Supplementary Materials for Learn-to-Decompose: Cascaded Decomposition Network for Cross-Domain Few-Shot Facial Expression Recognition

Xinyi Zou[1], Yan Yan[1⋆], Jing-Hao Xue[2], Si Chen[3], and Hanzi Wang[1]

[1] Xiamen University, China
[2] University College London, UK
[3] Xiamen University of Technology, China

1 Detailed Architecture of the LD module

The proposed CDNet cascades several learn-to-decompose (LD) modules with shared parameters to obtain weighted expression prototypes progressively. We illustrate the detailed architecture of the proposed LD module in Fig. 1. Each LD module contains a decomposition block and a weighting block to obtain an expression prototype and its corresponding weight, respectively. The decomposition block includes a transformation layer, a batch normalization layer, and an activation layer. The weighting block consists of a three-layer perceptron network. The dropout layer is used for regularization.

2 Datasets

Our CDNet is trained on multiple *basic* expression datasets and tested on the *compound* expression dataset. Such a cross-domain setting evaluates the generalization ability of our model (trained on multiple source domains) on a novel compound FER task in the target domain.

Basic Expression Datasets We use five basic expression datasets to form the training set, including three in-the-lab datasets (CK+, MMI, and Oulu-CASIA), and two in-the-wild datasets (RAF-DB and SFEW). **CK+** involves 327 video sequences annotated with seven basic expression categories. **MMI** contains 326 video sequences (205 frontal-view sequences are used) with six basic expression categories. **Oulu-CASIA** consists of 2,880 video sequences (480 normal indoor illumination sequences are used) with six basic expression categories. Three peak frames of each sequence in the above in-the-lab datasets are selected in our experiments. **RAF-DB** includes a basic subset and a compound subset. The basic subsets with seven basic expression annotations are employed as a part of the training set. **SFEW** is also labeled with seven basic expression categories. All the samples in these basic expression datasets are used for training.

[⋆] Corresponding author. 🖂 : yanyan@xmu.edu.cn

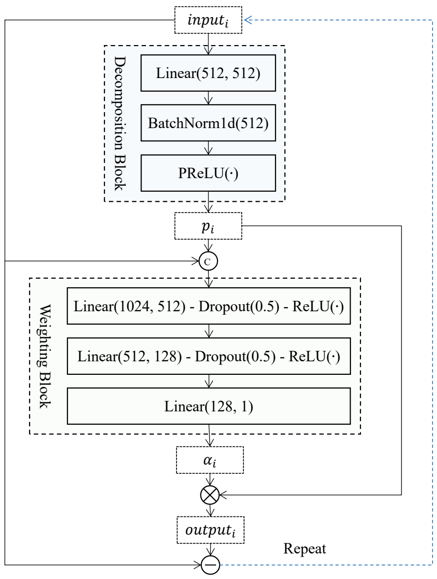


Fig. 1. Detailed architecture of the proposed LD module. Each LD module contains a decomposition block and a weighting block. CDNet cascades LD modules with shared parameters to obtain weighted expression prototypes progressively.

Compound Expression Datasets Three compound expression datasets are used to evaluate the performance of the learned model. To ensure the disjointness of base classes and novel classes, only the compound expression subsets of the compound expression datasets are used for testing, denoted as CFEE_C, EmotioNet_C, and RAF_C. **CFEE_C** is derived from the CFEE dataset. It is an in-the-lab dataset and annotated with 15 compound expressions for 230 subjects. **EmotioNet_C** comes from the EmotioNet challenge. The samples are collected in-the-wild and annotated with ten compound expression categories. **RAF_C** is the compound subset of RAF-DB with 11 compound expression categories. All the samples in these compound expression subsets are used for testing.

3 Influence of Different Balance Weights

We evaluate the influence of balance weights in the fine-tuning stage. First, we fix λ_r^f as 1.0, and set λ_d^f from 0.0 to 1.0. The results are given in Table 1(a). We can see that, without considering the domain classification loss, our CDNet achieves the worst recognition accuracy in most cases. This is because of the domain gap between the training set and the testing set. Meanwhile, a large value of λ_d^f will also influence the result (see the first row in Table 1(a)). Our CDNet obtains the best performance when λ_d^f is set to 0.01.

We also evaluate the results with the different values of λ_r^f. The results are shown in Table 1(b), where λ_d^f is fix as 0.01 and λ_r^f is varied from 0.0 to 2.0. We

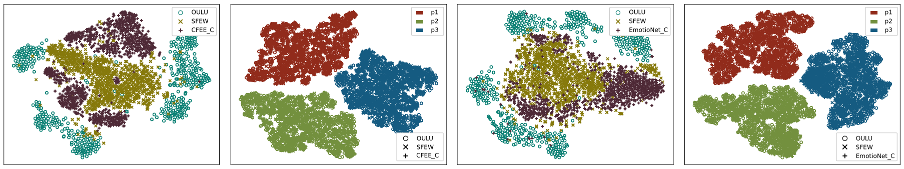
Table 1. Influence of different balance weights. The average accuracy (%) of 5-way few-shot classification tasks is used for evaluation with different values of λ_d^f and λ_r^f.

(a) Influence of different λ_d^f

λ_d^f	CFEE_C		EmotioNet_C		RAF_C	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
1	55.47	68.42	54.56	62.74	45.75	62.52
0.1	56.77	68.57	54.73	62.81	45.83	62.75
0.01	**56.99**	**68.98**	**55.16**	**63.03**	**46.07**	**63.03**
0.001	56.34	68.66	54.84	62.7	45.96	62.60
0	56.06	67.99	54.38	62.27	44.73	61.10

(b) Influence of different λ_r^f

λ_r^f	CFEE_C		EmotioNet_C		RAF_C	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
2.0	56.94	68.42	54.45	62.71	45.76	62.67
1.5	56.65	68.51	54.88	62.85	46.01	62.82
1.0	**56.99**	**68.98**	**55.16**	**63.03**	**46.07**	**63.03**
0.5	55.87	68.92	54.66	62.17	45.49	62.22
0.0	55.17	67.85	53.01	61.60	43.78	61.00



(a) Feature (b) Prototype (c) Feature (d) Prototype

Fig. 2. Visualization results with and without the decomposition design. (a), (c) denote the learned holistic features from the baseline method, where different colors denote different domains. (b), (d) denote the learned expression prototypes from our cascaded decomposition, where different colors and markers denote different prototypes and domains, respectively.

can see that the regularization plays a critical role in the final performance. A large or a small value of λ_r^f will both affect the recognition accuracy. Our model achieves the best performance when λ_r^f is set to 1.0.

4 Visualization of Decomposition

To validate the importance of the decomposition design, we visualize the holistic feature obtained by the baseline method and the learned prototypes obtained by our cascaded decomposition in Fig. 2. The results given by two source domains (an in-the-lab dataset (OULU) and an in-the-wild dataset (SFEW)) and two target domains (an in-the-lab dataset (CFEE_C) and an in-the-wild dataset (EmotioNet_C)) are shown. Note that the expression categories are disjoint between the source and target domains.

The holistic features extracted by images from the source and target domains are significantly different (see Fig. 2(a) and Fig. 2(c)). In contrast, the learned prototypes from these domains are indistinguishable (e.g., the first learned prototypes (marked in red) from different domains (marked in different markers) are closely distributed in Fig. 2(b) and Fig. 2(d)). Therefore, the learned prototypes are generic to different expression categories and domains, and are of great significant to reconstruct a transferable feature space that can help recognize novel compound expression categories in the cross-domain FSL setting.