MACHINE LEARNING FOR PERCEIVING FACIAL MICRO-EXPRESSION
YANTE LI

MACHINE LEARNING FOR PERCEIVING FACIAL MICRO-EXPRESSION

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Technology and Natural Sciences of the Doctoral Programme Committee of Information Technology and Electrical Engineering, Linnanmaa, on 25 May 2022, at 12 noon

UNIVERSITY OF OULU, OULU 2022
Li, Yante, **Machine learning for perceiving facial micro-expression.**
University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering
*Acta Univ. Oul. C* 827, 2022
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

**Abstract**

Emotion analysis plays an important role in humans' daily lives. Facial expression is one of the major ways to express emotions. Besides the common facial expressions we see every day, emotion can also be expressed in a special format, micro-expression. Micro-expressions (MEs) are involuntary facial movements that come about in reaction to emotional stimulus, which reveal people’s hidden feelings in high-stakes situations and have many potential applications, such as clinical diagnosis, ensuring national security, and conducting interrogations. However, ME recognition becomes challenging due to the low intensity, short duration and small-scale datasets.

This thesis is a thorough summary of the important subjects for ME recognition, consisting of five papers corresponding to the progress of my research. Firstly, the automatic ME recognition system based on deep learning is introduced. Secondly, the Micro-expression Action Unit (ME-AU) detection is described, which plays an important role in facial behavior analysis. Thirdly, the robust ME recognition with AU detection is illustrated that verifies the contribution of AU detection to ME recognition.

The contributions of this study can be classified into three categories: (1) A deep ME recognition approach with the apex frame is proposed, which would be capable of demonstrating that deep learning can achieve impressive performance of ME recognition with the apex frame; (2) We break the ground of the ME-AU study and provide the baselines and novel transfer learning methods for the future study of ME-AU detection; (3) A unified framework for ME recognition with AU detection based on contrastive learning is proposed for verifying the AU contribution to robust ME recognition.

Lastly, we summarize the contributions of the work, and propose future plans about ME studies based on the limitations of the current work.

**Keywords:** action unit detection, affective, deep learning, machine learning, micro-expression

Tämä opinnäytetyö on kattava yhteenveto mikroilmeiden tunnistuksen kannalta tärkeistä aiheista, ja se koostuu viidestä tutkimukseni vastaavasta artikkelista. Ensimmäiseksi otetaan käyttöön syväoppimiseen perustuva automaattinen mikroilmeentunnistusjärjestelmä. Toiseksi esitellään mikroilmeiden aktioyksikkö tunnistus, jolla on tärkeä rooli kasvojen käyttäytymisen analysoinnissa. Kolmannaksi esitetään robusti mikroilmeiden tunnistus aktioyksikköjen avulla, joka vahvistaa aktioyksikköjen tuloksen mikroilmeiden tunnistukseen.

Tämän tutkimuksen tulokset voidaan luokitella kolmeen osaan: (1) Mikroilmeiden tunnistukseen ehdotetaan perusteellista lähestymistapaa videon apektikohdan avulla, mikä osoittaa, että syväoppiminen voi edistää mikroilmeiden tunnistusta videoon apektin ansiosta; (2) Avamme uuden uran mikroilmeiden aktioyksikköjen tutkimuksele ja tarjoamme perustason ja uusia siirrymisoppimismenetelmiä tulevaa mikroilmeiden aktioyksikköjen tunnistusta varten; (3) Ehdotamme yhtenäistä kehystä, jolla mikroilmeet voidaan tunnistaa aktioyksikköjen ja kontrastii- sen oppimisen avulla ja jolla voimme vahvistaa aktioyksikköjen merkityksen vahvassa mikroilmeiden tunnistuksessa.

Lopuksi teemme yhteenvedon työn tuloksista ja ehdotamme tulevaisuuden suunnitelmaa mikroilmeiden tutkimuksille nykyisen työn rajoitusten perusteella.

**Asiasanat:** affektiivinen, mikroilme, piirteiden ilmaisu
To my family and friends.
Acknowledgements

The research work of the thesis was carried out at the Center of Machine Vision and Signal Analysis (CMVS) at the University of Oulu between 2017 and 2022.

My research work on micro-expression analysis is conducted under the supervision of Academy Professor Guoying Zhao. I would like to express my deepest appreciation and thanks to Guoying. She gives me all of the freedom to grow into an independent researcher. She not only supports my research, but also gives me valuable suggestions on my life and career. I would also like to thank Prof. Xiaohua Huang for giving me technical suggestions on my research.

I would like to acknowledge all my co-authors Prof. Xiaohua Huang, Msc. Wei Peng for their valuable discussions and comments. Thanks to the members in our center, past and present, for creating a wonderful research environment.

I want to acknowledge the Infotech Oulu Doctoral Program for its financial support. As well, I wish to acknowledge the CSC IT Center for Science, Finland, for computational resources.

In particular, I shall thank Professor Moi Hoon Yap from Manchester Metropolitan University and Professor Yingli Tian from the City University of New York for their valuable comments and suggestions which help to improve the quality of this thesis.

I would also like to thank Professor Robert Jenssen from UiT-the Arctic University of Norway for serving as the opponent in my defense.

Finally, I would like to express my gratitude to my parents for their support during all the stages of my life. And I would also like to thank my friends in Oulu, who have accompanied and supported me forever and made my life in Oulu happy and interesting.

In Oulu, Finland, 17th of December, 2021
Yante Li
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>Micro-expression</td>
</tr>
<tr>
<td>SMIC</td>
<td>Spontaneous Micro facial expression database</td>
</tr>
<tr>
<td>CASME</td>
<td>Chinese Academy of Sciences micro-expression database</td>
</tr>
<tr>
<td>CASME II</td>
<td>The 2nd version of CASME</td>
</tr>
<tr>
<td>SAMM</td>
<td>Spontaneous Micro-Facial Movement Dataset</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>FE</td>
<td>Facial Expression</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>3D-FFT</td>
<td>3D Fast Fourier Transformation</td>
</tr>
<tr>
<td>fps</td>
<td>Frames per second</td>
</tr>
<tr>
<td>LBP</td>
<td>Local Binary Pattern</td>
</tr>
<tr>
<td>LBP-TOP</td>
<td>LBP on Three Orthogonal Planes</td>
</tr>
<tr>
<td>LBP-SIP</td>
<td>Local Binary Pattern with Six Intersection Points</td>
</tr>
<tr>
<td>STCLQP</td>
<td>Spatio-Temporal Completed Local Quantized Pattern</td>
</tr>
<tr>
<td>NMAE</td>
<td>Normalized Mean Absolute Error</td>
</tr>
<tr>
<td>NSE</td>
<td>Normalized Standard Error</td>
</tr>
<tr>
<td>ROIs</td>
<td>Regions of Interests</td>
</tr>
<tr>
<td>RHOOF</td>
<td>Optical Flow Histogram based on ROIs</td>
</tr>
<tr>
<td>OS-ROI</td>
<td>Optical Strain based on ROIs</td>
</tr>
<tr>
<td>GI</td>
<td>Global Information</td>
</tr>
<tr>
<td>LI</td>
<td>Local Information</td>
</tr>
<tr>
<td>FC</td>
<td>Fully Connected layer</td>
</tr>
<tr>
<td>MIL</td>
<td>Multiple Instance Learning</td>
</tr>
<tr>
<td>EMME</td>
<td>Extended Magnified ME database</td>
</tr>
<tr>
<td>NG</td>
<td>Without eyeglasses</td>
</tr>
<tr>
<td>WG</td>
<td>With eyeglasses</td>
</tr>
<tr>
<td>LOSO</td>
<td>Leave One Subject Out</td>
</tr>
<tr>
<td>LOVO</td>
<td>Leave One Video Out</td>
</tr>
<tr>
<td>LPQ</td>
<td>Local Phase Quantization</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>HOG</td>
<td>Histograms of Oriented Gradients</td>
</tr>
<tr>
<td>LGGBP</td>
<td>Local Gabor Binary Pattern</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>DRML</td>
<td>Deep Region and Multi-label Learning</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long-short-term mode</td>
</tr>
<tr>
<td>CA</td>
<td>Channel Attention</td>
</tr>
<tr>
<td>SA</td>
<td>Spatial Attention</td>
</tr>
<tr>
<td>SCA</td>
<td>Spatio-Channel Attention</td>
</tr>
<tr>
<td>SGD</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>TIM</td>
<td>Temporal Interpolation Model</td>
</tr>
<tr>
<td>ASP</td>
<td>Attentive Similarity-Preserving distillation</td>
</tr>
</tbody>
</table>
List of original publications

This thesis is based on the following articles, which are referred to in the text by their Roman numerals (I–V):


IV Yante Li & Guoying Zhao (2021). Micro-expression Action Unit Detection with Dual-view Attentive Similarity-Preserving Knowledge Distillation. IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 01-08. IEEE.


The present author is the first author in all of the publications. The experiments and writing were carried out by the present author, while valuable suggestions and discussions were given by the co-authors.
Contents

Abstract .......................... 9
Tiivistelmä .......................... 9
Acknowledgements .................. 11
List of abbreviations ............... 11
List of original publications ......... 13
Contents ............................ 15
1 Introduction ......................... 17
   1.1 Background and motivation .......... 17
   1.2 Micro-expression databases .......... 18
   1.3 Objectives ........................ 20
   1.4 Evaluation metrics ................ 21
   1.5 Summary of original papers .......... 22
   1.6 Organization of the thesis .......... 23
2 Micro-expression recognition based on deep learning ........ 25
   2.1 Introduction ........................ 25
   2.2 Related work ........................ 26
       2.2.1 Review of ME apex frame spotting studies .... 26
       2.2.2 Review of ME recognition studies .......... 27
   2.3 ME apex frame spotting ............. 28
       2.3.1 A method for ME apex frame spotting ..... 28
       2.3.2 Experiments ..................... 30
   2.4 ME recognition via joint local and global information learning .... 33
       2.4.1 A joint local and global information learning method ... 33
       2.4.2 Model training ................... 36
       2.4.3 Experiments ..................... 37
   2.5 A framework for ME recognition with detected apex frame .... 40
   2.6 Summary .......................... 41
3 Micro-expression action unit detection ........ 43
   3.1 Introduction ........................ 43
       3.1.1 Facial action coding system and action units .... 43
       3.1.2 Challenges of ME-AU detection .......... 44
   3.2 Related work ........................ 45
       3.2.1 Traditional approaches .............. 45
       3.2.2 Deep learning approaches ............ 46
3.3 ME-AU detection with attention ........................................... 47
  3.3.1 The framework of ME-AU detection with attention ............. 47
  3.3.2 Channel attention module ............................................... 49
  3.3.3 Spatial attention module ............................................... 50
  3.3.4 The fusion of channel and spatial attention modules .......... 51
  3.3.5 Model training .......................................................... 52
  3.3.6 Experiments ............................................................. 52

3.4 ME-AU detection with dual-view knowledge distillation ............ 57
  3.4.1 Overview ............................................................... 58
  3.4.2 Dual-view co-training for robust AU detection .................... 60
  3.4.3 Attentive similarity-preserving knowledge distillation ......... 62
  3.4.4 Model training ........................................................ 64
  3.4.5 Experiments ........................................................... 65

3.5 Summary ................................................................. 68

4 Micro-expression recognition with action units 71
  4.1 Introduction ............................................................. 71
  4.2 Related work ............................................................ 72
  4.3 ME recognition with AU detection based on contrastive learning 72
    4.3.1 Intra-contrastive learning ......................................... 72
    4.3.2 Inter-contrastive learning ......................................... 74
    4.3.3 ME-AU detection objective ........................................... 75
    4.3.4 A framework for ME recognition with AU detection .......... 75
    4.3.5 Model training ........................................................ 76
    4.3.6 Experiments ........................................................... 76
  4.4 Summary ................................................................. 79

5 Summary 81
  5.1 Contributions .......................................................... 81
    5.1.1 ME recognition with deep learning ................................. 81
    5.1.2 ME-AU detection ...................................................... 82
    5.1.3 ME recognition with AU detection ................................ 83
  5.2 Limitations and future work ............................................ 83

References 85

Original publications 95
1 Introduction

1.1 Background and motivation

Emotion plays an important role in humans’ daily lives. Everything we do, everything we say, somehow reflects some of our emotions (Manzoor, Mohsin, Mahira, & Mudasir, 2017). To better understand human behavior, it is necessary to analyze emotions through emotional data including voice, text, gestures, facial expressions and so on. However, the emotions can be suppressed by people’s high motivation, leading to the concealment of their true feelings.

In the last century, the study of Ekman and Friesen (1969) reported the finding of micro-expressions (MEs), which may occur in high-stakes situations when people try to hide their true feelings for either avoiding loss or gaining advantage. MEs are involuntary, fast, and difficult to control through one’s willpower. Therefore, MEs can be used as an important clue for analyzing hidden emotions and have many potential applications, such as clinical diagnosis, education, business dealing and interrogation.

Because of the wide potential applications of ME analysis, ME recognition has attracted the attention of researchers in the field of computer vision. The first ME recognition research can be traced to work of Pfister, Li, Zhao, and Pietikäinen (2011) which utilized a Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) (G. Zhao & Pietikainen, 2007). Following Pfister et al. (2011)’s work, various methods (X. Li et al., 2018; Y. Liu et al., 2016) based on geometry and appearance features have been proposed for improving the performance of ME recognition. Most of the early ME research only focuses on studying MEs based on hand-crafted features.

In recent years, deep learning has been successfully applied to various tasks, such as object detection (L. Liu et al., 2020) and facial expression recognition (S. Li & Deng, 2020) and has been verified as effective for learning discriminative representations. Some researchers (Patel, Hong, & Zhao, 2017; Quang, Chun, & Tokuyama, 2019; Quang et al., 2019) have begun to exploit deep neural networks for ME recognition. However, the deep learning-based ME recognition becomes challenging due to subtle facial movements and small-scale ME databases which are far from sufficient to train a robust model.

Another challenge of ME recognition is the ambiguities in ME interpretation, e.g., the lowering of brow may refer to tense or disgust. Recent research (Davison, Merghani, & Yap, 2018) demonstrates that encoding expressions through facial Action Units (AUs) coded by the Facial Action Coding System (FACS) is effective for resolving the issue of
ambiguity. FACS is a comprehensive system for taxonomizing all visually discernible facial movements. It was originally created by Hjortsjö and Carl-Herman (1969) with 23 facial movements and subsequently developed by E. Friesen and Ekman (1978) and Hager, Ekman, and Friesen (2002). According to FACS, the facial expressions can be broken down into individual components of facial muscle movements, called Action Units. AUs work as the building blocks to formulate multiple facial expressions and play an important role in human emotion understanding. Ekman declared that he would ever never discover MEs without FACS which allowed him to look so precisely at everything the faces do. In other words, it is very essential to explore AUs for deeply interpreting the facial behavior of MEs.

