37.56	35.78	38.73	45.56	44.86
SWD [45]	53.49	48.36	35.78	47.44	50.02	47.02
ECAN [27]	55.65	44.12	28.46	42.31	41.53	42.41
AGRA [22]	67.44	47.89	41.74	52.27	**59.41**	53.75
CGLRL [24]	68.22	46.95	46.79	59.15	54.30	55.08
Ours	**76.74**	47.72	**46.79**	**64.79**	54.70	**58.13**

4.3. Ablation Studies

Next, we conducted ablation studies to assess the individual contributions of different components within our proposed algorithm framework. Our framework comprises several components: Pseudo-Label Generation (PLG), Label Dynamic Weight Matching (LDWM), and Complementary Label Generation (CLG). To evaluate the performance contributions of each component to the CD-FER task, we conducted experiments using a variant of ResNet50 as the backbone network, RAF-DB as the source domain, and the CK+, JAFFE, SFEW2.0, FER2013, and ExpW datasets as the target domains.

We established a baseline where only labeled source domain data was used for separate classification learning. The results are shown in the baseline row in Table 2. Then, we introduced the CLG and LDWM modules separately, building upon the PLG component. The final row represents the accuracy of the complete framework implementation. By comparing the second and third rows, it is evident that both CLG and LDWM contribute to the performance improvement of CD-FER. In particular, a noticeable decline in performance was observed across all datasets when LDWM was removed, resulting in a drop in average accuracy of 0.82.

Table 2. Ablation study for each module.The best results are highlighted in bold.

Module	**CK+**	**JAFFE**	**SFEW2.0**	**FER2013**	**ExpW**	**Mean**
Baseline	73.64	59.15	52.29	56.88	68.93	62.18
PLG + CLG	88.37	66.67	55.73	60.56	72.72	68.81
PLG + LDWM	88.37	67.13	56.65	60.93	73.04	69.22
Ours	**88.37**	**68.54**	**56.88**	**61.10**	**73.25**	**69.63**

Recognizing the presence of class imbalance, we introduced the F1 score as a performance metric to account for this imbalance. Figure 3 shows the F1 scores on various datasets, demonstrating the robustness and effectiveness of our model across different domains.

Additionally, we assessed the contribution of local features by conducting an experiment in which only global features were used for training. The results in Table 3 highlight the importance of incorporating local information in enhancing cross-domain recognition performance.

Table 3. Accuracy of our approach using holistic features (HFs) and ours for adaptation on the CK+, JAFFE, SFEW2.0, FER2013, and ExpW datasets. The best results are shown in bold.

Method	**CK+**	**JAFFE**	**SFEW2.0**	**FER2013**	**ExpW**	**Mean**
Ours HFs	88.37	68.08	55.96	60.08	72.65	69.03
Ours	**88.37**	**68.54**	**56.88**	**61.10**	**73.25**	**69.63**

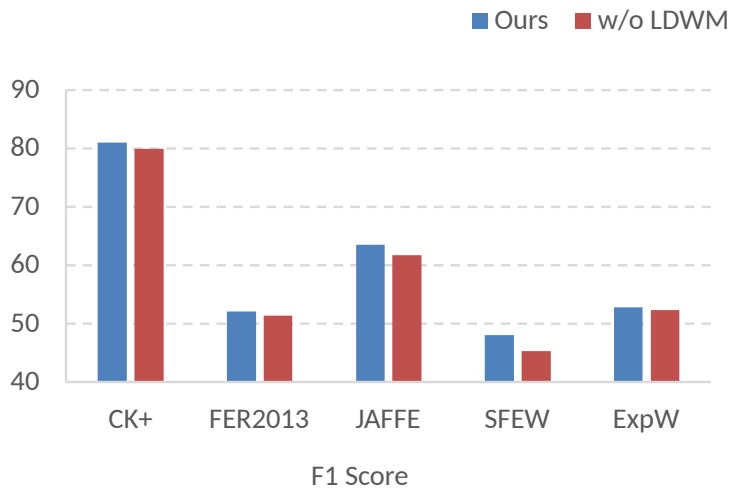

Figure 3. F1 scores on various datasets.

4.4. Parameter Analysis

The hyperparameter τ (temperature) plays a crucial role in the feature learning process for the target domain. Initially, we set τ to 1.0 as the initial value, equivalent to not introducing any hyperparameter. Subsequently, we increased τ in increments of 0.1 while observing the impact on the experimental results for each dataset. Notably, we observed that these fluctuations in the hyperparameter significantly affected the CK+ and FER2013 datasets. Hence, we selected these two datasets for experimentation to determine the optimal range of the hyperparameter τ. The temperature parameter τ serves as a balancing factor between learning the characteristics of challenging samples and preserving the features that have already been learned. Specifically, CK+ exhibited a decline in performance when τ dropped below a certain threshold, while FER2013 performed better with τ values closer to 1.0. We conducted experiments with different τ values on the CK+ and FER2013 datasets to investigate this phenomenon, as illustrated in Figure 4. We found that when τ exceeded or equaled 1.5, the CK+ dataset achieved its best performance. In contrast, the performance of FER2013 fluctuated while remaining at a relatively high level with τ values close to 1.0.

