Learning Deep Hierarchical Features with Spatial Regularization for One-Class Facial Expression Recognition

Bingjun Luo[1], Junjie Zhu[1], Tianyu Yang[1], Sicheng Zhao[2], Chao Hu[3]*, Xibin Zhao[1], Yue Gao[1]

[1]BNRist, KLISS, School of Software, Tsinghua University
[2]BNRist, Tsinghua University
[3]Central South University
luobingjun@gmail.com, zhujj18@mails.tsinghua.edu.cn, goatskyfish@gmail.com, schzhao@gmail.com, huchao@csu.edu.cn, {zxb, gaoyue}@tsinghua.edu.cn

Abstract

Existing methods on facial expression recognition (FER) are mainly trained in the setting when multi-class data is available. However, to detect the alien expressions that are absent during training, this type of methods cannot work. To address this problem, we develop a Hierarchical Spatial One Class Facial Expression Recognition Network (HS-OCFER) which can construct the decision boundary of a given expression class (called normal class) by training on only one-class data. Specifically, HS-OCFER consists of three novel components. First, hierarchical bottleneck modules are proposed to enrich the representation power of the model and extract detailed feature hierarchy from different levels. Second, multi-scale spatial regularization with facial geometric information is employed to guide the feature extraction towards emotional facial representations and prevent the model from overfitting extraneous disturbing factors. Third, compact intra-class variation is adopted to separate the normal class from alien classes in the decision space. Extensive evaluations on 4 typical FER datasets from both laboratory and wild scenarios show that our method consistently outperforms state-of-the-art One-Class Classification (OCC) approaches.

Introduction

As more and more intelligent devices step into our daily life, ubiquitous computing environments are gradually coming into reality (Xi et al. 2022a). However, most existing Human-Computer Interaction (HCI) works are designed to understand and handle the explicit instructions of users, while ignoring their internal psychological and emotional states (Zeng et al. 2009). Such kinds of interactions lack emotional intelligence and present great challenges for building user-friendly HCI systems. This problem prompts researchers to turn their attention to a powerful emotional indicator, facial expression (FE), which makes it possible for HCI to uncover the subtleties of the users' affective behavior and deal with their emotional changes. With the help of facial expression recognition (FER), HCI systems can gain the ability to provide more warm and humanized service (Wen et al. 2016; Xi et al. 2022b).

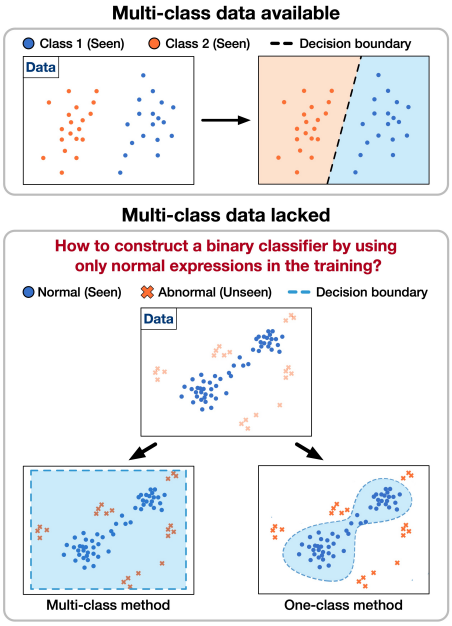

Figure 1: When multi-class expression data is available during training, ordinary FER methods can operate well. But in real applications, multi-class data is not always available. Under such a case, the FER system cannot be trained in the ordinary fashion. Therefore, one-class classification methods are needed to draw a precise decision boundary of the normal expression class so that the alien expression beyond the boundary can be easily detected during inference.

Generally, the objective of FER is to extract and interpret subtle FE-related representation from the input image (Zhang and Tjondronegoro 2011). As shown in Fig. 1, FER is often modeled as a multi-class classification task with adequate multi-class training data in most of the recent works. However, in some real application scenarios with a lack of multi-class data, these FER methods do not always work. This situation is quite common in the open world due to the great costs and difficulties of the collection and annotation of FE data. For example, panic disorder is an anxiety disorder characterized by reoccurring unexpected intense fear. Psy-

*Corresponding author.

chologists hope to monitor the course of this disease by automatically detecting the patient's fearful expression during the attack. Since the attacks of panic disorder are unexpected and short in time, it is almost impossible to collect the fearful expression from the patient during the model training. Thus, the lack of fear data will directly lead ordinary multi-class FER methods to not be trained normally.

The above-mentioned real scenarios require an FER system that can be trained from the normal expression which is easy to observe, and then detect the alien expression which is hard to collect. This leads to a new task of One-Class Classification (OCC) for FER. OCC is a machine learning paradigm to detect the alien class that falls outside of the training data. Though showing good performance in many other vision tasks, existing OCC approaches do not seem to fit well in the FER task. The main reason is that current OCC methods tend to concentrate on the high-level semantic extraction of the input image and neglect the hierarchical spatial information in subtle facial representations, which greatly limits their representation power and leads to the underfitting problem in the complicated FER task.

In this paper, we study one class facial expression recognition. Specifically, we design a novel method, termed HS-OCFER, to detect the alien expression with the help of hierarchical and spatial facial information. First, we construct the hierarchical bottleneck modules to enhance the representation ability of the auto-encoder backbone and extract rich latent features from various levels of the network. Comprised of low-level facial texture, middle-level muscle activation, and high-level semantic information, these features provide critical information for the FER task. Second, we propose to employ multi-scale spatial regularization by performing facial landmark detection on the extracted representation. This regularization term can be viewed as a constraint to guide the network towards emotional information in subtle facial representations. Third, we adopt the intra-class variation compacting in the decision space, which minimizes the volume of normal expression's hypersphere. To accommodate different tasks and features, we summarize the hybrid loss function and propose a decision-level fusion strategy for inference.