Furthermore, the Emotional Facial Action Coding System designed by W. V. Friesen, Ekman, et al. (1983) and the Facial Action Coding System Affect Interpretation Dictionary proposed by Ekman, Rosenberg, and Hager (1998) consider emotion-related AUs and define the mapping between AUs and emotions. However, current ME databases do not follow strict links for facial expressions and their psychological interpretations. Therefore, it is meaningful to study how to utilize the AU information for ME analysis.

1.2 Micro-expression databases

Different from macro-expressions, MEs are brief involuntary facial expressions, particularly occurring under high stakes. The above ME characteristics make it difficult to collect MEs, leading to small-scale ME databases. It is better to use a high-speed
Table 1. Spontaneous ME databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Resolution</th>
<th>Frame rate</th>
<th>Samples</th>
<th>subjects</th>
<th>Expression</th>
<th>AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMIC HS/NIR/VIS</td>
<td>640 × 480</td>
<td>100/25/25</td>
<td>164/71/71</td>
<td>16/8/8</td>
<td>3</td>
<td>○</td>
</tr>
<tr>
<td>CASME</td>
<td>640 × 480</td>
<td>60</td>
<td>195</td>
<td>35</td>
<td>8</td>
<td>●</td>
</tr>
<tr>
<td>CASME II</td>
<td>640 × 480</td>
<td>200</td>
<td>247</td>
<td>35</td>
<td>5</td>
<td>●</td>
</tr>
<tr>
<td>SAMM</td>
<td>2040 × 1088</td>
<td>200</td>
<td>159</td>
<td>32</td>
<td>7</td>
<td>●</td>
</tr>
<tr>
<td>Composite ME</td>
<td>1280 × 720</td>
<td>200</td>
<td>442</td>
<td>68</td>
<td>3</td>
<td>○ ●</td>
</tr>
</tbody>
</table>

1 ○ represents labeled and ● represents unlabeled.

camera to collect MEs for conducting ME recognition and the minimum frame rate of the camera would be 60 fps. Most existing ME databases are collected in laboratory environment. During experiments, the participants are required to keep a “poker face” and to report their emotions after watching each stimulant video. The example samples of the ME databases are shown in Figure 1. The details of the databases (shown in Table 1) used in this thesis are introduced as the following:

**SMIC** (X. Li, Pfister, Huang, Zhao, & Pietikainen, 2013) includes three subsets: SMIC-HS, SMIC-VIS and SMIC-NIR. SMIC-VIS and SMIC-NIR contain 71 samples recorded by normal speed cameras with visual (VIS) and near-infrared light range (NIR), respectively. SMIC-HS recorded by high-speed cameras with 100 fps is used for ME recognition. SMIC-HS collects 164 ME samples from 16 subjects. These samples are divided into three categories based on self-reports: positive, negative, and surprise.

**CASME** (W. Yan, Wu, Liu, Wang, & Fu, 2013) contains 159 ME clips from 19 subjects. It is recorded at 60 fps. Samples in the CASME database are categorized into seven ME emotions partly based on AUs and also taking account of participants’ self-report and the video episodes. The specific emotions are listed as follows: happiness, disgust, sadness, surprise, fear, tenseness, repression, and contempt. Furthermore, the apex frames that contribute the most of emotion information are also labeled.

**CASME II** (W. J. Yan, Li, et al., 2014) is an improved version of the CASME database. The samples in CASME II are recorded by a high-speed camera at 200 fps. The samples increased to 247 MEs with 26 valid participants. There are five kinds of ME expressions: happiness, surprise, disgust, repression, and others.

**SAMM** (Davison, Lansley, Costen, Tan, & Yap, 2018) contains 159 ME samples collected by a high-speed camera at 200 fps in controlled lighting conditions designed to prevent flickering. Unlike other databases lack of ethnic diversity, the 32 participants
are from 13 different ethnicities. SAMM is coded using the FACS. It includes the ME emotion classes *happy, sad, surprise, angry, disgust, fear, contempt, and other.*

**The composite database** (See, Yap, Li, Hong, & Wang, 2019) is proposed by the 2nd Micro-Expression Grand Challenge (MEGC2019), which merges samples from SMIC-HS (X. Li et al., 2013), CASME II (W. J. Yan, Li, et al., 2014), and SAMM (Davison, Lansley, et al., 2018) databases. In this way, the composite database can evaluate methods on data with different natures. To unify the different emotion labels in the three databases, the emotion labels in the composite database are re-annotated as *positive, negative, and surprise.*

### 1.3 Objectives

The main objective of the thesis is building effective models for ME recognition based on annotated video clips. Firstly, we study the apex frame contribution for ME recognition and build an automatic ME apex frame spotting and recognition system. The work demonstrates that the apex frame has an important contribution to ME analysis and only an apex frame can work well for ME recognition. Given that deep learning has achieved considerable performance in facial expression recognition on massive facial images, we further develop a joint feature learning architecture coupling local and global information for ME recognition with the single apex frame.

Besides, there are ambiguities in ME interpretation. Encoding facial expressions via AUs has been found to be effective in resolving the ambiguity issue among different expressions. Therefore, AU detection plays an important role in emotion analysis. While a number of AU detection methods have been proposed for common facial expressions, there is very limited study of ME-AU detection. ME-AU detection is challenging because of the small quantity of ME databases and the low intensity of MEs. Thus, the second objective of this thesis is to design methods for robust ME-AU detection and expecting to contribute to the community. Specifically, a novel ME-AU detection method is proposed utilizing the self high-order statistics of spatio-wise and channel-wise features to capture subtle regional changes. As this is the first work aimed at ME-AU detection, we provide the baseline results of ME-AU detection to the research community. On the other hand, considering the small-scale ME databases, a novel dual-view attentive similarity-preserving distillation method is proposed for robust ME-AU detection by leveraging massive facial expressions in the wild. Through such an attentive similarity-preserving distillation method, we break the domain shift problem and essential AU knowledge from common facial AUs is efficiently distilled.
Finally, the contribution of AU information to ME recognition is studied. In order to efficiently utilize the AU information, the contrastive learning strategy is further studied to fully explore subtle AU information and a multi-task learning framework is applied to achieve robust ME recognition through leveraging AU information obtained by an AU detection task.

1.4 Evaluation metrics

This section introduces the evaluation metrics for ME recognition and ME-AU detection: accuracy and F1 score. **Accuracy** is the common evaluation metric for ME recognition. In general, the accuracy metric measures the ratio of correct predictions over the total evaluated samples. The accuracy is defined as

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

in which TP, TN, FP, FN represent the value of true positive, true negative, false positive, and false negative samples, respectively.

However, the accuracy is susceptible to bias data. The **F1 score** solves the bias problem by considering the precision and recall to reveal the true classification performance. The precision is calculated as the ratio of true positive examples among the examples that the model classified as positive,

\[
\text{Precision} = \frac{TP}{TP + FP}.
\]

The recall is the fraction of examples classified as positive, among the total number of positive examples. The recall is defined as

\[
\text{Recall} = \frac{TP}{TP + FN}.
\]

The F1 score is the harmonic mean of the precision and the recall, which is calculated by

\[
F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.
\]

For ME-AU detection with imbalanced data, the F1-score is utilized to interpret the algorithm’s performance.
1.5 Summary of original papers

Five articles are included in the thesis on the topic of ME analysis. Papers I and II concentrate on ME recognition based on deep learning. Papers III-V focus on ME-AU detection.

- A novel ME apex frame detection method is proposed which locates the apex frame by estimating pixel-level change rates in the frequency domain. With frequency information, it performs more effectively on apex frame spotting than the currently existing apex frame spotting methods based on the available spatio-temporal change information. The ME is only recognized based on the detected magnified apex frame with deep model. The results indicate that the apex frame can substantially contribute to ME recognition. The method has been published in Paper I.

- Since not all regions make the same contribution to ME recognition and some regions do not even contain any emotional information, a joint feature learning architecture combined local and global information is proposed to recognize MEs. Leveraging the local and global information enables our model to learn discriminative ME representations and suppresses the negative influence of unrelated regions to MEs. The contribution has been published in Paper II.

- A framework for ME-AU detection is proposed. Due to the small quantity of ME databases and low intensity of MEs, ME-AU detection becomes challenging. To alleviate these problems, a novel ME-AU detection method is designed by utilizing self high-order statistics of spatio-wise and channel-wise features which can be considered as spatial and channel attentions, respectively. Through such spatial and channel attention module, we expect to utilize rich relationship information of facial regions to increase the AU detection robustness on limited ME samples. The details of the model has been provided in Paper III.

- To address the small-scale issue of the ME database, a dual-view attentive similarity-preserving distillation method is proposed for robust ME-AU detection by leveraging massive facial expressions in the wild. Such an attentive similarity-preserving distillation method can break the domain shift problem and essential AU knowledge from common facial AUs is efficiently distilled. Furthermore, a semi-supervised co-training approach is developed to construct a generalized teacher network for learning discriminative AU representation. Extensive experiments illustrate that our proposed knowledge distillation method can effectively distill and transfer the cross-domain knowledge for robust ME-AU detection. The description of the model has been provided in Paper IV.
We propose a novel ME-AU detection method by learning the intra- and inter-contrastive information among MEs to alleviate the low intensity problem of MEs. Through the intra-contrastive learning module, the difference between the onset and apex frames is enlarged and utilized to obtain the discriminative representation for low-intensity AU detection. In addition, considering the subtle difference between ME-AUs, the inter-contrastive learning is designed to automatically explore and enlarge the difference between different AUs to enhance ME-AU detection robustness. Intensive experiments on ME databases demonstrate the effectiveness and generalization ability of our proposed method. This contribution has been published in Paper V.

1.6 Organization of the thesis

The thesis is organized into five chapters as follows:

Chapter 1 briefly introduces the background of ME analysis, and also indicates the aims and objectives of this thesis, the contents of the five original articles, and the outline of the thesis.

Chapter 2 introduces the work about ME recognition based on deep learning, including the ME apex frame spotting method, the ME recognition method, and the fully automatic ME recognition system. This part of the contents includes methods and results that were originally presented in papers I and II.

Chapter 3 explores AU detection in MEs. In this chapter, the topic will be introduced by illustrating challenges in ME-AU detection and discussing how our attentive modules and transfer learning methods proposed in Paper III and IV overcome the challenges.

Chapter 4 focuses on ME recognition with AU information. This chapter introduces a contrastive learning method for robust AU detection described in Paper V and develops it into a multi-task learning framework for extracting ME and AU feature simultaneously. In this way, the discriminative AU information can be utilized for robust ME recognition.

Chapter 5 summarizes the contributions of all the works and discusses the limitations and possible future work.
2 Micro-expression recognition based on deep learning

2.1 Introduction

MEs are involuntary facial movements that react to emotional stimulus (Ekman & Friesen, 1971a). MEs can reveal people’s hidden feelings in high-stakes situations and have many potential applications in different fields, such as clinical diagnosis, national security, and interrogations. Different from ordinary facial expressions that we see daily, MEs have short duration (1/25 to 1/3 second), low intensity, and occur with sparse facial action units (Ekman, 2009). All of the above characteristics make MEs difficult to detect and recognize.

Generally, two main tasks are included in ME analysis: spotting and recognition. The spotting task aims to identify ME occurrence (J. Li, Soladie, Seguier, Wang, & Yap, 2019), while the recognition task classifies the MEs into specific emotion categories (Kim, Baddar, & Ro, 2016; Liong, See, Wong, & Phan, 2018; Xu, Zhang, & Wang, 2017). Most of the current research on ME recognition utilizes whole video clips (X. Huang, Zhao, Hong, & Zheng, 2016; Mayya, Pai, & Pai, 2016; S.-J. Wang et al., 2015; G. Zhao & Pietikainen, 2007). Ekman (1993) declared that ‘snapshot taken at a point when the expression is at its apex can easily convey the emotion message’. This means that, the apex frame can contribute major information to facial expression recognition. Recently, Liong et al. (2018) discovered that the redundancy information in ME clips could decrease the performance of ME recognition. In contrast, the onset, apex, and offset frames provide useful information to ME classification. Moreover, Liong, See, Wong, and Phan (2016) proposed a bi-weighted orientation optical flow feature extracted on the spotted apex frame for ME recognition. However, so far there are few studies that analyze the contribution of the apex frame to ME recognition.

On the other hand, as deep learning technology has achieved considerable performance in facial expression recognition (F. Zhang, Zhang, Mao, & Xu, 2018), some researchers have begun to exploit deep neural networks for ME recognition (Patel et al., 2017; Peng, Wang, Chen, Liu, & Fu, 2017). However, their proposed methods dramatically degraded the performance compared to hand-crafted methods (X. Huang et al., 2016). This is explained by the fact that ME databases are very small and the changes in MEs are subtle.
Motivated by the above-mentioned observations (Ekman, 1993; Liong et al., 2018; Patel et al., 2017; Peng et al., 2017), this section studies ME recognition based on apex frame with deep learning. Firstly, it revisits the ME spotting and ME recognition studies. Then, we introduce the novel ME apex frame detection method based on frequency in Section 2.3. An ME recognition method with deep learning is described in section 2.4. Finally, an ME recognition framework with the detected apex frame is discussed.

2.2 Related work

In the related work section, research studies concerning ME analysis are reviewed in the two subsections. Subsection 2.2.1 is about ME spotting methods and Subsection 2.2.2 is about ME recognition methods.

2.2.1 Review of ME apex frame spotting studies

ME spotting aims to identify ME occurrence or finding the onset, apex and offset frames (Kim et al., 2016; Liong et al., 2018; Xu et al., 2017). Specifically, the onset frame and offset frame are the starting and the end frames in an ME clip, and the apex frame is the frame with the largest intensity. As this section studies ME recognition based on apex frame with deep learning, we focus on spotting ME apex frames in ME clips. Due to the subtle and rapid characteristics of MEs, it is difficult to locate the apex frames accurately. At the beginning, most ME spotting methods located ME apex frames by computing the feature difference between frames. Moilanen, Zhao, and Pietikäinen (2014) proposed spotting MEs based on the Chi-Square distance of the LBP in fixed-length sliding windows. X. Li et al. (2018) proposed a training-free method based on the feature difference contrast and peak detection to spot MEs. On the other hand, Liong, See, Wong, and Le (2015) used the binary search strategy with a local binary pattern and optical flow on several interesting facial sub-regions to spot the apex frame in ME clips. And Ma, An, Wu, and Yang (2017) further improved the performance of apex frame spotting by utilizing the histogram of oriented optical flow.

Inspired by the use of deep Convolutional Neural Networks (CNNs) in action detection (Vahdani & Tian, 2021), Z. Zhang, Chen, Meng, Liu, and Fu (2018) firstly utilized a CNN to extract features for ME spotting. Then, a feature matrix processing method based on a sliding window was applied to search the apex frame in MEs. Tran, Vo, Hong, and Zhao (2019) proposed employing an LSTM network to capture the local and global correlation of the spatial-temporal feature to predict the ME apex frame.
Moreover, an end-to-end deep framework was proposed to spot ME by using 2D+1D spatio-temporal CNN (S.-J. Wang, He, Li, & Fu, 2021).