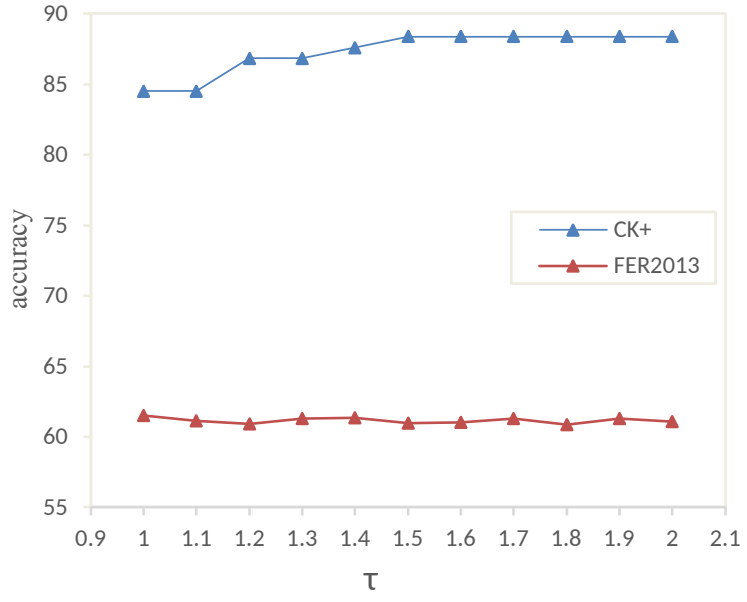

Figure 4. Performance with respect to τ on CK+ and FER2013.

5. Conclusions

In this paper, we have introduced a robust global–local feature learning approach coupled with dynamic label weighting to enhance the performance of unsupervised cross-domain facial expression recognition. Our method comprises the PCLG module, which generates high-quality labels for learning domain-invariant features, and the LDWM module, which adjusts label weights to mitigate class distribution disparities between the source and target domains. We have demonstrated the efficacy of these proposed components through extensive experiments and ablation studies, consistently achieving superior performance on publicly available datasets compared to state-of-the-art methods.

While this work represents a significant advancement in cross-domain facial expression recognition, it may not show the best performance in certain scenarios. For example, when changing the backbone network in conjunction with the source domain, the performance of the proposed model is not as robust as previously observed. Moreover, introducing hyperparameters necessitates additional computational resources in order to explore suitable ranges for these hyperparameters. Future research should explore more sophisticated data augmentation techniques, robust label prediction methods, and appropriate weight calculation formulas without introducing hyperparameters.

Author Contributions: Methodology, Y.G. and X.B.; Software, Y.G., Y.C. and X.B.; Validation, B.L. and W.Z.; Investigation, S.L.; Resources, W.Z.; Writing—original draft, Y.C. and B.L.; Writing—review & editing, Y.G. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

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

References

1. Lu, C.; Zong, Y.; Zheng, W.; Li, Y.; Tang, C.; Schuller, B.W. Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2217–2230. [CrossRef]
2. Zhang, T.; Zong, Y.; Zheng, W.; Chen, C.L.P.; Hong, X.; Tang, C.; Cui, Z.; Zhao, G. Cross-database micro-expression recognition: A benchmark. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 544–559. [CrossRef]
3. Zhang, S.; Zhang, Y.; Zhang, Y.; Wang, Y.; Song, Z. A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition. *Electronics* **2023**, *12*, 3595. [CrossRef]
4. Yan, L.; Shi, Y.; Wei, M.; Wu, Y. Multi-feature fusing local directional ternary pattern for facial expressions signal recognition based on video communication system. *Alex. Eng. J.* **2023**, *63*, 307–320. [CrossRef]
5. Li, S.; Deng, W.H. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
6. Sun, Z.; Zhong, H.H.; Bai, J.T.; Liu, M.Y.; Hu, Z.P. A discriminatively deep fusion approach with improved conditional gan (im-cgan) for facial expression recognition. *Pattern Recognit.* **2023**, *135*, 109157. [CrossRef]
7. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2852–2861.
8. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013; pp. 117–124.
9. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
10. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
11. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011; pp. 2106–2112.
12. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From facial expression recognition to interpersonal relation prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569. [CrossRef]
13. Yan, H. Transfer subspace learning for cross-dataset facial expression recognition. *Neurocomputing* **2016**, *208*, 165–173. [CrossRef]