In summary, the contributions of this paper are threefold:

(1) We construct a novel method HS-OCFER to detect alien expression that falls out of the training expression class. It is innovative to propose the deep OCC method on FER, as far as we know.

(2) To balance the network's representation power appropriately between underfitting and overfitting, we propose hierarchical feature extraction with multi-scale spatial regularization and compacting intra-class variation. By jointly optimizing the three parts, the network is well-guided to extract more detailed FE-related features that are momentous for FER and construct a more precise decision boundary between the normal expression and alien expression.

(3) We conduct extensive experiments on 4 representative FER datasets including the lab-controlled CFEE and KDEF, and in-the-wild ExpW and RAF-DB. The results demonstrate the superiority of the proposed HS-OCFER method compared to the state-of-the-art OCC approaches.

Related Work

Facial Expression Recognition. As a complicated vision task, the key to a well-performed FER method is superior representation learning ability with separable features (Corneanu et al. 2016). To solve this problem, **conventional methods** propose various algorithms to extract non-image representation of the facial image, including Gabor feature (Lyons et al. 1998), Local Binary Pattern (Shan, Gong, and McOwan 2009), and Optical flow (Cohn et al. 1998). Recently, more and more recognition methods are proposed with the development of deep neural networks and collection of large-scale datasets (Fan et al. 2021, 2022). The **deep methods** designed for FER includes DLP-CNN (Li, Deng, and Du 2017), IL-CNN (Cai et al. 2018), SCN (Wang et al. 2020), ADDL (Ruan et al. 2020), MA-Net (Zhao, Liu, and Wang 2021), and FDRL (Ruan et al. 2021). Deep methods have achieved great improvement in ordinary FER task due to their strong power of representation learning, especially with various environmental factors in the wild.

One Class Classification. OCC, also known as novelty detection, is an unsupervised learning task of detecting samples out of the distribution from training data. In OCC, the model is trained on the data of only one class (named normal class) and ought to detect the samples that lie out of the training samples (named alien class) during inference (Perera, Oza, and Patel 2021; Pimentel et al. 2014). Different from well-researched multi-class classification, it is much tougher for an OCC classifier to learn the distinction among different classes and extract more discriminative features for decision. There are generally two types of OCC methods: non-deep and deep methods. **Non-deep methods** focus on calculating the optimal margin of the training data and learning a data-enclosing region in the sample space (Schölkopf et al. 1999; Tax and Duin 2004), which are proficient in handling structured data of a relatively small scale but may get stuck in the curse of dimensionality when coping with high-dimensional images in the wild. **Deep methods** are proposed to overcome this curse in an end-to-end manner, including the discriminative (Oza and Patel 2019, 2018) and generative deep models (Zaheer et al. 2020; Sabokrou et al. 2018).

Problem Definition

Our goal is to build a FER model that is trained with the expressions of only one class (called normal class) and can detect the expressions from alien classes during inference. Supposing c_0 is the normal expression class that can be observed during training, and $\mathbb{C} = \{c_0\} \cup \{c_k\}_{k=1}^{K}$ is the set of expression classes the model may encounter during inference, including K alien classes. Due to the fact that we do not know how many expression classes will occur, K is set as a scalar random variable. Let $\boldsymbol{x}$ and $\boldsymbol{s}$ denote the input expression image and spatial regularization data, with y as the corresponding label. Given training data $\{\boldsymbol{x}_n, \boldsymbol{s}_n, y_n\}_{n=1}^{N}$, we want to learn a feature extractor $f(\boldsymbol{x}; \boldsymbol{\Theta}_f)$ to extract deep features $\boldsymbol{z}$ from the input data, where $\boldsymbol{\Theta}_f$ is the weight parameter matrix of f. After that, an alien score function $S(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\Theta}_S)$ is needed to infer its probability score to be

from the alien expression classes $\{c_k\}_{k=1}^{K}$, where $\mathbf{\Theta}_S$ is the weight parameter matrix of S. As most existing research, our work focuses on the above feature extractor f and alien score function S (Hu et al. 2020; Perera, Nallapati, and Xiang 2019; Ruff et al. 2018). To overcome the influence of thresholding process and provide a calibration independent measurement for the given alien score, Area Under the Curve (AUC) of the Receiver Operating Characteristic is adopted as the evaluation metric.

Our Model

In this section, we present the proposed model named Hierarchical Spatial One Class Facial Expression Recognition (HS-OCFER).

Model Overview

The main framework of our model is shown in Fig 2. Specifically, the proposed HS-OCFER consists of three main components. Firstly, we devise hierarchical bottleneck modules to extract the detailed feature hierarchy from the input image. By bridging the representation gap between the encoder and decoder, the modules can relieve the pressure of underfitting and achieve better feature extraction. Secondly, we employ multi-scale spatial regularization using FE-related information to avoid the potential overfitting problem and facilitate the robustness of feature extraction in the real world. Thirdly, we adopt compact intra-class variation on the hierarchical decisive feature to construct a compact decision space for OCC task. Finally, we conclude with a summarized loss function for optimization and propose a decision-level fusion inference algorithm to incorporate the evaluation of hierarchical feature compactness and image reconstruction quality. The whole framework is trained in an end-to-end manner.

Hierarchical Feature Extraction

As a typical model for unsupervised feature extraction, autoencoder (AE) is frequently adopted in recent OCC works (Ruff et al. 2018; Chen et al. 2021). Generally, AE consists of an encoder network, a decoder network, and a much smaller bottleneck layer in the middle. Benefiting from the typical image reconstruction task and encoder-bottleneck-decoder architecture, AE is capable of learning effective latent representation in various OCC problems.