However, all above methods merely concerned the subtle spatial change between neighboring frames, but omitted the rapid change of frames along the temporal domain. In contrast, our proposed ME apex frame spotting method based on 3D Fast Fourier Transform (FFT) not only analyzes rapid changes of ME in the frequency domain, but also leverages the spatial and continuous temporal information. Specific information on the proposed method is described in Section 2.3.

### 2.2.2 Review of ME recognition studies

The spontaneous ME recognition research can be traced to the work of Pfister et al. (2011). Following Pfister et al. (2011)’s work, X. Li et al. (2013) proposed recognizing ME by using a Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) and classical classifiers. For increasing the efficiency of LBP-TOP for ME recognition, Y. Wang, See, Phan, and Oh (2014a) proposed a spatio-temporal descriptor with six intersection points (LBP-SIP), also suppressed the redundancy information of LBP-TOP. In order to improve the performance of ME recognition, certain spatio-temporal descriptors have been proposed, e.g., the Spatio-Temporal Completed Local Quantized Pattern (STCLQP) (X. Huang et al., 2016) and a histogram of image gradient orientation (X. Li et al., 2018). Likewise, other feature types like main direction main optical (Y. Liu et al., 2016) and tensor independent color space (S. Wang, Yan, Li, & Zhao, 2014) methods were proposed. All the aforementioned methods are mostly based on the whole video clip. However, there remains a query over which frame could significantly contribute to ME recognition. Liong et al. (2016) attempted to use apex frames for ME recognition, but unfortunately, this system based on apex frames cannot making an improvement, and in contrast, it still behaved worse than the state-of-the-art methods (X. Li et al., 2018; Y. Liu et al., 2016) throughout the whole video clip. Even so, using an apex frame could obtain high efficiency to some extent in a real-world application.

In recent years, deep learning has achieved promising performance in many research fields (F. Zhang et al., 2018). It has also been used in ME recognition (Alexey et al., 2015; Peng et al., 2017). The work of Patel et al. (2017) was the first to transfer deep CNN models from objects and facial expressions to small ME databases. However, its recognition rate on the CASME II database is 47.3%, which is worse than hand-crafted descriptors. Peng et al. (2017) proposed a dual-template CNN model based on optical flows extracted from the ME sequences for ME recognition. The optical flow information
over the whole video should first be extracted and then fed into CNN. Actually, the extraction of optical flow led to heavy computation in real-world applications, which seriously degraded the efficiency of the dual-template CNN model. J. Li, Wang, See, and Liu (2018) proposed a novel automatic ME analysis algorithm utilizing the Flownet 2.0 (Alexey et al., 2015). With the benefit of Flownet, Peng et al. (2017) improved the performance of dual-template CNN, but it is inferior to classical methods (X. Li et al., 2018). There is still considerable room for improvement in the ME recognition performance. Effective deep learning-based approaches that could counter for limited and subtle MEs should be found, and more robust ME recognition frameworks need to be explored. Our proposed approach about ME recognition is described in Section 2.4 and a constructed robust ME framework based on a spotted apex frame is introduced in Section 2.5.

### 2.3 ME apex frame spotting

In this section, we first introduce the ME apex frame spotting method. Then experimental results together with evaluation metrics are detailed. This work is originally published in Papers I and II.

#### 2.3.1 A method for ME apex frame spotting

As previously discussed in Section 2.2.1, the subtle change of ME leads to a hard locating apex frame in the spatio-temporal domain. According to our empirical experience (Y. Li, Huang, & Zhao, 2018), the frequency can clearly express the subtle but rapid pixel changes in ME sequences. Thus, we spot the apex frames in the frequency domain instead. The basic idea is to represent each ME frame with the frequency components at short intervals, and to locate the apex frame by comparing the frequencies. Figure 2 depicts the flowchart of the proposed ME apex frame spotting method.
According to Dikpal, Ramamoorthi, and Curless (2012), it is found that the frequency is sensitive to illumination variations. Prior to analyzing frequency, the gray-scale invariant LBP (W. J. Yan, Wang, Chen, Zhao, & Fu, 2014) is used to extract the texture map of the ME frame, which suppresses the influence of illumination change to frequency. Subsequently, the frequency of sequential video frames is obtained at a specified interval. For more details, the facial area is divided into equal-sized blocks (6 × 6 in the experiments). Afterwards, the video blocks are transformed into frequency domain through 3D Fast Fourier Transformation (3D-FFT) with a sliding time window. Given the sliding window of length $T$, for the $i$-th interval, the frequency values for the interval are computed on blocks by 3D-FFT. The frequency value of the $j$-th block in the $i$-th interval is obtained as

$$F_{b_{ij}}(u, v, q) = \int_{-\frac{L_b}{2}}^{\frac{L_b}{2}} \int_{-\frac{W_b}{2}}^{\frac{W_b}{2}} f_{b_{ij}}(x, y, z) \times e^{j2\pi(ux+vy+qz)} dx dy dz,$$  

(5)

where $(u, v, q)$ represents the position in the frequency domain; $L_b$ and $W_b$ represent the height and width of the $j$-th block $b_{ij}$ in the $i$-th interval, respectively and $j = \{1, 2, \ldots, 36\}$.

Based on the observation of Y. Li et al. (2018), the apex frame with rapid pixel change is related to the higher frequency. On the other hand, MEs with subtle changes contain useless low-frequency information. Thus, a high-band frequency (HBF) filter is antecedently used to filter the higher frequency and reduce the influence of unchanging pixels in the frames. The HBF filter $H_{b_{ij}}$ is defined as

$$H_{b_{ij}}(u, v, q) = \begin{cases} 
1 & \text{if } \sqrt{u^2 + v^2 + q^2} \geq D_0 \\
0 & \text{if } \sqrt{u^2 + v^2 + q^2} < D_0 
\end{cases},$$  

(6)

where $D_0$ is the threshold.

The proposed 3DF-N obtains the high-frequency components of the $j$-th block in the $i$-th interval according to Equation 3,

$$G_{b_{ij}}(u, v, q) = F_{b_{ij}}(u, v, q) \times H_{b_{ij}}(u, v, q).$$  

(7)

Due to sparse facial changes caused by MEs, the occurrence of apex frame leads to higher frequency in some specific blocks. To reduce redundancy information, 3DF-N uses the specific blocks with the $N$ largest frequency values, and then sums up the high-frequency value $G_{b_{ij}}$ in the $i$-th video interval by the following formulation,
\[ A_i = \sum_{j=1}^{N} \sum_{u=1}^{T} \sum_{v=1}^{L} \sum_{w=1}^{W} | G_{b_i}(u,v,q) |, \]

where \( A_i \) represents the frequency amplitude of the \( i \)-th interval. \( A_j \) indicates the range of rapid facial movements at the \( i \)-th interval. In the same way, 3DF-N can obtain frequency information of all the video intervals. The interval with maximum amplitude indicates the frames with the most obvious facial movement, which is defined as

\[ A_{pi} = \max(A_i), \]

where \( A_{pi} \) represents the interval with the most rapid facial movements. The middle of the interval can be viewed as the apex frame.

### 2.3.2 Experiments

#### Evaluation metrics

The Normalized Mean Absolute Error (NMAE) and Normalized Standard Error (NSE) are chosen to report the effectiveness of the apex frame spotting method.

NMAE is the average normalized frame distance between the spotted apex frame and the ground-truth,

\[ NMAE = \frac{1}{K} \sum_{i=1}^{K} e_i', \]

\[ e_i' = \frac{|e_i|}{\text{len}}, \]

where \( e_i \) is the frame distance between the spotted apex frame and the ground-truth apex frame of the \( i \)-th sample. \( \text{len} \) is the average length of the samples in the database and \( K \) is the number of samples in the databases.

NSE represents the standard deviation of the sample mean distribution,

\[ NSE = \sqrt{\frac{(e_i' - \overline{e_i'})^2}{K}}, \]

where \( \overline{e_i} \) is the average of \( e_i' \).

#### The influence of specific blocks with the largest frequency values

In order to see the influence of specific blocks with the largest frequency values represented by \( N \), 3DF-N methods are evaluated on different \( N \) blocks, in which \( N \) is
blocks correspond to the first $N$ largest frequency amplitudes. The results with various $N$s are illustrated in Figure 3. 3DF-N consistently improves the 3DF-36 when the blocks with lower frequency amplitude are abandoned. It is concluded that the high-frequency signal contributes more valuable information to apex frame spotting. As seen in Figure 3, when $N$ is 28, 14, and 23 for CASME, CASME II, and SAMM, respectively, 3DF-N achieves the best performance by considering NMSE and NSE jointly. Although, the NSE on SAMM is not the lowest when $N = 23$, it slightly decreases the performance by 0.0003 compared to when $N = 14$. The difference of $N$ for the three databases is likely caused by the different properties of the databases including the recording rates and image resolution.

Performance evaluation for apex frame spotting

Tables 2 and 3 report the comparative results in terms of NMAE and NSE, respectively. The 2DF method computes the frequency in the X-T and Y-T dimensions, and then sums the frequency magnitudes in X-T and Y-T dimensions up to represent the final change rate. 3DF-36 and 3DF-N represent the proposed apex frame spotting method based on all 36 blocks and maximum $N$ blocks, respectively. OS-N computes optical strain on maximum $N$ blocks.

As shown in Table 2, the proposed 3DF-N consistently outperforms LBP (W. J. Yan, Wang, et al., 2014) by 0.2439, 0.0566, and 0.3011 in terms of NMAE on CASME, CASME II and SAMM, respectively. 3DF-N improves the OS-N consistently with gains of 0.1014, 0.0207, and 0.1414 in terms of NMAE on CASME, CASME II, and SAMM, respectively. The increasing results demonstrate that our proposed 3DF-N outperforms optical flow on ME apex frame spotting by a large margin. The results
Table 2. The NMAE (The less the better) of apex frame spotting, where 2DF, 3DF-36 and 3DF-N are the proposed methods for apex frame spotting. Reprinted with permission, Paper II © 2020 IEEE.

<table>
<thead>
<tr>
<th>Database</th>
<th>CASME</th>
<th>CASME II</th>
<th>SAMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP (W. J. Yan, Wang, et al., 2014)</td>
<td>0.3462</td>
<td>0.2037</td>
<td>0.4364</td>
</tr>
<tr>
<td>OS-ROI (Liong et al., 2015)</td>
<td>0.1824</td>
<td>0.1964</td>
<td>0.2550</td>
</tr>
<tr>
<td>RHOOF (Ma et al., 2017)</td>
<td>0.1644</td>
<td>0.1656</td>
<td>N/A</td>
</tr>
<tr>
<td>OS-N</td>
<td>0.2037</td>
<td>0.1678</td>
<td>0.2767</td>
</tr>
<tr>
<td>2DF</td>
<td>0.1399</td>
<td>0.1954</td>
<td>0.1567</td>
</tr>
<tr>
<td>3DF-36</td>
<td>0.1089</td>
<td>0.1687</td>
<td>0.1412</td>
</tr>
<tr>
<td>3DF-N</td>
<td>0.1023</td>
<td>0.1471</td>
<td>0.1353</td>
</tr>
</tbody>
</table>

*N/A - no results reported.

Table 3. The NSE (The less the better) of apex frame spotting, where 2DF, 3DF-36 and 3DF-N are the proposed methods for apex frame spotting. Reprinted with permission, Paper II © 2020 IEEE.

<table>
<thead>
<tr>
<th>Database</th>
<th>CASME</th>
<th>CASME II</th>
<th>SAMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP (W. J. Yan, Wang, et al., 2014)</td>
<td>0.0223</td>
<td>0.0158</td>
<td>0.0197</td>
</tr>
<tr>
<td>OS-ROI (Liong et al., 2015)</td>
<td>0.0100</td>
<td>0.0118</td>
<td>0.0156</td>
</tr>
<tr>
<td>RHOOF (Ma et al., 2017)</td>
<td>0.0110</td>
<td>0.0159</td>
<td>N/A</td>
</tr>
<tr>
<td>OS-N</td>
<td>0.0147</td>
<td>0.0094</td>
<td>0.0178</td>
</tr>
<tr>
<td>2DF</td>
<td>0.0137</td>
<td>0.0119</td>
<td>0.0108</td>
</tr>
<tr>
<td>3DF-36</td>
<td>0.0094</td>
<td>0.0116</td>
<td>0.0111</td>
</tr>
<tr>
<td>3DF-N</td>
<td>0.0085</td>
<td>0.0080</td>
<td>0.0107</td>
</tr>
</tbody>
</table>

*N/A - no results reported.
indirectly indicate that it is reasonable to spot apex frames in the frequency domain. Furthermore, it can be found that 3DF-36 and 3DF-N work better than 2DF. The results show that both spatial and temporal changes make contributions to apex frame spotting. Moreover, 3DF-N outperforms 3DF-36 by 0.0066, 0.0216, and 0.0059 in terms of NMAE on CASME, CASME II and SAMM, respectively. It shows that reducing the redundant information can improve apex frame performance. By comparison to the state-of-the-art methods (Liong et al., 2015; Ma et al., 2017), 3DF-N achieves the best performance on all three databases. Compared to RHOOF (Ma et al., 2017) based on optical flow histogram on ROIs, 3DF-N improves the spotting performances on the CASME and CASME II databases by 37.78% and 11.17% in terms of NMAE, respectively. In Table 3, compared to LBP, 3DF-N reduces the NSE by 0.0138, 0.0078, and 0.0090 in the CASME, CASME II, and SAMM databases, respectively. 3DF-N achieves the best robustness compared to LBP, OS-ROI (Liong et al., 2015) and RHOOF (Ma et al., 2017).

2.4 ME recognition via joint local and global information learning

This section introduces a joint learning framework for ME recognition. The empirical experience and quantitative analysis in the work of X. Huang, Zhao, Zheng, and Pietikäinen (2012) show that eyeglasses have a seriously negative influence on the performance of ME recognition. It is of importance to suppress the influence of outliers such as eyeglasses. On the other hand, only specific AUs are triggered when facial expression occurs (Z. Wang & Peng, 2019). Compared to the motionless regions, the local region related to AUs may contribute more information to ME recognition. A novel local and global learning framework called LGCcon is proposed to emphasize local informative region learning among global information for robust ME recognition. The performance of our proposed LGCcon is compared to the state-of-the-art methods in the last part of this section.

2.4.1 A joint local and global information learning method

For the majority of MEs, not all facial regions contribute to ME recognition. In order to emphasize emotion learning from informative regions and reduce the influence of outliers, the proposed LGCcon discovers that local facial regions contribute ME information and learns the local and global facial information jointly to increase the discrimination and robustness of features against the problem of outliers. Besides, multi-constraints on local and global information learning are developed to raise the
Fig. 4. The proposed ME recognition framework LGCon. GI and LI paths extract global features of the whole face and local features of sub-regions, respectively. Reprinted with permission, Paper II © 2020 IEEE.

discrimination of local and global representations, respectively. Furthermore, Centerloss (Wen, Zhang, Li, & Qiao, 2016) is employed to enhance inter-class dispersion and intra-class compactness for ME recognition.