14. Miao, Y.Q.; Araujo, R.; Kamel, M.S. Cross-domain facial expression recognition using supervised kernel mean matching. In Proceedings of the International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012; Volume 2, pp. 326–332.
15. Sun, Z.; Chiong, R.; Hu, Z.P.; Dhakal, S. A dynamic constraint representation approach based on cross-domain dictionary learning for expression recognition. *J. Vis. Commun. Image Represent.* **2022**, *85*, 103458. [CrossRef]
16. Ni, T.; Zhang, C.; Gu, X. Transfer model collaborating metric learning and dictionary learning for cross-domain facial expression recognition. *IEEE Trans. Comput. Soc. Syst.* **2020**, *8*, 1213–1222. [CrossRef]
17. Wang, C.; Ding, J.; Yan, H.; Shen, S. A Prototype-Oriented Contrastive Adaption Network for Cross-Domain Facial Expression Recognition. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022; pp. 4194–4210.
18. Bozorgtabar, B.; Mahapatra, D.; Thiran, J.P. ExprADA: Adversarial domain adaptation for facial expression analysis. *Pattern Recognit.* **2020**, *100*, 107111. [CrossRef]
19. Liang, G.; Wang, S.; Wang, C. Pose-aware adversarial domain adaptation for personalized facial expression recognition. *arXiv* **2020**, arXiv:2007.05932.
20. Yang, H.; Zhang, Z.; Yin, L. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 15–19 May 2018; pp. 294–301.
21. Xie, Y.; Chen, T.; Pu, T.; Wu, H.; Lin, L. Adversarial graph representation adaptation for cross-domain facial expression recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1255–1264.
22. Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; Lin, L. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9887–9903. [CrossRef]
23. Li, Y.; Gao, Y.; Chen, B.; Zhang, Z.; Zhu, L.; Lu, G. JDMAN: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition. In Proceedings of the ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 3312–3320.
24. Xie, Y.; Gao, Y.; Lin, J.; Chen, T. Learning Consistent Global-Local Representation for Cross-Domain Facial Expression Recognition. In Proceedings of the International Conference on Pattern Recognition, Montreal, QC, Canada, 21–25 August 2022; pp. 2489–2495.
25. Zheng, W.; Zong, Y.; Zhou, X.; Xin, M. Cross-domain color facial expression recognition using transductive transfer subspace learning. *IEEE Trans. Affect. Comput.* **2016**, *9*, 21–37. [CrossRef]
26. Zong, Y.; Zheng, W.; Huang, X.; Shi, J.; Cui, Z.; Zhao, G. Domain regeneration for cross-database micro-expression recognition. *IEEE Trans. Image Process.* **2018**, *27*, 2484–2498. [CrossRef]
27. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [CrossRef]
28. Lu, S.; Liu, M.; Yin, L.; Yin, Z.; Liu, X.; Zheng, W. The multi-modal fusion in visual question answering: A review of attention mechanisms. *PeerJ Comput. Sci.* **2023**. [CrossRef]
29. Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; Yang, Y. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2507–2516.
30. Tsai, Y.H.; Sohn, K.; Schulter, S.; Chandraker, M. Domain adaptation for structured output via discriminative patch representations. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1456–1465.
31. Zhang, X.; Huang, D.; Li, H.; Zhang, Y.; Xia, Y.; Liu, J. Self-training maximum classifier discrepancy for EEG emotion recognition. *CAAI Trans. Intell. Technol.* **2023** [CrossRef]
32. Gao, D.; Wang, H.; Guo, X.; Wang, L.; Gui, G.; Wang, W.; Yin, Z.; Wang, S.; Liu, Y.; He, T. Federated Learning Based on CTC for Heterogeneous Internet of Things. *IEEE Internet Things J.* **2023** [CrossRef]
33. Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; Zhao, J. PiCO: Contrastive Label Disambiguation for Partial Label Learning. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021; pp. 1–18.
34. Zheng, D.; Xiao, J.; Chen, K.; Huang, X.; Chen, L.; Zhao, Y. Soft Pseudo-Label Shrinkage for Unsupervised Domain Adaptive Person Re-identification. *Pattern Recognit.* **2022**, *127*, 108615. [CrossRef]
35. Wang, J.; Zhang, X.L. Improving pseudo labels with intra-class similarity for unsupervised domain adaptation. *Pattern Recognit.* **2023**, *138*, 109379. [CrossRef]
36. Rizve, M.; Duarte, K.; Rawat, Y.; 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, Virtual, 3–7 May 2021; pp. 1–20.
37. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-Training With Noisy Student Improves ImageNet Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10684–10695.
38. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; pp. 596–608.
39. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised Data Augmentation for Consistency Training. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; pp. 6256–6268.

40. Yi, C.; Chen, H.; Xu, Y.; Liu, Y.; Jiang, L.; Tan, H. ATPL: Mutually enhanced adversarial training and pseudo labeling for unsupervised domain adaptation. *Knowl. Based Syst.* **2022**, *250*, 108831. [CrossRef]
41. Zhang, B.; Wang, Y.; Hou, W.; WU, H.; Wang, J.; Okumura, M.; Shinozaki, T. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; pp. 18408–18419.
42. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 87–102.
43. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional adversarial domain adaptation. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 1647–1657.
44. Xu, R.; Li, G.; Yang, J.; Lin, L. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1426–1435.
45. Lee, C.Y.; Batra, T.; Baig, M.H.; Ulbricht, D. Sliced wasserstein discrepancy for unsupervised domain adaptation. In Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10285–10295.