However, conventional AE model encounters a crucial challenge of underfitting when utilized in the one-class FER task, which accounts for its unpromising results in the experiment. In conventional AE, the latent feature is only extracted from the top layer of the encoder network, which emphasizes the high-level semantics while neglecting the lower-level visual representations, like low-level facial texture and middle-level muscle activation. Recent research has shown that these hierarchical representations are significantly involved with the formation of facial expressions and play an important role in FER (Du, Tao, and Martinez 2014; Zhao, Liu, and Wang 2021). Causing such valuable information to vanish through the network, the conventional bottleneck layer inappropriately limits the representation power of the whole model and leads to the underfitting problem for complicated representation learning tasks like FER.

To alleviate such an underfitting problem, we propose a series of hierarchical bottleneck modules to preserve multi-level hidden representations for better reconstruction and extract hierarchical latent features from different layers. As shown in Fig. 2, these modules are introduced as shortcut paths between the encoder and decoder in all hidden levels, which takes the feature map of the hidden layer in the encoder as its input and allows such intermediate information to flow to the decoder without further encoding. Specifically, the i-th ($i = 1, \cdots, l$) hierarchical feature extraction module firstly reduces the channels of the input feature map through convolutional and ReLU layers and derives its distilled abstract $\boldsymbol{\phi}_i = \text{ReLU}(\text{Conv}_1(\boldsymbol{x}_i))$ where $\boldsymbol{x}_i$ is the feature map of the i-th encoding layer. Afterward, there are two paths split for different purposes. For one path, further convolutional and max-pooling operations are conducted on the abstract to obtain the hierarchical latent feature $\boldsymbol{z}_i = \text{Pooling}(\text{Conv}_2(\boldsymbol{\phi}_i))$ in the i-th level. For the other path, the abstract representation is detailed through another convolutional layer and turns into a recovered feature map $\boldsymbol{x}'_i = \text{Conv}_3(\boldsymbol{\phi}_i)$ of the same size as $\boldsymbol{x}_i$. The recovered $\boldsymbol{x}'_i$ is then fused with the output $\tilde{x}_{i+1}$ of the $(i+1)$-th decoding layer using element-wise sum and fed into the i-th decoding layer. Similar to common AE, given the output image of the decoder $\tilde{\boldsymbol{x}}$, the goal of hierarchical feature extraction is defined as:

$$\mathcal{L}_{recons} = \sum_{(\boldsymbol{x},\boldsymbol{s},y)} \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2 \tag{1}$$

Multi-scale Spatial Regularization

The extraction of hierarchical features strengthens the representation power of the model, but brings another vital problem, i.e., overfitting. With the enriched latent feature comes massive redundant information that is not related to the FER task (e.g. various backgrounds, illumination changes, arbitrary pose variations), while subtle facial expression changes can be easily neglected. Moreover, due to the highly varied environment in the real world, the influence of surroundings may cause a lot of disturbance in the feature extraction. Therefore, directly applying the hierarchical reconstruction feature with no spatial regularization will lead to sub-optimal results due to those extraneous factors.

To solve such an overfitting problem, we propose to learn robust multi-scale spatial regularization from the FE-related information, e.g., facial landmark coordinates, which encodes rich spatial locations for FER task (Lv et al. 2019; Gopalan, Bellamkonda, and Chaitanya 2018; Wang et al. 2020). Specifically, given the landmark coordinates of the input image, the spatial regularization term is performed by a subtask of facial landmark detection. As shown in informed research, different convolutional layers in the network are proficient in extracting the features on different scales (Selvaraju et al. 2017). For example, the lower layers tend to focus on small-scale textures, while the higher layers usually reflect large-scale structures. Therefore, we obtain the multi-scale spatial feature of the model by concatenating the feature hierarchies $\{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_l\}$ with the latent fea-

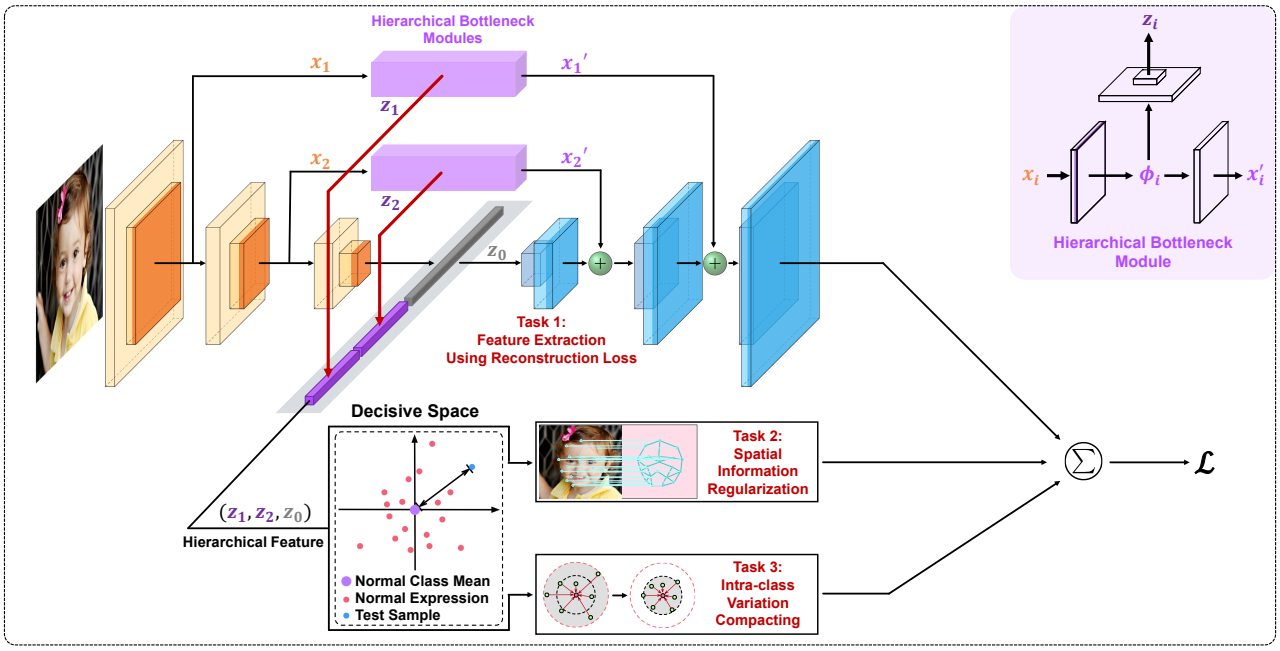

Figure 2: Main framework of the proposed HS-OCFER, which comprises three main tasks including hierarchical feature extraction, spatial information regularization, intra-class variation compacting.