Figure 4 illustrates the framework of LGCon. The backbone of LGCon is based on VGG-16 CNN architecture (Parkhi, Vedaldi, & Zisserman, 2015). LGCon consists of the Global Information path (GI) and Local Information path (LI), which extract global and local features, respectively. Specifically, the GI aims to extract contextual features from the whole facial image. Meanwhile, the LI aims to extract features from the local region which contributes the most ME information. Below, the details of the LI and GI in LGCon are presented.

As Figure 4 shows, given a facial image \( I \), it passes through 16 convolutional layers and three fully connected (FC) layers. The feature of the last FC layer is represented as \( \phi_G \). For the GI, the score of ME based on the whole face is defined as \( S_G \)

\[
S_G(\theta; I) = w^\theta_G \cdot \phi_G(I),
\]

where \( \phi_G \) is the feature extracted from the whole facial region \( I \). The dimension of \( \phi_G \) corresponds to the number of ME categories. \( w^\theta_G \) is the global weight for ME category \( \theta \).

Given the score \( S_G(\theta; I) \) for ME based on global information, the Softmax function is used to compute the probability \( p_G \)

\[
p_G(\theta; I) = \frac{\exp(S_G(\theta; I))}{\sum_{\theta \in E} \exp(S_G(\theta; I))}.
\]
Thus, the loss function for the GI path is defined as

\[ L_G = -\frac{1}{M} \sum_{i=1}^{M} \log(p_G(\theta = l_i|I_i)), \quad (15) \]

where \( M \) represents the batch size and \( l_i \) represents the true label of image \( I_i \).

On the other hand, as seen in Figure 4, the LI is proposed to extract the information on the local regions containing ME emotion (e.g., cheek raiser). For the sake of simplicity, \( \tau \) is defined as a region in \( I \), \( R(\tau; I) \) is the set of candidates for the sub-regions in the whole set of regions in \( I \). As the facial structure is fixed and symmetrical, LGCcon obtains \( R(\tau; I) \) by a sliding window with the height being one-third of the face height and the face width. The step size of the sliding window is one-sixth of the facial height, and six ROIs are obtained in one facial image. Then, the ROI pooling layer is used to extract local features \( \phi_L \) for all of the \( R(\tau; I) \).

Inspired by multiple instance learning (MIL) (Gkioxari, Girshick, & Malik, 2015) resolving the problem of inaccurate annotations through data in the form of bags with positive or negative labels (Maron & Lozano-Pérez, 1998), LI regarded the set of candidate sub-regions \( R(\tau; I) \) as a ‘bag’ of instances in ME recognition. For each ME image, at least one local region contributes emotional information to ME recognition. The most informative region can be seen as the positive instance for the corresponding ME category. The LI path recognizes MEs based on the positive instance which contributes the most emotional information through a maximum operation. Therefore, the score and probability of MEs based on the local information are defined in Equations 16 and 17, respectively.

\[ S_L(\theta; \tau, I) = \max_{\tau \in R(\tau; I)} w^\theta_L \cdot \phi_L(\tau; I), \quad (16) \]

where \( \phi_L \) is the feature extracted from local face regions \( R(\tau; I) \). The dimension of \( \phi_L \) is the number of ME categories. \( w^\theta_L \) is the local weight for ME category \( \theta \).

\[ p_L(\theta; \tau, I) = \frac{\exp(S_L(\theta; \tau, I))}{\sum_{\theta \in \Theta} \exp(S_L(\theta; \tau, I))}, \quad (17) \]

where \( S_L \) and \( p_L \) represent the score and probability of MEs based on local information, respectively.

Based on \( p_L(\theta; \tau, I) \), the loss function for the LI path is defined as \( L_L \). That is

\[ L_L = -\frac{1}{M} \sum_{i=1}^{M} \log(p_L(\theta = l_i|\tau_i, I_i)). \quad (18) \]

where \( l_i \) is the true label of ROI \( \tau_i \) in facial image \( I_i \).
Finally, the scores based on the global information and local information learning are combined to jointly estimate the final ME probability.

\[
S_{LG}(\theta; \tau, I) = S_G(\theta; I) + S_L(\theta; \tau, I), \quad (19)
\]

\[
p_{LG}(\theta; \tau, I) = \frac{\exp(score_{LG}(\theta; \tau, I))}{\sum_{\theta \in E} \exp(score_{LG}(\theta; \tau, I))}, \quad (20)
\]

where \( S_{LG} \) and \( p_{LG} \) represent the joint score and probability of MEs, respectively. Specifically, the feature representations \( \phi_L \) and \( \phi_G \), and the weight vectors \( w^G_\theta \) and \( w^L_\theta \) in Equations 13 and 16 are learned jointly for all ME categories. The loss function of the joint local and global information learning is represented as \( L_{LG} \),

\[
L_{LG} = -\frac{1}{M} \sum_{i=1}^{M} \log(p_{LG}(\theta = l_i|\tau_i, I_i)). \quad (21)
\]

However, based on the previously described framework, the features are not sufficiently discriminative. Due to small ME databases, the possible training identities are very limited and not diversified. For enhancing the discrimination ability of ME features, Centerloss (Wen et al., 2016) is employed to strengthen inter-class dispersion and intra-class compactness. Centerloss is defined as

\[
L_C = \frac{1}{2} \sum_{i=1}^{M} \left\| x_i - c_\theta \right\|^2_2, \quad (22)
\]

where \( x_i \) represents the sample in the class, while the \( c_\theta \) represents the center of samples belonging to ME class \( \theta_i \).

During the training process, the \( L_L \) and \( L_{LG} \) are also used as the constraints to restrict the learning procedure based on local and global information, respectively. They aim to promote the discrimination ability of local and global representations. Therefore, the final loss function \( L \) is formulated as

\[
L = L_{LG} + \lambda_L \cdot L_L + \lambda_L \cdot L_{LG}, \quad (23)
\]

where \( \lambda_C \), \( \lambda_L \), and \( \lambda_G \) balance the loss functions. \( \lambda_C \) is set to 0.008. \( \lambda_L \) and \( \lambda_G \) are set to 0.7 for faster training convergence.

### 2.4.2 Model training

LGCcon is built based on VGG and R*CNN (C. Zhang, Platt, & Viola, 2006) and fine-tuned on the VGG-FACE model (Parkhi et al., 2015). In the training stage, the
Table 4. Ablation study on micro-expression recognition accuracy (%) and F1 score of the proposed methods. Reprinted with permission, Paper II © 2020 IEEE.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CASME ACC</th>
<th>CASME II ACC</th>
<th>SAMM ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG</td>
<td>44.44</td>
<td>61.73</td>
<td>34.71</td>
</tr>
<tr>
<td>LGcon</td>
<td>53.21</td>
<td>63.79</td>
<td>33.38</td>
</tr>
<tr>
<td>LGC</td>
<td>48.54</td>
<td>61.08</td>
<td>35.52</td>
</tr>
<tr>
<td>LGCcon</td>
<td>60.82</td>
<td>65.02</td>
<td>40.90</td>
</tr>
</tbody>
</table>

losses $L_{LG}$, $L_{C}$, $L_{L}$, and $L_{G}$ are trained jointly. The learning rate is set to 0.00001 and the batch size is 64. To avoid over-fitting, the dropout rate is set to 0.8.

As the MEs have a low intensity and are difficult to recognize, the apex frames are magnified to train the ME classifier. The Eulerian magnification method (H. Wu, Shih, Shih, Guttag, & Freeman, 2012) is used to magnify the subtle motion of apex frames. Here, it enlarges the difference between different ME categories for enhancing the performance of recognition. The level of motion magnification is set to 30 in our framework. In addition, due to the small sampling size of the ME database, the new data augmentation strategy is exploited to train a good model. Although the ME is rapid, the neighboring five frames to the apex frame are very similar to the apex frame, especially the magnified one. The apex frame and the two frames before and after the apex frame are chosen for training, such that the ME database is augmented five times. For the sake of simplicity, the extended database is named the Extended Magnified ME (EMME) database.

### 2.4.3 Experiments

This section reports the results of ME recognition on the CASME, CASME II, and SAMM databases. In the experiments, the leave-one-subject-out cross validation (LOSO) protocol is used. The recognition accuracy and F1 score are used as performance metrics.

**Ablation study**

To reveal the contribution of each module, the accuracy and F1 score of LGCcon with different configurations are evaluated. Table 4 reports the comparison results. The proposed backbone LG obtains 0.50, 0.62, and 0.22 in terms of F1 score on the CASME, CASME II, and SAMM databases, respectively.
Table 5. Micro-expression recognition comparisons of accuracy (%) on the subjects with and without eyeglasses. NG represents the subjects without eyeglasses and WG represents the subjects with eyeglasses. Reprinted with permission, Paper II © 2020 IEEE.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CASME</th>
<th>CASME II</th>
<th>SAMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WG</td>
<td>NG</td>
<td>WG</td>
</tr>
<tr>
<td>VGGMag</td>
<td>62.96</td>
<td>56.45</td>
<td>61.50</td>
</tr>
<tr>
<td>LGCcon</td>
<td>63.88</td>
<td>56.45</td>
<td>64.40</td>
</tr>
</tbody>
</table>

(1) LGCcon includes constraints $L_L$ and $L_G$ on local and global information learning. The experiments show that LGCcon increases the F1 score by 0.03 on average on all three databases in comparison to LG. It indicates that constraints on local and global information learning can control the learning and lead to a better fusion result.

(2) The LGC consists of the basic LG with the Centerloss $L_C$. As seen in Table 4, LGC performs better than basic LG on the CASME and SAMM databases. LGC also achieves comparable results on the CASME II database. Based on the results, it is inferred that the Centerloss improves the discriminative ability of the ME feature.

(3) Our framework LGCcon is designed by combining Centerloss and constraints with local and global information learning. LGCcon achieves an accuracy of 60.82%, 65.02%, and 40.90% on the CASME, CASME II, and SAMM databases, respectively. Compared to LG, the addition of $L_G$, $L_L$, and $L_C$ losses improves the recognition accuracy by 16.38%, 3.29%, and 6.19% on the CASME, CASME II, and SAMM databases, respectively. It validates the effectiveness of multi-constraints and Centerloss.

*The effectiveness with outliers*

In order to evaluate the effectiveness of LGCcon with outliers, LGCcon is further studied on participants with eyeglasses. Table 5 reports the comparisons between LGCcon and VGGMag (Y. Li et al., 2018) on subjects with and without eyeglasses. The local information on ME recognition is not considered in VGGMag. In Table 5, it is seen that LGCcon outperforms VGGMag with 2.9% when they recognize the ME of the subjects with eyeglasses on the CASME II database. It narrows the performance gap between subjects with and without eyeglasses by 2.02%. Moreover, LGCcon outperforms VGGMag by 0.92% and 0.87% on subjects with eyeglasses on the CASME and SAMM databases, respectively. The results indicate that joint learning local and global information not only improves the discrimination of the ME feature, but also reduces the influence of outliers to some extent.
Performance comparisons with different frames

In order to validate the importance of the apex frame, a comparison of ME recognition performances based on the apex frame with the other frames is conducted. One frame from the ME clip between the onset and apex frames is randomly selected. Figure 5 shows ME recognition accuracy with a function of training frames on the CASME database. When the selected frame is closer to the apex frame, LBP and LGCon gain improvements by around 10% in terms of accuracy. These results indicate that the apex frame in ME clips contributes more important information to ME recognition, compared with the other frames.

Comparison with the state-of-the-art algorithms

This subsection compares the proposed LGCon with state-of-the-art methods LBP-SIP (Y. Wang et al., 2014a), FHOFO (Happy & Routray, 2017), STCLQP (X. Huang et al., 2016), HIGOMag (X. Li et al., 2018), 3D-FCNN (J. Li et al., 2018) Bi-WOOF (Liong et al., 2018), Selective (Patel et al., 2017), TIM-DCNN (Mayya et al., 2016), CNNLSTM (Kim et al., 2016), VGGMag (Y. Li et al., 2018), STRCN-A and STRCN-G (Z. Xia, Hong, Gao, Feng, & Zhao, 2019) and TSCNN (Song et al., 2019). The baseline is LBP-TOP (X. Li et al., 2013). Table 6 summarizes the compared results. It is seen that LGCon surpasses the existing deep learning methods based on whole ME sequence (Kim et al., 2016; Patel et al., 2017). Besides, LGCon achieves promising results compared to the hand-crafted methods using ME sequences (Happy & Routray,
Table 6. Micro-expression recognition accuracy and F1 score of the proposed methods and the state-of-the-art methods. Reprinted with permission, Paper II © 2020 IEEE.

<table>
<thead>
<tr>
<th>Methods</th>
<th>CASME ACC</th>
<th>CASME F1</th>
<th>CASME II ACC</th>
<th>CASME II F1</th>
<th>SAMM ACC</th>
<th>SAMM F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>40.35</td>
<td>0.26</td>
<td>40.65</td>
<td>0.33</td>
<td>34.56</td>
<td>0.25</td>
</tr>
<tr>
<td>LBP-SIP</td>
<td>36.84</td>
<td>0.33</td>
<td>46.56</td>
<td>0.45</td>
<td>36.76</td>
<td>0.21</td>
</tr>
<tr>
<td>FHOF0</td>
<td>65.99</td>
<td>0.54</td>
<td>55.86</td>
<td>0.52</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>STCLQP</td>
<td>57.31</td>
<td>0.50</td>
<td>58.39</td>
<td>0.58</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Bi-WOOF</td>
<td>N/A</td>
<td>N/A</td>
<td>59.67</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HiGOMag</td>
<td>N/A</td>
<td>N/A</td>
<td>67.21</td>
<td>N/A</td>
<td>41.91</td>
<td>N/A</td>
</tr>
<tr>
<td>3D-FCNN†</td>
<td>54.44</td>
<td>N/A</td>
<td>59.11</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>TIM-DCNN†</td>
<td>N/A</td>
<td>N/A</td>
<td>64.90</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>VGGMag</td>
<td>60.23</td>
<td>0.58</td>
<td>63.21</td>
<td>0.59</td>
<td>36.00</td>
<td>0.25</td>
</tr>
<tr>
<td>CNNLSTM</td>
<td>N/A</td>
<td>N/A</td>
<td>60.96</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Selective</td>
<td>N/A</td>
<td>N/A</td>
<td>47.30</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>STRCN-A ‡</td>
<td>40.93</td>
<td>0.35</td>
<td>45.26</td>
<td>0.38</td>
<td>32.85</td>
<td>0.24</td>
</tr>
<tr>
<td>STRCN-G ‡</td>
<td>59.65</td>
<td>0.57</td>
<td>63.37</td>
<td>0.62</td>
<td>53.48</td>
<td>0.36</td>
</tr>
<tr>
<td>TSCNN</td>
<td>73.88</td>
<td>0.72</td>
<td>80.97</td>
<td>0.81</td>
<td>71.76</td>
<td>0.69</td>
</tr>
<tr>
<td>LGCon</td>
<td>60.82</td>
<td>0.60</td>
<td>65.02</td>
<td>0.64</td>
<td>40.90</td>
<td>0.34</td>
</tr>
</tbody>
</table>

*N/A - no results reported.
† employing LOVO which leaves one video out.
‡ re-implemented in PyTorch with the same training protocol and data augmentation as LGCon.

Moreover, TSCNN (Song et al., 2019) utilized the dynamic temporal information of optical flow between onset, apex, and offset frames, and the static spatial information of apex frames. Although the accuracy of LGCon is a bit lower than TSCNN, LGCon, based on only static apex frame information, can deal with the situation when the onset frame, offset frame, and temporal information are missing. Even though many methods (Song et al., 2019; Z. Xia et al., 2019) indicated that optical flow-based methods always outperform the appearance-based methods, our proposed LGCon, which is considered an appearance-based method, achieves competitive performance compared with optical flow-based methods. The results further verify the effectiveness of LGCon.

2.5 A framework for ME recognition with detected apex frame

In this section, we propose a complete ME recognition system which recognizes the ME category with the spotted apex frame. The flow of the proposed ME recognition
system is shown in Figure 6. Given a ME clip as the input, the system first processes the video to locate the apex frame with the 3DF-N apex spotting method described in Section 2.3. Then the detected apex frame is inputted to the LGCcon network introduced in Section 2.4 to recognize the MEs.

Table 7 reports the results of LGCcon with the detected apex frame based on 3DF-N, namely LGCconD. Compared to LGCcon, LGCconD slightly degrades the performance by 0.06, 0.04 and 0.11 in terms of F1 score on the CASME, CASME II, and SAMM databases, respectively, which suggests that our proposed apex frame spotting method is reliable.

### 2.6 Summary

This chapter focuses on the ME recognition study. The state-of-the-art progress about ME studies are firstly reviewed, including the recent studies of ME spotting and recognition on spontaneous ME databases. Then, we introduce the three parts of our works, including (1) a ME apex frame spotting method in frequency domain; (2) a ME recognition approach based on deep learning with apex frame; and (3) a ME recognition system which recognizes MEs with the detected apex frame.

Recently, with the development of deep learning, it has achieved considerable performance in various fields including facial expression recognition. However, the ME recognition based on deep learning is still relatively poor, due to the fact that
ME databases are very small and the changes in MEs are subtle. Paper I studies the contribution of apex frame in MEs with deep learning. Considering the subtle and rapid muscle movements of MEs, a new method to locate the apex frame accurately through analyzing MEs in frequency domain is proposed. Then, the MEs are recognized with deep model only based on the magnified apex frame. The experimental results demonstrate that our proposed method is effective compared to the state-of-the-art methods and the only apex frame can work well for ME recognition.

Paper II is an extension version of Paper I. The work in Paper II is substantially extended in five aspects: (1) The apex frame spotting method is further improved through reducing redundancy information, which is achieved by locating regions with large change rates; (2) To suppress the influence of outliers and motionless regions for ME recognition, a MIL-based method is proposed to automatically detect the most important information on the face; (3) To further gain discriminative representation ability, a local maximum and global context joint learning framework is designed to adaptively embed local and global information; (4) A robust ME recognition framework with detected apex frame is achieved; and (5) Intensive experiments demonstrate the effectiveness and generalizability of LGCon.

Inspired by the local region contribution for ME recognition, we plan to continue ME research on fine-grained level with facial action unit analysis which describes all discernible facial movements. The detailed information is introduced in Section 4.
3 Micro-expression action unit detection

3.1 Introduction

3.1.1 Facial action coding system and action units

Facial expressions are one of the major ways that humans convey emotions and play an important role in human’s life. Various studies have been conducted on automatic facial expression analysis considering its practical importance in medical treatment, sociable robots, and many other human-computer interaction systems. As early as the twentieth century, Ekman and Friesen (1971b) proposed the definition of six basic emotions including anger, fear, sadness, disgust, surprise and happiness, based on the assumption of the universality of human emotion display. Later, Lazarus and Lazarus (1994) and Cowen and Keltner (2017) extended the primary six emotions to 15 and 27 specific emotions, respectively, such as depression, fatigue, joy, relief, and pain. Currently, facial expressions of the discrete emotions are most commonly studied due to the simplicity of the representation. However, the discrete emotions cannot explain the full range of facial expressions and advanced research on neuroscience and psychology argue that the discrete emotions are not culturally universal (Jack, Garrod, Yu, Caldara, & Schyns, 2012).

Besides discrete emotions, another effective emotion description model is based on the Facial Action Coding System (FACS) (E. Friesen & Ekman, 1978). Different from the discrete emotions, the FACS is a comprehensive, objective, anatomically based system for describing all visually facial movements. The FACS specifies 32 atomic facial muscle actions, named Action Units (AUs), 14 action descriptors referring to head pose and gaze direction, and miscellaneous actions such as blow, bite and jaw thrust.

According to FACS, facial behavior can be interpreted by a set of AUs. The specific information of common AUs is shown in Table 8. The FACS indicates that successful AU detection greatly facilitates the analysis of the complicated facial actions or expressions (W. Li, Abtahi, & Zhu, 2017). In other words, it is very essential to explore AUs for deeply interpreting the facial behavior of expressions. Currently, AU detection has played an indispensable role in analyzing macro-expression (Michiel, 1982; K. Zhao, Chu, & Zhang, 2016). To the best of our knowledge, few work is conducted on analyzing AUs for MEs due to the ME challenges, which is specifically discussed in the Subsection 3.1.2.
Table 8. Action units defined in FACS.

<table>
<thead>
<tr>
<th>AU</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU1</td>
<td>Inner Brow Raise</td>
</tr>
<tr>
<td>AU2</td>
<td>Outer Brow Raiser</td>
</tr>
<tr>
<td>AU4</td>
<td>Brow Lowerer</td>
</tr>
<tr>
<td>AU5</td>
<td>Upper Lid Raise</td>
</tr>
<tr>
<td>AU6</td>
<td>Cheek Raise</td>
</tr>
<tr>
<td>AU7</td>
<td>Lids Tight</td>
</tr>
<tr>
<td>AU9</td>
<td>Nose Wrinkle</td>
</tr>
<tr>
<td>AU10</td>
<td>Upper Lip Raiser</td>
</tr>
<tr>
<td>AU12</td>
<td>Lip Corner Puller</td>
</tr>
<tr>
<td>AU14</td>
<td>Dimpler</td>
</tr>
<tr>
<td>AU15</td>
<td>Lip Corner Depressor</td>
</tr>
<tr>
<td>AU17</td>
<td>Chin Raiser</td>
</tr>
<tr>
<td>AU25</td>
<td>Lips Part</td>
</tr>
<tr>
<td>AU26</td>
<td>Jaw Drop</td>
</tr>
</tbody>
</table>

3.1.2 Challenges of ME-AU detection

Compared to macro-expression AU detection, ME-AU detection becomes more difficult. It may be explained by as followed: Firstly, ME-AU detection suffers from much lower intensity and shorter duration of AU occurrence. Secondly, compared to facial AU databases such as BP4D database (X. Zhang et al., 2014) (328 videos and about 140,000 frames in total), ME database contains very small number of samples. Lastly, few AUs co-exist in MEs. In fact, the ME-AU correlations are very weak, comparing with macro-expressions. Following the research of K. Zhao, Chu, la Torre, Cohn, and Zhang (2016), we have analyzed the cross-correlation between AUs in micro- and macro-expressions in Figure 7. Specifically, K. Zhao, Chu, la Torre, et al. (2016) defined AUs with over moderate positive correlations (correlation coefficient $≥ 0.40$) as positive correlations. As seen in Figure 7, only one pair of AUs (AU1 and AU2) has positive correlation in the ME database. Different from macro-expressions, the AU correlation makes a small contribution to ME-AU detection.

In this chapter, we explore the robust ME-AU detection methods. Firstly, Section 3.2 reviews previous AU detection methods. In Section 3.3, the work in Paper III is introduced which proposes a ME-AU detection method with spatial and channel attention. Section 3.4 presents the work in Paper IV which alleviates the small-scale ME
Fig. 7. Illustration of relation matrix of AU labels. The lower left matrix represents the AU cross-correlation in CASME II database, while the upper right one represents the macro-expression AU cross-correlation studied on more than 350,000 valid frames with AU labels in CK+ (Lucey et al., 2010), GFT (Sayette et al., 2012) and BP4D databases (X. Zhang et al., 2014). Red and yellow rectangles indicate the positive correlations between AUs in micro- and macro-expression database, respectively. Reprinted with permission, Paper III © 2021 Elsevier.

3.2 Related work

Research concerning AU detection on macro-expressions are briefly reviewed in this section. The traditional AU detection approaches are described in subsection 3.2.1 and the deep learning-based AU detection approaches are introduced in subsection 3.2.2.

3.2.1 Traditional approaches

AU detection with traditional approaches could be categorized into appearance feature-based approaches, geometry-based approaches and hybrid feature-based approaches.

Appearance features describe the colour and texture of a facial region. The appearance feature-based AU detection approaches focus on extracting global features from the whole facial images overall or local features from regions of interest on face. As different facial regions have diverse contributions to AUs analysis, the local features are
widely used for AU detection. To date, many effective appearance features have been proposed, such as spatial filter representation (Gianluca, Stewart, C., Paul, & J., 1999; Jacob & W, 2006), histogram representation (Albert, Bir, & Songfan, 2011; Ellis & B, 1982), and data-driven representation (Quan, 2013). However, appearance features are sensitive to non-frontal head poses and illumination changes.

Geometry features describe the geometric information of the faces and effectively represent the various local facial changes (Cohn, Zlochower, Lien, & Kanade, 1999; Tian, Kanade, & Cohn, 2001). The geometry feature-based AU detection is based on the deformation of facial components or the location of landmarks between facial images (Kotsia & Pitas, 2006; Pantic & Bartlett, 2007). In this way, the identity bias is able to be reduced in some degree. Furthermore, compared to appearance features, the computation of geometry features is relatively simpler and the dimensionality of feature is lower. However, some AUs are only related with texture changes, e.g. the AU6 refers to skin wrinkling around eyes and cheek raise (Valstar & Pantic, 2011). This kind of AU is difficult to be identified with solely geometric representations.

Considering the strength of various kinds of features, extensive research combines different features to enhance AU detection performance, denoted as hybrid feature-based approaches (Jeremie, Kevin, & Mohamed, 2015; Jérémie, Kévin, & Mohamed, 2016). Li Tian, Kanade, and Cohn (2002) combined geometric features and Gabor wavelet features to improve the discriminative AU representation. Senechal et al. (2012) combined local gabor binary patterns features (Almaev & Valstar, 2013) with active appearance model coefficients (Matthews & Baker, 2004) and a multi-kernel SVM were utilized for automatic AU analysis.

### 3.2.2 Deep learning approaches

In recent years, deep learning methods have been extensively studied in AU detection of macro-expressions (S. Li & Deng, 2020; K. Zhao, Chu, & Zhang, 2016), due to their strong nonlinear representational power. W. Li et al. (2017) proposed a local convolutional neural network for learning AUs on cropped regions of interest centered at facial landmarks. However, such network seriously suffers from the unstable facial landmark detection. Moreover, Deep Region and Multi-label Learning method (DRML) (K. Zhao, Chu, & Zhang, 2016) was proposed to obtain important facial regions by exploiting a region layer. It extracted facial structure information to obtain promising AU detection results with subtle movements. Furthermore, W. Li, Abtahi, Zhu, and Yin (2018) proposed a local feature learning method embedding facial landmark based attention map on cropped regions. Moreover, W. Li et al. (2017) concatenated CNN
features from different facial parts to represent AUs. These works strongly suggest that local information learning can improve AU detection performance. However, all these methods focused on local regions but ignored the correlation of different facial regions. Since certain facial regions become active simultaneously, AU detection can be benefited by exploiting the correlation of local regions. Niu, Han, Yang, Huang, and Shan (2019) used Long-short-term model (LSTM) structure on facial areas to obtain the relationship of individual local face regions. It considered only first-order statistic information and needed other modules to model local regional relationship.

Different from macro-expression AUs, ME-AUs suffer from low facial intensity and small quantity, thus, it is more difficult to capture subtle regional changes. The work of Oncel, Fatih, and Peter (2006) indicated that higher-order statistics have stronger modeling capability than first-order statistics. Considering that the high-order representation can better describe local regional deformation and the latent semantic information contribute important information to AU analysis, Paper III proposes a Spatio-Channel Attention (SCA) mechanism extracting local regional changes and the relationship of local facial regions simultaneously for obtaining discriminative and robust ME-AU representation. The details of SCA is described in Section 3.3. Furthermore, aiming at solving the small-scale ME database issue, a Dual-view Attentive Similarity-Preserving Knowledge Distillation (DVASP) method is proposed in Paper IV to leverage the massive facial images, which is introduced in Section 3.4.

### 3.3 ME-AU detection with attention

This section introduces the spatio-channel attention mechanism for ME-AU detection. The overview framework is introduced in subsection 3.3.1. Then the detailed process of spatial, channel attention and the fusion of spatial and channel attention modules are introduced in subsections 3.3.2, 3.3.3, and 3.3.4, respectively. Subsection 3.3.5 describes the model training strategy. The experimental results are discussed in the subsection 3.3.6.

#### 3.3.1 The framework of ME-AU detection with attention

The ME-AUs have very low intensity and weak correlations, whereas fast changes in temporal domain. The temporal information makes an important contribution to ME-AU detection. Additionally, ResNet (He, Zhang, Ren, & Sun, 2016; Niu, Han, Yang, et al., 2019) has been demonstrated to have a strong ability for local features generation with convolutional layers. Therefore, SCA is built on 3D residual network (Res3D) (Kensho,
Hirokatsu, & Yutaka, 2018) with the temporal information, as shown in Figure 8. Before passing to a series of 3D residual blocks, the input facial images are aligned. The last 3D residual block outputs a local feature $F$ with $h \times w \times t \times c$ dimension, where $h$ and $w$ are spatial height and width, respectively, $t$ represents the temporal length, and $c$ is the number of channels.