ture $\boldsymbol{z}_0$ in the conventional bottleneck layer. Then through a fully-connected layer, the detection result of the landmark coordinates can be derived as $\tilde{\boldsymbol{s}}$. Following the protocol of numerical coordinate regression, given the landmark ground truth $\boldsymbol{s}$, the goal of multi-scale spatial regularization can be formulated as:

$$\mathcal{L}_{reg} = \sum_{(\boldsymbol{x},\boldsymbol{s},y)} \|\boldsymbol{s} - \tilde{\boldsymbol{s}}\|_2^2 \tag{2}$$

Compact Intra-class Variation

After extracting robust hierarchical features with multi-scale spatial regularization, a key point for OCC problem is to construct a separable decision space in which the images of the normal expression can be easily distinguished from the alien ones. Hence, in such space, different images of the normal expression are expected to have a similar feature representation and a collection of the features must be placed in a compact hypersphere during training.

In adaptation to the one-class task, we construct a separable feature space with compact intra-class variation. Similar to the spatial regularization, we concatenate the hierarchical features from the bottleneck layers to form the decisive feature $\boldsymbol{z}_d = (\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_l)$. To limit the feature distribution of the normal expression, we assume that $\boldsymbol{z}_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \cdot \boldsymbol{I})$ where $\boldsymbol{\mu}$ is a variable prototype center and $\boldsymbol{\sigma}^2 \cdot \boldsymbol{I}$ is a constant covariance matrix for the given normal class. In order to perform the parameter estimation of $\boldsymbol{\mu}$, we have to minimize the log-likelihood function of $\boldsymbol{z}_d$. From that, the Euclidean distance between the decisive feature $\boldsymbol{z}_d$ and the prototype center $\boldsymbol{\mu}$ is proved to be the minimization target of compact intra-class variation, i.e., the feature points of the normal expression class should be distributed in the neighborhood of their prototype $\boldsymbol{\mu}$. Using the Maximum Likelihood Estimation method, the goal of compact intra-class variation is defined as:

$$\mathcal{L}_{compact} = \sum_{(\boldsymbol{x},\boldsymbol{s},y)} \|\boldsymbol{z}_d - \boldsymbol{\mu}\|_2^2 \tag{3}$$

where $\boldsymbol{\mu}$ is a trainable parameter.

Model Optimization and Inference

Comprising the above three tasks, the overall optimization target function of the proposed model can be briefly summarized as

$$\mathcal{L} = \mathcal{L}_{recons} + \lambda \cdot \mathcal{L}_{reg} + \gamma \cdot \mathcal{L}_{compact} \tag{4}$$

where $\mathcal{L}_{recons}$ is the image reconstruction loss, $\mathcal{L}_{reg}$ is the spatial regularization loss, $\mathcal{L}_{compact}$ is the compact variation loss, with λ and γ as the weight parameters. By optimizing the target function, the proposed model can appropriately balance its representation power between underfitting and overfitting with the help of hierarchical reconstruction information and spatial regularization, and utilize the power to construct a competent decision space with compact intra-class variation for one-class FER task.

During inference, we propose a decision-level fusion algorithm to calculate the alien scores of the test samples based on hierarchical feature compactness and image reconstruction quality. Since the features of the normal samples are compactly constricted by the compact variation loss during training, it is supposed that the decisive feature points of the normal expression tend to be located closer to the prototype center of training samples than the alien ones. In a similar way, the reconstructed images of the normal samples

are believed to have lower errors than the alien ones. Therefore, given the input image $\boldsymbol{x}$, its centripetal distance to the prestored feature prototype $\boldsymbol{c}$ and reconstruction error to the output image $\tilde{\boldsymbol{x}}$ are both considered great candidates for the alien score function. For a more comprehensive evaluation of the test sample, we adopt the weighted fusion method to combine the decision information of the feature compactness and reconstruction quality and generate the final alien score as the model output.

Experiment Setup

In this section, we introduce the detailed experimental setup, including the datasets, task setting, baselines, and implementation details.

Datasets

In the experiment, we utilize four typical and representative FER datasets as our benchmark, including **laboratory-controlled** CFEE (Du, Tao, and Martinez 2014) and KDEF (Lundqvist, Flykt, and Öhman 1998), and **in-the-wild** ExpW (Zhang et al. 2018) and RAF-DB (Li and Deng 2019). The scale of FER datasets is generally smaller than many other computer vision tasks. This can be the difficulty of data collection and annotation caused by the subjectivity and subtleness of expressions. Furthermore, the distributions of FER datasets are also highly varied due to different environmental factors like lighting. The inadequate samples and diverse distribution cause great challenges for one-class FER.

Task Setting

All the aforementioned FER datasets have seven expression classes, from which we follow the previous OCC works (Perera, Nallapati, and Xiang 2019; Ruff et al. 2018; Hu et al. 2020; Chen et al. 2021) and create seven setups in the experiments respectively. For each of these four FER datasets, we use the training set of each expression class in turn as the normal class c_0 to train the c_0-th model and then test it on the full test set of all classes in which the rest of the expression classes except c_0 are regarded as the alien classes. In order to alleviate the impact of dataset partitioning, we randomly split each dataset into 5-folds and report the average performance over 5 runs.