For obtaining robust global feature, traditional CNNs directly feed the output of convolutional layers to a global average/maximum pooling layer (Oncel et al., 2006). However, this kind of operation fails to extract local regional information of structured objects such as face (Dinesh, Zhiwu, Danda, & Luc, 2018; Niu, Han, Yang, et al., 2019; Oncel et al., 2006). In contrast, a Channel Attention (CA) module is introduced to extract second-order statistics of the feature generated from 3D residual blocks. As the ME-AUs occur in few facial regions, the CA module can better capture regional changes for ME-AU detection through embedding high-order statistics on feature channels. Additionally, it is beneficial to consider the relationship of facial regions for ME-AU detection. To this end, SCA firstly designs a spatial attention (SA) module with spatial second-order statistics to automatically explore the intrinsic relationship of local facial regions. Subsequently, the spatial and channel attention representations are fused through learning to increase the discriminative ability. Overall, SCA can capture both regional changes and relationship of individual facial regions. The flowchart of SCA is shown in Figure 9.
Fig. 9. The illustration of SCA module. Given the input feature $F'$, the channel-wise $C_{ch}$ and spatio-wise $C_{sp}$ covariance matrices are computed through C-cov and S-cov, respectively. C-cov and S-cov represent channel-wise and spatio-wise covariance computation, respectively. The channel attention weight $W_{ch}$ and spatial attention weight $W_{sp}$ are produced by a linear convolution and non-linear activation, respectively. $W_{sp}$ and $W_{ch}$ are embedded on $F'$ through multiplication. Finally, the $F'_{sp}$ and $F'_{ch}$ are fused through learning. Specifically, for the CA module, the dimension reduction is implied firstly to improve the computation efficiency. Reprinted with permission, Paper III © 2021 Elsevier.

3.3.2 Channel attention module

The CA module aims to better capture regional changes of MEs through embedding second-order statistics. As shown in Figure 9, the CA module has focusing on second-order statistics along the channel dimension of feature maps and steers attention to significant channels.

The input of CA module is a feature $F$ with $h \times w \times t \times c$, which is the output of last 3D residual block. In practice, as the temporal dimension $t$ is embedded into 1D through 3D residual blocks, for simplifying the computation, $F$ is reformulated into a new feature $F'$ of $h \times w \times c$. For reducing computational cost, we further suppress the channel dimension of $F'$ to $c'$, where $c' < c$. Here, we denote the reduced dimension feature as $F'_{re}$. Apparently, $F'_{re}$ can be seen as $h \times w$ feature maps with size of $c'$. In the implementation, we reshape the $F'_{re}$ into feature maps $F'_{ch} = \{f'_{ch1}, f'_{ch2}, \ldots, f'_{chd}\}$, where $f'_{chi} \in \mathbb{R}^{c'}$ and $d = h \times w$. With $F'_{ch}$, the pairwise channel correlation $C_{ch}$ is computed as

\[
C_{ch} = \frac{1}{d-1} \sum_{i=1}^{d} (f'_{ch1} - \overline{f'_{ch}}) (f'_{ch1} - \overline{f'_{ch}})^T,
\]

where $\overline{f'_{ch}} = \frac{1}{d} \sum_{i=1}^{d} f'_{chi}$. 

\[49\]
The channel covariance correlation computation is denoted as ‘C-cov’ in Figure 9. According to the work of Tao, Cai, Zhang, Xia, and Zhang (2019), each row in \( C_{ch} \) represents the statistical dependency of the channel with all channels. In order to reserve the structural information, instead of quadratic operations involved changing the data order, the row-wise convolution is performed for \( C_{ch} \) by regarding each row as a group in group convolution (Krizhevsky, Sutskever, & Hinton, 2012). As shown in Figure 10, \( C_{ch} \) is convoluted with \( c' \) filter groups of size \( k \times 1 \times c' \) to obtain \( F'_{g} \). The filter group is represented as \( G = g_1, g_2, \ldots, g_{c'} \). Specifically, following the implementation of Gao, Xie, Wang, and Li (2019), \( k \) is set to 4.

Furthermore, we perform the second convolution on \( F'_{g} \). To completely exploit feature inter-dependencies from the aggregated information of covariance matrix, the sigmoid function is used as a nonlinear activation (Gao et al., 2019). It can obtain the channel attention map \( W_{ch} \). For each channel, its feature \( F'_{ch} \) is element-wise multiplied with \( W_{ch} \). Currently, we can obtain the ME-AU feature with channel attention. Specifically, individual channels of \( F' \) are emphasized or suppressed in a soft manner. The CA module can capture the second-order statistical dependency of the holistic image and steer attention to significant channels to improve the representation capability of ME-AUs.

3.3.3 Spatial attention module

One key aspect of SCA network is to design a spatial attention module to compute spatial pairwise feature correlations of the holistic image. As the AUs may simultaneously occur, the correlation of facial regions will make contributions to enhance the robustness of ME-AU detection. For obtaining the relationship of facial regions and steering...
attention on active AU regions for ME-AU detection, we propose a SA module which encodes nonlinear pair-wise feature dependency at all spatial positions.

For SA module, we compute pairwise correlations of feature $F'$ at all spatial positions. Given the input $F'$, which is denoted as $F'_{sp} = [f'_{sp1}, f'_{sp2}, \ldots, f'_{spC}]$, where $f'_{spi} \in \mathbb{R}^d$, the pairwise spatial correlation is computed as

$$C_{sp} = \frac{1}{C-1} \sum_{i=1}^{C} (f'_{sp_i} - \bar{f}_{sp})(f'_{sp_i} - \bar{f}_{sp})^T,$$

(25)

where $\bar{f}_{sp} = \frac{1}{C} \sum_{i=1}^{C} f'_{spi}$.

It is noted that $C_{sp} \in \mathbb{R}^{d \times d}$, where $d = h \times w$. Each row in $C_{sp}$ means statistical correlation of one feature with all features. $C_{sp}$ is also fed into row-wise convolution to extract the structural correlation representation. Subsequently, a convolution followed by sigmoid is applied to output the weight vector $W_{sp}$. Through inverse reshaping, we can obtain an $h \times w$ spatial attention map $W'_{sp}$. Lastly, the spatial attention map embeds nonlinear spatial correlations of the holistic image by multiplying with $F'$. By this means, we obtain the spatio-attention AU feature $F'_{sp}$ through taking account of the relationship of different facial regions. Traditional CNNs fail to capture the relationship of individual regions due to limited receptive field size. Different from traditional CNNs, our proposed SA module can consider holistic dependency of features at distant positions and steer attention on active AUs in MEs.

### 3.3.4 The fusion of channel and spatial attention modules

Aforementioned, the CA and SA modules independently consider statistical correlations of AU features along channel-wise and spatial-wise dimensions. It is believed that fusing them contributes to characterizing both regional facial changes and the relationship of facial regions. Moreover, fusing CA and SA modules can enhance the discriminative ability of features. Here, we learn the fusion weights $w_{sp}$ and $w_{ch}$ to fuse the outputs of CA and SA modules adaptively and steer attention on salient features. The fusion operation is defined as

$$F'_{sc} = w_{sp}F'_{sp} + w_{ch}F'_{ch},$$

(26)

where $F'_{ch}$ and $F'_{sp}$ represent the outputs of CA and SA modules, respectively.

As the previous discussion in Subsection 3.1.2 implies, the AUs in MEs co-exist and their correlation is weak. Considering them, each AU detection can be treated as a specific task. The loss for each task is defined as a binary cross-entropy loss, that is

$$L_a = y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i),$$

(27)
where $y_i$ is the ground truth for the $i$-th AU, with 1 denoting occurrence of the AU and 0 denoting no occurrence. $\hat{y}_i$ is the predicted probability of the occurrence of $i$-th AU. The overall loss is formulated as the sum of all task losses,

$$L_{\text{total}} = \sum_{i=1}^{M} L_{a_i},$$

where $M$ is number of AU categories.

### 3.3.5 Model training

In our experiments, we employ the cropped face images provided by the databases. The input is aligned RGB ME sequential images. All models were pre-trained on Kinetics (Will et al., 2017) and UCF-101 (Khurram, Zamir, & Shah, 2012) databases. We randomly crop $112 \times 112$ images to augment data. As the length of ME sequences is short and varies. The ME sequences are interpolated to 50 through Temporal Interpolation Model (TIM) (Zhou, Zhao, & Pietikäinen, 2011) and sampled to 10 by average sampling. During training, the networks are optimized using stochastic gradient descent (SGD) with a weight decay of 0.001 and momentum of 0.9. The initial learning rate is set to 0.01, divided by 10 every 30 epochs until 60 epochs, unless specified otherwise. The dropout of 0.5 is used. All implementations are based on Pytorch.

The proposed ME-AU detection are evaluated on the CASME II, CASME, and SAMM databases. The common AUs with the number of samples greater than 14 are utilized in the experiments. Following common experimental settings for AU detection Niu, Han, Yang, et al. (2019), we use subject independent four-fold cross validation. Each time two folds are used for training. The other two folds are for validation and testing, respectively.

### 3.3.6 Experiments

#### Metrics

ME-AU detection is a multi-label binary classification problem. In our evaluation, binary F1 scores are computed for eight AUs in CASME II, four AUs in CASME and four AUs in SAMM databases according to the AU samples quantity and importance. The overall performance of the algorithm is described by the average F1 score.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AU14</th>
<th>AU15</th>
<th>AU17</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP-TOP</td>
<td>0.1057</td>
<td>0.4985</td>
<td>0.7324</td>
<td>0.0635</td>
<td>0.2386</td>
<td>0.2185</td>
<td>0.0000</td>
<td>0.1667</td>
<td>0.2530</td>
</tr>
<tr>
<td>LPQ-TOP</td>
<td>0.2877</td>
<td>0.3350</td>
<td>0.6525</td>
<td>0.1359</td>
<td>0.3631</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1667</td>
<td>0.2426</td>
</tr>
<tr>
<td>LBP-SIP</td>
<td>0.2308</td>
<td>0.3892</td>
<td>0.7354</td>
<td>0.0888</td>
<td>0.2143</td>
<td>0.2979</td>
<td>0.4318</td>
<td>0.4287</td>
<td>0.3521</td>
</tr>
<tr>
<td>HOG3D</td>
<td>0.2771</td>
<td>0.2769</td>
<td>0.7012</td>
<td>0.0000</td>
<td>0.0526</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1212</td>
<td>0.1786</td>
</tr>
<tr>
<td>Res3D18</td>
<td>0.3239</td>
<td>0.2157</td>
<td>0.8302</td>
<td>0.1933</td>
<td>0.4447</td>
<td>0.2070</td>
<td>0.4684</td>
<td>0.2264</td>
<td>0.3637</td>
</tr>
<tr>
<td>Res3D50</td>
<td>0.3229</td>
<td>0.2159</td>
<td>0.7410</td>
<td>0.1341</td>
<td>0.3388</td>
<td>0.0469</td>
<td>0.1333</td>
<td>0.1961</td>
<td>0.2661</td>
</tr>
<tr>
<td>Res3D101</td>
<td>0.2608</td>
<td>0.2510</td>
<td>0.6667</td>
<td>0.2065</td>
<td>0.1383</td>
<td>0.0714</td>
<td>0.0000</td>
<td>0.1296</td>
<td>0.2156</td>
</tr>
<tr>
<td>Res3D18+SCA</td>
<td>0.2857</td>
<td>0.4532</td>
<td>0.8877</td>
<td>0.2473</td>
<td>0.4792</td>
<td>0.3327</td>
<td>0.3954</td>
<td>0.5159</td>
<td>0.4496</td>
</tr>
<tr>
<td>Res3D50+SCA</td>
<td>0.1429</td>
<td>0.4488</td>
<td>0.7801</td>
<td>0.2105</td>
<td>0.4332</td>
<td>0.1449</td>
<td>0.2296</td>
<td>0.2911</td>
<td>0.3351</td>
</tr>
<tr>
<td>Res3D101+SCA</td>
<td>0.1350</td>
<td>0.1724</td>
<td>0.7497</td>
<td>0.2036</td>
<td>0.2596</td>
<td>[0.3150]</td>
<td>0.2000</td>
<td>0.1774</td>
<td>0.2904</td>
</tr>
</tbody>
</table>
### Table 10. F1 scores on the CASME database. Reprinted with permission, Paper III © 2021 Elsevier.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU4</th>
<th>AU9</th>
<th>AU14</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP-TOP</td>
<td>0.1296</td>
<td>0.4423</td>
<td>0.0000</td>
<td>0.2333</td>
<td>0.2013</td>
</tr>
<tr>
<td>LPQ-TOP</td>
<td>0.1719</td>
<td>0.6048</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1942</td>
</tr>
<tr>
<td>LBP-SIP</td>
<td>0.1507</td>
<td>0.5235</td>
<td>0.0000</td>
<td>0.1167</td>
<td>0.1977</td>
</tr>
<tr>
<td>HOG3D</td>
<td>0.0000</td>
<td>0.5321</td>
<td>0.1111</td>
<td>0.0000</td>
<td>0.1608</td>
</tr>
<tr>
<td>Res3D18</td>
<td>0.3510</td>
<td>0.4209</td>
<td>0.1506</td>
<td>0.1592</td>
<td>0.2704</td>
</tr>
<tr>
<td>Res3D18+SCA</td>
<td>0.3818</td>
<td>0.4134</td>
<td>0.2233</td>
<td>0.2574</td>
<td>0.3189</td>
</tr>
</tbody>
</table>

### Table 11. F1 scores on the SAMM database. Reprinted with permission, Paper III © 2021 Elsevier.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP-TOP</td>
<td>0.2652</td>
<td>0.1538</td>
<td>0.4603</td>
<td>0.2376</td>
<td>0.2792</td>
</tr>
<tr>
<td>LPQ-TOP</td>
<td>0.1818</td>
<td>0.1538</td>
<td>0.4614</td>
<td>0.2376</td>
<td>0.2587</td>
</tr>
<tr>
<td>LBP-SIP</td>
<td>0.2144</td>
<td>0.0556</td>
<td>0.0400</td>
<td>0.0000</td>
<td>0.1675</td>
</tr>
<tr>
<td>HOG3D</td>
<td>0.0000</td>
<td>0.1667</td>
<td>0.2330</td>
<td>0.0833</td>
<td>0.1208</td>
</tr>
<tr>
<td>Res3D18</td>
<td>0.1726</td>
<td>0.1813</td>
<td>0.3637</td>
<td>0.3923</td>
<td>0.2775</td>
</tr>
<tr>
<td>Res3D18+SCA</td>
<td>0.3289</td>
<td>0.1297</td>
<td>0.4876</td>
<td>0.4218</td>
<td>0.3419</td>
</tr>
</tbody>
</table>