Baselines

To have comprehensive coverage of related works, we choose the following non-deep and deep OCC methods as the compared baselines. For **non-deep methods**, OCSVM (Schölkopf et al. 1999) and SVDD (Tax and Duin 2004; Zeng et al. 2006) are chosen as non-deep baselines. With the usage of the kernel function, they can be extended into the corresponding kernel methods as OCSVM-k and SVDD-k. For **deep methods**, the state-of-the-art deep OCC methods can be divided into two categories: discriminative methods and generative methods. For discriminative methods, we choose DSVDD (Ruff et al. 2018), HRN (Hu et al. 2020), and MKD (Salehi et al. 2021) for comparison. For generative methods, we choose OCGAN (Perera, Nallapati, and Xiang 2019), OGNet (Zaheer et al. 2020), and IGD (Chen et al. 2021) as baselines.

The detailed properties of all the baselines and our method are summarized in Table 1. Specifically, Feature Hierarchy is the feature extraction structure from different levels, and Spatial Reg. denotes the regularization term of the spatial information to avoid overfitting.

Method	Backbone	Feature Hierarchy	Spatial Reg.	Compact Space
OCSVM	SVM	✗	✗	✗
OCSVM-k	SVM	✗	✗	✗
SVDD	SVM	✗	✗	✓
SVDD-k	SVM	✗	✗	✓
DSVDD	AE	✗	✗	✓
OCGAN	GAN	✓	✗	✗
OGNet	GAN	✓	✗	✗
HRN	MLP	✗	✗	✗
MKD	CNN	✓	✗	✗
IGD	AE	✗	✗	✓
Ours	AE	✓	✓	✓

Table 1: The categorization of the comparison methods.

Implementation Details

We build our model using *PyTorch* library. Before training, the location of facial regions and landmarks is loaded from the original dataset. The images of the facial regions are then cropped into the size of 224×224 pixels and pre-processed using L_1-norm global contrast normalization. During the training process, we choose Adam as the main optimizer and initialize its learning rate as 10^{-3}, with an exponential decay factor of 0.99. The number of training epochs is set as 200 and the mini-batch size is set as 128. The fusion weight parameter α is 0.01 for CFEE and KDEF and 0.005 for ExpW and RAF-DB. All experiments are performed on NVIDIA RTX 3070 GPU card. The code and supplementary materials are provided at https://github.com/KyleL99/HS-OCFER.

Results and Analysis

In this section, we present the statistics and analysis of the experiments. First, we evaluate the performance of the proposed method as compared to state-of-the-art OCC baselines. Second, we conduct an ablation study to analyze the profits of different components. Then we design experiments of different ratios to explore the impact of imbalanced datasets and verify the robustness of the proposed method. Finally, we visualize the attention maps to further explain the effectiveness of the proposed method. Owing to the space constraints of the main paper, the detailed results on KDEF and RAF-DB are presented in ***Supplementary Materials***.

Comparison with State-of-the-arts

We present the results of all the comparison methods on CFEE and ExpW in Table 2 and Table 3 respectively. Based on the comparison results, it can be observed that:

(1) Compared with other one-class classification methods, our method obtains superior performances in all the 4

Method	Neutral	Happy	Sad	Angry	Surprised	Disgusted	Fearful	Mean
OCSVM	$55.55_{\pm 0.41}$	$59.38_{\pm 0.56}$	$52.90_{\pm 0.29}$	$53.12_{\pm 0.55}$	$59.36_{\pm 0.93}$	$50.42_{\pm 0.22}$	$47.97_{\pm 1.00}$	54.10
OCSVM-k	$70.13_{\pm 0.49}$	$69.60_{\pm 0.49}$	$63.00_{\pm 0.49}$	$63.58_{\pm 0.37}$	$63.27_{\pm 0.77}$	$63.51_{\pm 0.18}$	$55.56_{\pm 0.81}$	64.09
SVDD	$53.92_{\pm 0.07}$	$59.92_{\pm 0.33}$	$54.67_{\pm 0.55}$	$51.20_{\pm 0.75}$	$56.31_{\pm 0.56}$	$59.47_{\pm 0.61}$	$51.52_{\pm 0.62}$	55.29
SVDD-k	$67.98_{\pm 0.47}$	$68.72_{\pm 0.72}$	$60.82_{\pm 0.94}$	$61.82_{\pm 0.86}$	$63.30_{\pm 1.00}$	$62.29_{\pm 0.42}$	$53.45_{\pm 0.86}$	62.63
OCGAN	$59.87_{\pm 0.85}$	$61.95_{\pm 0.65}$	$56.88_{\pm 0.94}$	$57.24_{\pm 0.21}$	$53.92_{\pm 0.91}$	$59.42_{\pm 0.94}$	$50.70_{\pm 0.71}$	57.14
DSVDD	$59.08_{\pm 0.57}$	$61.36_{\pm 0.76}$	$52.79_{\pm 0.47}$	$56.63_{\pm 0.51}$	$57.42_{\pm 0.60}$	$56.14_{\pm 0.47}$	$43.81_{\pm 0.53}$	55.32
OGNet	$52.21_{\pm 0.55}$	$54.40_{\pm 0.51}$	$53.99_{\pm 0.41}$	$53.36_{\pm 0.57}$	$53.61_{\pm 0.84}$	$51.68_{\pm 0.77}$	$51.45_{\pm 1.27}$	52.96
HRN	$54.23_{\pm 0.32}$	$55.04_{\pm 0.41}$	$54.44_{\pm 0.54}$	$55.09_{\pm 0.30}$	$54.04_{\pm 0.47}$	$54.41_{\pm 0.34}$	$54.31_{\pm 0.19}$	54.51
MKD	$67.80_{\pm 0.66}$	$77.02_{\pm 0.48}$	$59.11_{\pm 0.79}$	$62.72_{\pm 0.58}$	$61.91_{\pm 1.27}$	$66.49_{\pm 0.27}$	$53.68_{\pm 0.67}$	64.82
IGD	$66.21_{\pm 0.44}$	$84.45_{\pm 0.36}$	$56.54_{\pm 0.70}$	$57.30_{\pm 0.63}$	$62.65_{\pm 0.88}$	$61.75_{\pm 0.60}$	$56.85_{\pm 0.54}$	63.68
Ours	$\mathbf{74.22}_{\pm 0.71}$	$\mathbf{87.70}_{\pm 0.57}$	$\mathbf{64.29}_{\pm 1.60}$	$\mathbf{64.36}_{\pm 1.27}$	$\mathbf{79.47}_{\pm 1.21}$	$\mathbf{68.29}_{\pm 1.14}$	$\mathbf{61.24}_{\pm 1.37}$	**71.37**