### Comparisons of methods

**Handcrafted features.** As this is the first work for ME-AU detection, we provide the baseline of ME-AU detection. Tables 9, 10, and 11 show the results on CASME II, CASME, and SAMM databases, respectively. Specifically, handcrafted features including LBP-TOP (G. Zhao & Pietikainen, 2007), LPQ-TOP (Päivärinta, Rahtu, & Heikkilä, 2011), LBP-SIP (Y. Wang, See, Phan, & Oh, 2014b), and HOG3D (Alexander, Marcin, & Cordelia, 2008) features are extracted on $5 \times 5$ blocks derived from frames. For LBP-TOP and LBP-SIP, the radii were set to $(3, 3, 3)$. One-vs-rest Linear SVM is used to train a single classifier per AU class. We test the classification penalty factor $C$ from 0.001 to 1000 on validation database and choose the best results for fair comparisons. Among the handcrafted features, the LBP-TOP achieves the best performance on CASME and SAMM databases in terms of average F1 score. From Tables 9 and 10, handcrafted features achieve relatively promising performance on AU4 (Brow lower) which refers to clear motions. However, most handcrafted features failed working on AU9 (Nose wrinkle) in MEs. The possible reason is that AU9 has blur motions and only relates with subtle appearance change on nose.
Table 12. Ablation study of F1 scores on the SAMM database. The baseline is Res3D18 and SCA$_{learn}$ is the fusion method developed in our paper. Reprinted with permission, Paper III © 2021 Elsevier.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.1726</td>
<td>0.1813</td>
<td>0.3637</td>
<td>0.3923</td>
<td>0.2775</td>
</tr>
<tr>
<td>Baseline+SA</td>
<td>0.0833</td>
<td>0.2397</td>
<td>0.4278</td>
<td>0.3726</td>
<td>0.2809</td>
</tr>
<tr>
<td>Baseline+CA</td>
<td>0.0781</td>
<td>0.1579</td>
<td>0.4660</td>
<td>0.4145</td>
<td>0.2791</td>
</tr>
<tr>
<td>Baseline+SCA$_{max}$</td>
<td>0.1765</td>
<td>0.1841</td>
<td>0.3910</td>
<td>0.4171</td>
<td>0.2913</td>
</tr>
<tr>
<td>Baseline+SCA$_{avg}$</td>
<td>0.1372</td>
<td>0.1250</td>
<td>0.4582</td>
<td>0.3764</td>
<td>0.2742</td>
</tr>
<tr>
<td>Baseline+SCA$_{concat}$</td>
<td>0.0560</td>
<td>0.0417</td>
<td>0.2817</td>
<td>0.3557</td>
<td>0.1838</td>
</tr>
<tr>
<td>Baseline+SCA$_{learn}$</td>
<td>0.3289</td>
<td>0.1297</td>
<td>0.4876</td>
<td>0.4218</td>
<td>0.3419</td>
</tr>
</tbody>
</table>

**Deep features.** The ME databases have small quantity. Although we did augmentation during training and pre-trained on large databases, the sample variety is still limited compared to common databases (X. Zhang et al., 2014). As reported in Tables 9, 10 and 11, the basic deep learning frameworks achieve comparable results to those from the handcrafted features. For example, the Res3D18 obtained 0.3637, 0.2704, and 0.2775 in terms of F1 scores on CASME II, CASME, and SAMM databases, respectively. Moreover, as seen from Table 9, Res3D18 also obtained much higher F1 scores than handcraft features on most AUs.

**Effect of the depth of network**

The same framework Res3D with different depths also have different performances. As shown in Table 9, Res3D18 outperforms Res3D50 and Res3D101 by 0.0976 and 0.1505 in terms of average F1 score on the CASME II database. In general, the AU detection performance drops when the network is further deepened. The possible reason is that the ME-AU databases have limited subjects and a very deep network may suffer from over-fitting.

**Ablation study**

In this subsection, we provide ablation study to investigate the effectiveness of each part in our SCA module. The baseline is Res3D18 exploring only first-order statics for ME-AU detection. To verify the effectiveness of SA and CA, we add SA and CA modules on Res3D18, separately.
Spatial attention. As shown in Table 12, the SA module outperforms the baseline by 0.0034 in terms of average F1 score on the SAMM database. The results demonstrate that the spatial relationship of local regions is useful for ME-AU prediction and our SA module is effective for modeling spatial correlation information.

Channel attention. The CA module also contributes to the improvement of the average F1 score by 0.0016 on SAMM, as shown in Table 12. The results indicate that second-order statistics are more representative for subtle local facial changes. From the class activation maps shown in Figure 11, we can see that CA is more likely to focus on the informative regions.

Spatio-channel attention. When both SA and CA modules are included, the network can better describe local AU regions, and consider the relationship of facial regions, and thus is able to achieve better performances. We test the spatio-channel attention fusion methods by the commonly used operations of average/maximum (SCA\textsubscript{Avg}/SCA\textsubscript{Max}) and concatenation (SCA\textsubscript{Concat}), and our proposed learning fusion weights method (SCA\textsubscript{Learn}). The results of fusion methods is presented in Table 12.
Among different fusion methods, $\text{SCA}_{\text{Learn}}$ achieves the highest average F1 score on SAMM, as shown in Table 12. $\text{SCA}_{\text{Learn}}$ outperforms the second-best fusion method ($\text{SCA}_{\text{Max}}$) by 0.0506 in terms of average F1 score on the SAMM database. The results demonstrate the effectiveness of our proposed fusion method $\text{SCA}_{\text{Learn}}$. $\text{SCA}_{\text{Learn}}$ differs from $\text{SCA}_{\text{Max}}$ in an adaptive way: $\text{SCA}_{\text{Learn}}$ learns the adaptive fusion weights to effectively fuse the complementary information between channel and spatial attentions.

Visualization

Figure 11 shows some examples of class activation maps. It is seen that the model can focus more on the accurate regions of ME-AUs by using the SA and CA modules. For example, for AU14 (3rd row of Figure 11), the baseline network focuses on the AU-unrelated region, i.e., right cheek. After adding SA module, our network can capture local regional information and concentrate on part of the mouth. $\text{SCA}_{\text{Avg}}$ only focuses on the right dimple and $\text{SCA}_{\text{Max}}$ roughly captures the AU information from the whole mouth. In contrast to them, $\text{SCA}_{\text{Learn}}$ can focus on both dimples exactly. The class activation maps further indicate the effectiveness of our fusion method $\text{SCA}_{\text{Learn}}$. Figure 12 shows a failed ME-AU detection example. It may be caused by the very small active regions of the AU. The feature dependencies are insensitive in representing very small regions.

3.4 ME-AU detection with dual-view knowledge distillation

As the discussion in Section 3.1.2 implies, the ME-AU detection is more challenging because of small-scale databases. In this section, we introduce the details of Dual-View Attentive Similarity-Preserving (DVASP) knowledge distillation to alleviate the small-scale database issue through taking advantage of the massive facial images by transfer learning. The experimental results demonstrate that our proposed knowledge distillation method can effectively distill and transfer the cross-domain knowledge for robust ME-AU detection. The original work is presented in Paper IV.
3.4.1 Overview

Compared to macro-expressions, ME-AU detection is more difficult because of the subtle facial movements and small-scale databases which are far from enough to train a robust network for AU detection. Fortunately, there are large AU databases, e.g. EmotioNet (Fabian, Srinivasan, & Martinez, 2016), including massive facial images in the wild, regarded as macro-expression. Since AUs are relatively objective encoding ways and the micro- and macro-expressions are coded by FACS, their semantic information is consistent. Obviously, there should be a strong correlation between them. Therefore, the information of macro-expression AUs can be utilized to detect ME-AUs.

Knowledge distillation (Abdolmaged, Fahad, & Irfan, 2020) has been proved as an effective approach to transfer information from pre-trained high-capacity networks to achieve faster speeds and handle the problems caused by the lack of labeled data. Inspired by the success of knowledge distillation, Sun, Cao, Li, He, and Yu (2020) utilized Fitnets (Romero et al., 2015) to improve the ME recognition performance through directly mimicking the macro-expression representations. However, the appearance of micro- and macro-expressions is different due to low intensity of MEs, directly mimicking the representation space of the network pre-trained on macro-expressions is not reasonable, because of the domain shift. Several works (Y. Li, Huang, & Zhao, 2020; Y. Liu, Du, Zheng, & Gedeon, 2019) applied subtle motion magnification (Tae-Hyun et al., 2018; H. Wu et al., 2012) to narrow down the gap between macro- and micro-expressions. However, the magnification may introduce noises and deformation, as shown in Figure 13 (a), thus there is still a domain shift between micro- and macro-expressions (Romero et al., 2015; Tung & Mori, 2019).

Although the appearances of micro- and macro-expression are different, the correlation between the samples is consistent. Recent research (Tung & Mori, 2019) found that the correlation knowledge can be utilized for cross domain transfer learning. On the other hand, as AUs correspond to specific muscular activations of the face (E. Friesen & Ekman, 1978), it is crucial to focus on the specific regions, thus attention learning is
meaningful. Inspired by above observations, instead of mimicking the representation space of the teacher network, we propose a novel Attentive Similarity-Preserving distillation (ASP) to effectively supervise the training of a student network with a pre-trained teacher referring to domain shift without extra learning burdens. In this way, the important correlation knowledge can be transferred from macro-expressions to MEs for better AU detection on small ME databases.

Besides the transferring method, obtaining a generalized teacher network to supervise the student model is very important for successful knowledge distillation. To achieve this, a semi-supervised Dual-view Co-training (DVCT) approach is developed to make full use of the massive labeled and unlabeled facial images in the wild. As illustrated in Figure 13 (b), following common co-training methods (Niu, Han, Shan, & Chen, 2019; Qiao, Shen, Zhang, Wang, & Yuille, 2018), the DVCT generates two representations by way of different models which highlight different cues for AU detection. Learning such kind of dual-view representations can leverage unlabeled facial images and obtain generalized AU representation.

Finally, our proposed Dual-View Attentive Similarity-Preserving (DVASP) transfer learning framework can effectively transfer the dual-view AU knowledge learned from generalized AU representation for robust ME-AU detection. The whole framework is consisted of a pre-trained teacher network and a student network, as shown in Figure 14. The AU knowledge is transferred from the teacher network based on two Res34 networks with DVCT to the student network based on Res18. During the transfer learning process, the teacher parameters are frozen while the student parameters are updated. On the test stage, the teacher network is removed and the ME-AU detection is based on the student network.
network. In the following sections, the details of the DVCT and ASP are described in Subsection 3.4.2 and 3.4.3, respectively.

### 3.4.2 Dual-view co-training for robust AU detection

**Teacher network:** The DVCT is introduced for obtaining a generalized teacher network for AU detection on EmotioNet (Fabian et al., 2016). DVCT is a co-training method for semi-supervised learning. The co-training approach (Blum & Mitchell, 1998) assumes that each sample in the training set has two different views $v_1$ and $v_2$, and both views are sufficient to learn effective models to represent the sample (R. Xia, Wang, Dai, & Li, 2015). That means: (1) The features from different views are conditionally independent; (2) The models trained on different views tend to have consistent predictions.

As illustrated in Figure 15, our DVCT algorithm utilizes two Res34 networks to generate the dual-view features based on the above mentioned co-training principles. Here, the large-scale facial image database, EmotionNet $D$ is introduced since there are massive facial images with AU information. We represent the database as $D = L \cup U$, $L$ and $U$ are denoted as facial images with and without AU labels, respectively. As shown in Figure 15, for each image in $D$, the dual-view features are inputted to two fully connected layers ($FC'_1$ and $FC'_2$) denoted as AU classifiers, respectively.

On the database $L$ with AU labels, considering the AUs co-exist and there is strong inference between AUs in macro-expressions (Fabian et al., 2016; Niu, Han, Shan, & Chen, 2019), the multi-label sigmoid cross entropy loss is utilized to recognize $C$ AUs in EmotioNet. Furthermore, a selective learning strategy (Hand, Castillo, & Chellappa, 2018) is adopted to handle the AU imbalance. Let $p_{ij}$ denotes the probability for the
j-th AU of i-th view, the loss function for AU detection for the i-th view is defined as

$$L'_{vi} = -\frac{1}{C} \sum_{j=1}^{C} a_c(p'_j \log(\hat{p}'_{ij}) + (1 - p'_j) \log(1 - \hat{p}'_{ij})),$$

(29)

where $a_c$ is the balancing parameter calculated in each batch (Hand et al., 2018). $p'_j$ is the ground-truth probability for the j-th AU. Specifically, $p'_j = 1$ denotes occurrence of the AU and $p'_j = 0$ absence.

For the unsupervised database $U$ without AU labels, the training is based on the co-training assumption that the two-view networks have close predictions. Therefore, we utilize a co-regularization loss to minimize the distance between the two predicted probability distributions which is measured by Jensen-Shannon divergence (Endres & Schindelin, 2003). The co-regularization loss is formulated as

$$L'_{cr} = \frac{1}{C} \sum_{j=1}^{C} \left( H(\frac{\hat{p}'_{1j} + \hat{p}'_{2j}}{2}) - \frac{H(\hat{p}'_{1j}) + H(\hat{p}'_{2j})}{2} \right),$$

(30)

where $H(\hat{p}') = -(\hat{p}' \log \hat{p}' + (1 - \hat{p}') \log(1 - \hat{p}'))$ is the entropy.

Another key condition of successful co-training is that the multi-view features should be different and provide complementary information. However, minimizing Eqs. 29 and 30 only encourages the networks to output the same predictions. Therefore, it is necessary to guarantee the networks to be conditionally independent instead of collapsing to each other. To achieve this, we orthogonalize the weights of the AU classifiers of different views through a multi-view loss $L'_{mv}$.

$$L'_{mv} = \frac{1}{C} \sum_{j=1}^{C} \| W'_{1j}^T W'_{2j} \|,$$

(31)

where $W'_{ij}$ represents classifier weights for the j-th AU of the i-th view. The final loss function of the teacher network is formulated as

$$L_{teacher} = \frac{1}{2} \sum_{i=1}^{2} L'_i + \lambda'_{mv} L'_{mv} + \lambda'_{cr} L'_{cr},$$

(32)

where $\lambda'_{mv}$ and $\lambda'_{cr}$ are hyper-parameters for balancing the losses.

**Student network:** As ME databases have small number of samples, we distill the knowledge from the pre-trained dual-view teacher network and transfer it to a relatively shallow student network for ME-AU detection. During the transfer learning process, the teacher parameters are frozen while the student parameters are updated through two parts of loss functions: ME-AU detection loss and knowledge distillation loss.

In consistent with the teacher network, the student network also employs two-view networks. Here, we choose two Res18 networks as the feature generators for ME-AU
detection. Different from macro-expressions, the AU correlation is low in ME databases (Y. Li et al., 2020) and makes little contribution to ME-AU detection. Instead of multi-label learning (W. Li et al., 2017), multiple AU detection in MEs should be viewed as multiple specific tasks based on multi-task learning framework.

For each image in the ME database, two-view features $f_1$ and $f_2$ are generated by two Res18 networks. Then, $N$ classifiers can be learned to predict the probabilities of $N$ ME-AUs using the features $f_1$ and $f_2$. Let $\hat{p}_{ij}$ denotes the probability obtained by softmax function for the $j$-th AU of $i$-th view, the loss function for AU recognition for the $i$-th view is defined as

$$L_{vi} = -p_j \log(\hat{p}_{ij}) + (1 - p_j) \log(1 - \hat{p}_{ij});$$

$$L_{vs} = \frac{1}{N} \sum_{j=1}^{N} L_{vij},$$

(33)

where $p_j$ is the ground-truth probability for the $j$-th AU.