Table 2: Average $\text{AUC}_{\pm \text{STD}}\%$ over 5-fold cross-validation on lab-controlled CFEE. The best method is emphasized in bold.

Method	Neutral	Happy	Sad	Angry	Surprised	Disgusted	Fearful	Mean
OCSVM	$51.74_{\pm 0.17}$	$54.99_{\pm 0.24}$	$49.79_{\pm 0.12}$	$56.96_{\pm 0.13}$	$59.92_{\pm 0.36}$	$51.08_{\pm 0.49}$	$57.70_{\pm 0.21}$	50.93
OCSVM-k	$52.36_{\pm 0.15}$	$56.67_{\pm 0.18}$	$51.28_{\pm 0.27}$	$48.50_{\pm 0.24}$	$49.29_{\pm 0.16}$	$50.35_{\pm 0.29}$	$48.04_{\pm 0.33}$	54.60
SVDD	$52.77_{\pm 0.16}$	$57.80_{\pm 0.19}$	$51.05_{\pm 0.29}$	$45.71_{\pm 0.14}$	$44.59_{\pm 0.12}$	$49.63_{\pm 0.33}$	$44.21_{\pm 0.50}$	49.39
SVDD-k	$54.86_{\pm 0.11}$	$61.68_{\pm 0.14}$	$54.21_{\pm 0.38}$	$58.55_{\pm 0.18}$	$52.18_{\pm 0.22}$	$50.56_{\pm 0.30}$	$48.43_{\pm 0.40}$	54.35
OCGAN	$56.39_{\pm 0.16}$	$60.97_{\pm 0.18}$	$51.78_{\pm 0.30}$	$42.94_{\pm 0.13}$	$42.64_{\pm 0.25}$	$48.62_{\pm 0.23}$	$37.68_{\pm 0.47}$	48.72
DSVDD	$55.51_{\pm 0.08}$	$59.37_{\pm 0.31}$	$53.13_{\pm 0.31}$	$55.18_{\pm 0.14}$	$46.61_{\pm 0.32}$	$48.92_{\pm 0.21}$	$50.78_{\pm 0.49}$	52.79
OGNet	$51.67_{\pm 0.26}$	$51.76_{\pm 0.24}$	$49.09_{\pm 0.19}$	$47.30_{\pm 0.56}$	$53.36_{\pm 0.53}$	$51.35_{\pm 0.62}$	$49.09_{\pm 0.99}$	50.52
HRN	$48.65_{\pm 0.07}$	$54.90_{\pm 0.10}$	$52.04_{\pm 0.18}$	$48.61_{\pm 0.22}$	$46.51_{\pm 0.12}$	$48.08_{\pm 0.20}$	$47.62_{\pm 0.45}$	49.49
MKD	$47.88_{\pm 0.11}$	$63.78_{\pm 0.17}$	$49.32_{\pm 0.23}$	$58.05_{\pm 0.31}$	$57.87_{\pm 0.13}$	$49.43_{\pm 0.24}$	$50.48_{\pm 0.40}$	53.83
IGD	$52.48_{\pm 0.04}$	$62.70_{\pm 0.15}$	$53.78_{\pm 0.18}$	$59.02_{\pm 0.21}$	$54.47_{\pm 0.09}$	$50.29_{\pm 0.30}$	$52.57_{\pm 0.46}$	55.05
Ours	$\mathbf{56.67}_{\pm 0.29}$	$\mathbf{64.88}_{\pm 0.32}$	$\mathbf{54.63}_{\pm 0.17}$	$\mathbf{59.68}_{\pm 0.45}$	$\mathbf{60.23}_{\pm 0.31}$	$\mathbf{52.15}_{\pm 0.67}$	$\mathbf{58.95}_{\pm 0.84}$	**58.17**

Table 3: Average $\text{AUC}_{\pm \text{STD}}\%$ over 5-fold cross-validation on in-the-wild ExpW. The best method is emphasized in bold.

datasets, with performance improvement of 6.55% (CFEE), 3.12% (ExpW), 2.87% (KDEF), and 4.38% (RAF-DB) as compared to the top baselines. The performance improvements benefit from the advantages of the proposed HS-OCFER. First, hierarchical feature extraction and multi-scale spatial regularization can enhance the representation power of the model and guide it to concentrate more on FE-related facial regions. Second, compact intra-class variation can improve the performance based on the more compact spatial distribution of the given class's feature in the decision space. Furthermore, the decision-level fusion algorithm can integrate the decisive information of both feature compactness and reconstruction quality into a more comprehensive evaluation during inference.

(2) As for the non-deep baselines, the performance on two types of FER datasets shows a huge difference. In laboratory-controlled CFEE and KDEF, non-deep methods generally function well. However, in in-the-wild ExpW and RAF-DB, non-deep methods do not show competitive performance, which is believed to be caused by the relatively larger intra-class variation of the normal expression in the real world. As mentioned in Related Work, that is exactly the weakness of non-deep methods.

(3) As for the deep baselines, there is one common phenomenon. The top deep baseline methods achieve better performance on in-the-wild ExpW and RAF-DB due to their relatively stronger representation ability and the larger scale of data. Specifically, overall IGD and MKD are the strongest baselines, which work stably on all the four datasets and reach the second-best performance on most datasets. However, the prominent deficiency of them lies in that they do not employ FE-related spatial information to filter out the extraneous disturbance. The rest of deep baselines do not show a good result due to a lack of data distribution fitting and intra-class variation controlling.

(4) Neither of the baselines can reach the best performance on 4 datasets consistently. As introduced in Datasets, there is a huge distinction between these two types of datasets which represent totally different collection conditions and environments. It is a great challenge for a method to achieve good robustness and performance on all datasets.

Ablation Study

The proposed HS-OCFER method contains three novel components: hierarchical feature extraction, multi-scale spatial regularization, and compact intra-class variation. We conduct an ablation study to further verify their effectiveness. Since the spatial regularization term is based on the multi-scale features extracted from the hierarchical modules, it cannot be separated from hierarchical feature extraction. Therefore, we testify all the six possible combinations of three components, as shown in Table 4. Besides, in order to verify the strength of the proposed decision-level fusion inference algorithm, we conduct additional experiments on different decision rules based on the well-trained HS-OCFER network, as shown in Table 5.

Feature Hierarchy	Spatial Reg.	Compact Space	CFEE	ExpW
✗	✗	✗	$51.41_{\pm 0.69}$	$53.46_{\pm 0.40}$
✓	✗	✗	$51.28_{\pm 0.26}$	$53.80_{\pm 0.66}$
✗	✗	✓	$57.54_{\pm 0.56}$	$54.28_{\pm 0.52}$
✓	✓	✗	$56.22_{\pm 0.42}$	$55.30_{\pm 0.55}$
✓	✗	✓	$65.26_{\pm 0.88}$	$56.17_{\pm 0.78}$
✓	✓	✓	$\mathbf{71.37}_{\pm 0.28}$	$\mathbf{58.17}_{\pm 0.14}$

Table 4: Average $\text{AUC}_{\pm \text{STD}}\%$ of different components.

Inference Rule	CFEE	ExpW
Reconstruction Quality Only	$69.04_{\pm 0.42}$	$57.43_{\pm 0.65}$
Feature Compactness Only	$70.42_{\pm 0.60}$	$57.82_{\pm 0.66}$
Decision-level Fusion	$\mathbf{71.37}_{\pm 0.28}$	$\mathbf{58.17}_{\pm 0.14}$

Table 5: Average $\text{AUC}_{\pm \text{STD}}\%$ of different inference rules.

As the results show, we have the following observations: (1) The basic model and that with only hierarchical feature extraction are the worst methods in all cases. (2) Simply adding compact intra-class variation performs better than the basic model. Besides, continuing to add hierarchical feature extraction and multi-scale spatial regularization can further improve the performance of the model. This demonstrates the necessity of taking these two components into consideration when minimizing the volume of a data-enclosing hypersphere. (3) The decision-level fusion algorithm performs better than using either of the reconstruction quality and feature compactness while using only reconstruction quality is less prominent. (4) All three components contribute to OCC for FER. The HS-OCFER method that jointly combines the three components and employs the decision-level fusion algorithm performs the best in all cases. These observations demonstrate the effectiveness of the proposed method.

Different Ratio

In all of the above experiments, we set the instance number ratio of the normal class and each alien class as $1:1$ to follow the mainstream OCC works (Zaheer et al. 2020). But in real application scenarios with imbalanced data, the occurrence of alien expressions usually cannot be pre-determined. To explore the impact of different ratios and get a brief evaluation of the robustness under various ratio settings, we plot the performance comparison for different ratios from $1:10$ to $10:1$ in Fig. 3. It turns out that our model performs better than other baseline methods robustly.

Visualization

To demonstrate the interpretability of our model, the heat map generated by the Grad-Cam algorithm (Selvaraju et al. 2017) is used to visualize the significance of the spatial locations learned by the basic auto-encoder model and the proposed HS-OCFER method. By comparing the regions that are considered by different networks as being important for predicting a class, we attempt to see how this network is making good use of feature extraction. For a more intuitive demonstration of the remarkable FE-related regions, we highlight the regions of the representative Action Units (AU)

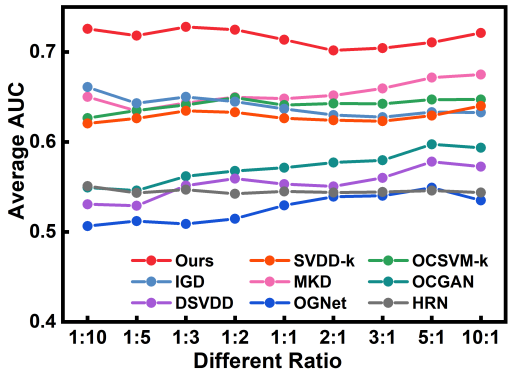

Figure 3: The performance for different ratios on CFEE.

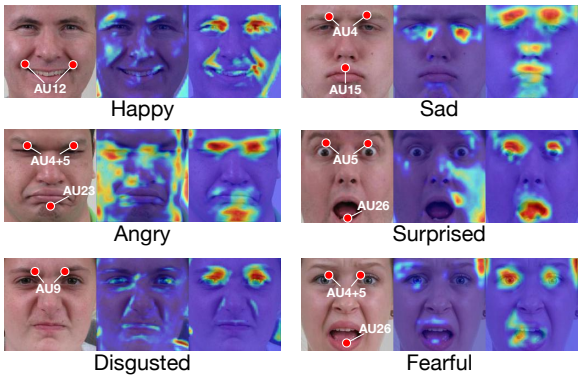

Figure 4: Visualization of the attention heatmap on some expression samples. From left to right in each group, there are: the test image with highlighted emotional AUs, the corresponding heatmap generated by the basic auto-encoder, and the proposed HS-OCFER.

(Ekman and Friesen 1978) of the corresponding expression class in each facial image. As shown in Figure 4, the proposed HS-OCFER method notices more attentive and discriminative regions that are related to FE, compared with the basic model. For example, AU12 (lip corner puller) is regarded as an exclusive action unit for happiness. In Figure 4, HS-OCFER focuses on the discriminative regions of this AU while the basic model focuses on other facial regions.

Conclusion

In this paper, by constructing a novel OCC method named HS-OCFER, we have shown that better one-class facial expression recognition can be achieved through joint learning with hierarchical feature extraction, multi-scale spatial regularization, and compact intra-class variation. Due to these three new components, the proposed HS-OCFER enhances its representation power of image reconstruction and extracts FE-related hierarchical visual information in different hidden layers. Comprehensive experiments demonstrate that the model consistently outperforms state-of-the-art classifiers on multiple small-scale laboratory-controlled and large-scale in-the-wild datasets.

Acknowledgments

The work was supported by the National Natural Science Funds of China (No. 62076146, 62021002, 61977062, 62177046, U1801263, U20A6003).

References

Cai, J.; Meng, Z.; Khan, A. S.; Li, Z.; O'Reilly, J.; and Tong, Y. 2018. Island loss for learning discriminative features in facial expression recognition. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 302–309.

Chen, Y.; Tian, Y.; Pang, G.; and Carneiro, G. 2021. Deep One-Class Classification via Interpolated Gaussian Descriptor. *arXiv preprint arXiv:2101.10043*.

Cohn, J. F.; Zlochower, A. J.; Lien, J. J.; and Kanade, T. 1998. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 396–401. IEEE.

Corneanu, C. A.; Simón, M. O.; Cohn, J. F.; and Guerrero, S. E. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1548–1568.

Du, S.; Tao, Y.; and Martinez, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15): E1454–E1462.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Fan, Z.; Chen, T.; Wang, P.; and Wang, Z. 2022. CADTransformer: Panoptic Symbol Spotting Transformer for CAD Drawings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10986–10996.

Fan, Z.; Zhu, L.; Li, H.; Chen, X.; Zhu, S.; and Tan, P. 2021. FloorPlanCAD: a large-scale CAD drawing dataset for panoptic symbol spotting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10128–10137.

Gopalan, N.; Bellamkonda, S.; and Chaitanya, V. S. 2018. Facial expression recognition using geometric landmark points and convolutional neural networks. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1149–1153. IEEE.

Hu, W.; Wang, M.; Qin, Q.; Ma, J.; and Liu, B. 2020. HRN: A holistic approach to one class learning. *Advances in Neural Information Processing Systems*.

Li, S.; and Deng, W. 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1): 356–370.

Li, S.; Deng, W.; and Du, J. 2017. 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*, 2852–2861.

Lundqvist, D.; Flykt, A.; and Öhman, A. 1998. Karolinska directed emotional faces. *Cognition and Emotion*.

Lv, C.; Wu, Z.; Wang, X.; and Zhou, M. 2019. 3D facial expression modeling based on facial landmarks in single image. *Neurocomputing*, 355: 155–167.

Lyons, M.; Akamatsu, S.; Kamachi, M.; and Gyoba, J. 1998. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, 200–205. IEEE.

Oza, P.; and Patel, V. M. 2018. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26(2): 277–281.

Oza, P.; and Patel, V. M. 2019. Active authentication using an autoencoder regularized cnn-based one-class classifier. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 1–8.

Perera, P.; Nallapati, R.; and Xiang, B. 2019. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2898–2906.

Perera, P.; Oza, P.; and Patel, V. M. 2021. One-class classification: A survey. *arXiv preprint arXiv:2101.03064*.

Pimentel, M. A.; Clifton, D. A.; Clifton, L.; and Tarassenko, L. 2014. A review of novelty detection. *Signal Processing*, 99: 215–249.

Ruan, D.; Yan, Y.; Chen, S.; Xue, J.-H.; and Wang, H. 2020. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2833–2841.

Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; and Wang, H. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7660–7669.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International Conference on Machine Learning*, 4393–4402.

Sabokrou, M.; Khalooei, M.; Fathy, M.; and Adeli, E. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3379–3388.

Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M. H.; and Rabiee, H. R. 2021. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14902–14912.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 1999. Estimating the support of a high-dimensional distribution. *Technical Report MSR-T R-99–87, Microsoft Research.*

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shan, C.; Gong, S.; and McOwan, P. W. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6): 803–816.

Tax, D. M.; and Duin, R. P. 2004. Support vector data description. *Machine Learning*, 54(1): 45–66.

Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6906.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, 499–515.

Xi, H.; Aussel, D.; Liu, W.; Waller, S. T.; and Rey, D. 2022a. Single-leader multi-follower games for the regulation of two-sided mobility-as-a-service markets. *European Journal of Operational Research*.

Xi, H.; He, L.; Zhang, Y.; and Wang, Z. 2022b. Differentiable road pricing for environment-oriented electric vehicle and gasoline vehicle users in the bi-objective transportation network. *Transportation Letters*, 14(6): 660–674.

Zaheer, M. Z.; Lee, J.-h.; Astrid, M.; and Lee, S.-I. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 14183–14193.

Zeng, Z.; Fu, Y.; Roisman, G. I.; Wen, Z.; Hu, Y.; and Huang, T. S. 2006. One-class classification for spontaneous facial expression analysis. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 281–286. IEEE.

Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1): 39–58.

Zhang, L.; and Tjondronegoro, D. 2011. Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, 2(4): 219–229.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5): 550–569.

Zhao, Z.; Liu, Q.; and Wang, S. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30: 6544–6556.