The final AU probabilities are decided by the sum of the two-view features $f_1$ and $f_2$, denoted as $f_s$. In consistent with $L_{vi}$, $\hat{p}_j$ is denoted as the predicted probability for the $j$-th AU and the loss of the AU probabilities based on $f_s$ is defined as

$$L_{vj} = -p_j \log(\hat{p}_j) + (1 - p_j) \log(1 - \hat{p}_j);$$

$$L_{vs} = \frac{1}{N} \sum_{j=1}^{N} L_{vj}.$$

(34)

Similar to teacher network, a multi-view loss $L_{mv}$ is adopted to encourage the different view features to be different while complementary with each other. The multi-view loss for $N$ ME-AU classifiers is defined as

$$L_{mv} = \frac{1}{N} \sum_{j=1}^{N} \frac{W_{ij}^T W_{2j}}{\|W_{ij}\| \|W_{2j}\|},$$

(35)

where $W_{ij}$ represents the parameters of the $j$-th AU’s classifier of the $i$-th view. Finally, the main loss for ME-AU detection is formulated as

$$L_{au} = \frac{1}{2} \sum_{i=1}^{2} L_{vi} + L_{vs} + \lambda_{mv} L_{mv},$$

(36)

where $\lambda_{mv}$ is a hyper-parameter to for balancing the losses.

### 3.4.3 Attentive similarity-preserving knowledge distillation

Tung and Mori (2019)’s research verified that images of the same category tend to activate similar channels in a trained network and the activation similarities across different images can capture useful semantics. Furthermore, AUs only relate with specific
facial regions (E. Friesen & Ekman, 1978). Although the appearances of corresponding macro- and micro-expression AUs are different, the active facial regions are consistent and the active regions contribute more important AU knowledge. Therefore, an ASP is designed to transfer the dual-view AU knowledge through preserving attentive pairwise similarities, as shown in Figure 16. In this way, useful correlation knowledge can be transferred from macro-expressions to MEs for better AU detection on small ME databases.

Recent research (Z. Huang & Wang, 2017; Nikos & Sergey, 2017) demonstrated that the absolute value of a hidden neuron activation can indicate the importance of the neuron. In other words, the spatial attention map can be constructed by analyzing the statistics of absolute activation values across channels. Here, the spatial attention are generalized through summing the squared activations along the channel dimension.

Specifically, given a ME instance, the activation tensors of the $i$-th view at a particular layer $l$ produced by the teacher and student networks can be denoted as $A_T^{(l)} \in \mathbb{R}^{c'_t \times h'_t \times w'_t}$ and $A_S^{(l)} \in \mathbb{R}^{c_t \times h_t \times w_t}$, respectively. $c'_t$ and $c_t$, $h'_t$ and $h_t$, and $w'_t$ and $w_t$ are the channel and spatial dimensions for teacher and student networks, respectively. $A_T^{(l)}$ and $A_S^{(l)}$ consist of $c'_t$ and $c_t$ feature planes with spatial dimensions $h'_t \times w'_t$ and $h_t \times w_t$. The spatial attention maps are computed by the following equations:

$$A_{Tt}^{(l)} = \sum_{k=1}^{c'_t} |A_T^{(l)}_{Tt,k}|^2,$$

$$A_{St}^{(l)} = \sum_{k=1}^{c_t} |A_S^{(l)}_{St,k}|^2,$$

where $A_T^{(l)}_{Tt,k} = A_T^{(l)}(k, \ldots)$ and $A_S^{(l)}_{St,k} = A_S^{(l)}(k, \ldots)$ are attention maps of teacher and student networks, respectively. Furthermore, the attentive activation maps are defined as

$$F_T^{(l)} = A_T^{(l)} A_T^{(l)}.$$
where $F_{S}^{(l)}$ and $F_{T}^{(l)}$ represent the attentive activation maps of teacher network and student networks, respectively. $F_{S}^{(l)}$ and $F_{T}^{(l)}$ put more weight to the most discriminative parts and tend to distill more valuable knowledge efficiently.

Given a mini-batch MEs, let $Q_{T}^{(l)} \in \mathbb{R}^{b \times c \times h' \times w'}$ denote the reshaped $F_{T}^{(l)}$ and $F_{S}^{(l)}$, where $b$ is the batch size. The ASP distillation loss is defined to guide the student towards the activation correlations induced in the teacher, through penalizing differences in the L2-normalized outer products of $Q_{T}^{(l)}$ and $Q_{S}^{(l)}$, which are computed through the equations below:

$$
\tilde{G}_{T}^{(l)} = Q_{T}^{(l)} \cdot Q_{T}^{(l)T}; \quad G_{T[p,:]}^{(l)} = \tilde{G}_{T[p,:]}^{(l)}/\| \tilde{G}_{T[p,:]}^{(l)} \|_2,
$$

$$
\tilde{G}_{S}^{(l)} = Q_{S}^{(l)} \cdot Q_{S}^{(l)T}; \quad G_{S[p,:]}^{(l)} = \tilde{G}_{S[p,:]}^{(l)}/\| \tilde{G}_{S[p,:]}^{(l)} \|_2,
$$

where $\tilde{G}_{T}^{(l)}$ and $\tilde{G}_{S}^{(l)}$ are $b \times b$ matrices. Specifically, the entry $(p, q)$ in $\tilde{G}_{T}^{(l)}$ and $\tilde{G}_{S}^{(l)}$ encode the activation similarities at the $l$-th layer elicited by the $p$-th and $q$-th images in the mini-batch. A rowwise L2 normalization is applied to obtain the normalized $G_{T}^{(l)}$ and $G_{S}^{(l)}$, where $[p, :]$ denotes the $p$-th row in a matrix. We define the knowledge distillation loss for the dual views as

$$
L_{asp} = \frac{1}{b^2} \sum_{i \in \Gamma} \| G_{T}^{(l)} - G_{S}^{(l)} \|_F^2; \quad L_{asp} = \frac{1}{2} \sum_{i=1}^{2} L_{asp},
$$

where $\Gamma$ collects the layer pairs (layers at the end of the same block of ResNet, as shown in Figure.14). $\| \cdot \|_F$ is the Frobenius norm.

The total student loss $L_{student}$ is defined as

$$
L_{student} = L_{au} + \alpha L_{asp},
$$

where $L_{asp}$ is the loss of student model guided by pre-trained teacher network, and $L_{au}$ is the loss of ME-AU detection. $\alpha$ is hyper-parameter to for balancing the two sub-tasks.

### 3.4.4 Model training

For the teacher network pre-trained on EmotioNet, the Adam optimizer with learning rate of 0.001 is applied to optimize the network by setting $\lambda_{mv} = 400$ and $\lambda_{cr} = 100$. For the student network, we utilize the apex frames obtained by Paper II. All the student networks are pre-trained on ImageNet (Jia et al., 2009). During training, all the images are resized to $240 \times 240$ and then randomly cropped to $224 \times 224$. In the pre-processing step, the MEs are magnified with ratio 30 (H. Wu et al., 2012), according to the research.
of Paper II. The learning rate is set to 0.01 until 40 epochs. We set $\lambda_{mv}$ in Eq. 36 to 100 and $\alpha$ in Eq. 44 to 100 for balancing the losses. Following common experimental settings for AU detection (Niu, Han, Yang, et al., 2019), all of the experiments use the subject independent four-fold cross validation.

### 3.4.5 Experiments

#### Metrics

Following the previous methods for AU detection (Niu, Han, Yang, et al., 2019), macro F1 score is applied as the evaluation metric for all the experiments. We report the macro F1 score for each AU and the averaged macro F1 score over all AUs.

#### Comparisons of methods

We conduct comparisons with ResNet (He et al., 2016), SCA (Y. Li, Huang, & Zhao, 2021), Dual-view co-training fine-tuning (DVFT) (Niu, Han, Shan, & Chen, 2019), as well as knowledge distillation methods based on Similarity-Preserving (Tung & Mori, 2019), Attention Transfer (Nikos & Sergey, 2017), Fitnets (Romero et al., 2015). The AU detection results on the CASME II database are shown in Table 13. Specifically, ‘DV’ represents dual-view co-training. ‘-18’ and ‘-34’ stand for the depth of student network. ‘Fit’, ‘AT’, ‘SP’, and ‘ASP’ denote the methods based on Fitnets, Attention Transfer, Similarity-Preserving, and our proposed Attentive Similarity-Preserving knowledge distillation, respectively. All the methods are implemented by us.

As shown in Table 13, the DVASP-18 improves the average F1 score of the baseline (Res18) by about 16% on CASME II. Moreover, compared to the state-of-the-art ME-AU detection method SCA, DVASP-18 enhances the performances by 0.058 in terms of average F1 score on CASME II. The results demonstrate the superiority of our proposed DAVSP-18. Specifically, our proposed knowledge distillation method ASP-18 outperforms Fit-18, AT-18, SP-18 by 0.037, 0.003, and 0.031 on CASME II in terms of average F1 score. The results indicate that ASP can effectively distill and transfer the AU knowledge across macro- and micro-expression domains and the attentive correlation encodes useful semantics for across domain learning.

In order to verify the transfer learning effectiveness of DVASP, the DVASP-18 is compared with the DVFT fine-tuning on the pre-trained teacher network. From Table 13, it can be seen that DVASP-18 outperforms DVFT on CASME II. Specifically, DVASP-18 reaches higher F1 scores in six out of eight AUs on the CASME II database,
<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AU14</th>
<th>AU15</th>
<th>AU17</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.629</td>
<td>0.670</td>
<td>0.626</td>
<td>0.682</td>
<td>0.725</td>
<td>0.483</td>
<td>0.888</td>
<td>0.704</td>
<td>0.666</td>
</tr>
<tr>
<td>DVASP-34</td>
<td>0.726</td>
<td>0.715</td>
<td>0.713</td>
<td>0.715</td>
<td>0.726</td>
<td>0.719</td>
<td>0.719</td>
<td>0.716</td>
<td>0.721</td>
</tr>
<tr>
<td>ASP-34</td>
<td>0.717</td>
<td>0.690</td>
<td>0.672</td>
<td>0.695</td>
<td>0.717</td>
<td>0.693</td>
<td>0.693</td>
<td>0.697</td>
<td>0.700</td>
</tr>
<tr>
<td>SV</td>
<td>0.589</td>
<td>0.762</td>
<td>0.827</td>
<td>0.487</td>
<td>0.593</td>
<td>0.642</td>
<td>0.481</td>
<td>0.629</td>
<td>0.626</td>
</tr>
<tr>
<td>Res34</td>
<td>0.551</td>
<td>0.692</td>
<td>0.832</td>
<td>0.554</td>
<td>0.628</td>
<td>0.567</td>
<td>0.522</td>
<td>0.629</td>
<td>0.622</td>
</tr>
<tr>
<td>Fit-18</td>
<td>0.655</td>
<td>0.743</td>
<td>0.791</td>
<td>0.468</td>
<td>0.617</td>
<td>0.624</td>
<td>0.539</td>
<td>0.576</td>
<td>0.627</td>
</tr>
<tr>
<td>Fit-34</td>
<td>0.575</td>
<td>0.661</td>
<td>0.732</td>
<td>0.572</td>
<td>0.634</td>
<td>0.563</td>
<td>0.609</td>
<td>0.527</td>
<td>0.609</td>
</tr>
<tr>
<td>SP-18</td>
<td>0.690</td>
<td>0.685</td>
<td>0.777</td>
<td>0.511</td>
<td>0.570</td>
<td>0.653</td>
<td>0.566</td>
<td>0.613</td>
<td>0.633</td>
</tr>
<tr>
<td>SP-34</td>
<td>0.565</td>
<td>0.580</td>
<td>0.814</td>
<td>0.552</td>
<td>0.574</td>
<td>0.580</td>
<td>0.625</td>
<td>0.73</td>
<td>0.628</td>
</tr>
<tr>
<td>AT-18</td>
<td>0.650</td>
<td>0.627</td>
<td>0.831</td>
<td>0.576</td>
<td>0.586</td>
<td>0.625</td>
<td>0.466</td>
<td>0.656</td>
<td>0.627</td>
</tr>
<tr>
<td>AT-34</td>
<td>0.590</td>
<td>0.602</td>
<td>0.807</td>
<td>0.528</td>
<td>0.635</td>
<td>0.594</td>
<td>0.476</td>
<td>0.767</td>
<td>0.625</td>
</tr>
<tr>
<td>SCA</td>
<td>0.641</td>
<td>0.421</td>
<td>0.507</td>
<td>0.808</td>
<td>0.567</td>
<td>0.619</td>
<td>0.702</td>
<td>0.590</td>
<td>0.606</td>
</tr>
<tr>
<td>DVASP-18</td>
<td>0.726</td>
<td>0.721</td>
<td>0.898</td>
<td>0.569</td>
<td>0.796</td>
<td>0.685</td>
<td>0.715</td>
<td>0.700</td>
<td>0.726</td>
</tr>
<tr>
<td>DVASP-34</td>
<td>0.666</td>
<td>0.704</td>
<td>0.881</td>
<td>0.643</td>
<td>0.691</td>
<td>0.653</td>
<td>0.435</td>
<td>0.752</td>
<td>0.678</td>
</tr>
</tbody>
</table>

Table 13: F1 scores on the CASME II database. Res18 is the baseline (He et al., 2016). Reprinted with permission, Paper IV © 2021 IEEE.
Table 14. Validation the effectiveness of DVCT. Reprinted with permission, Paper IV © 2021 IEEE.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AU14</th>
<th>AU15</th>
<th>AU17</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit-18</td>
<td>0.655</td>
<td>0.743</td>
<td>0.791</td>
<td>0.468</td>
<td>0.617</td>
<td>0.624</td>
<td>0.539</td>
<td>0.576</td>
<td>0.627</td>
</tr>
<tr>
<td>DV+Fit-18</td>
<td>0.685</td>
<td>0.785</td>
<td>0.831</td>
<td>0.594</td>
<td>0.681</td>
<td>0.604</td>
<td>0.556</td>
<td>0.716</td>
<td>0.682</td>
</tr>
<tr>
<td>SP-18</td>
<td>0.690</td>
<td>0.685</td>
<td>0.777</td>
<td>0.511</td>
<td>0.570</td>
<td>0.653</td>
<td>0.566</td>
<td>0.613</td>
<td>0.633</td>
</tr>
<tr>
<td>DV+SP-18</td>
<td>0.778</td>
<td>0.737</td>
<td>0.876</td>
<td>0.573</td>
<td>0.619</td>
<td>0.684</td>
<td>0.653</td>
<td>0.631</td>
<td>0.694</td>
</tr>
<tr>
<td>AT-18</td>
<td>0.650</td>
<td>0.627</td>
<td>0.831</td>
<td>0.576</td>
<td>0.586</td>
<td>0.625</td>
<td>0.466</td>
<td>0.656</td>
<td>0.627</td>
</tr>
<tr>
<td>DV+AT-18</td>
<td>0.792</td>
<td>0.828</td>
<td>0.893</td>
<td>0.503</td>
<td>0.657</td>
<td>0.564</td>
<td>0.567</td>
<td>0.697</td>
<td>0.688</td>
</tr>
<tr>
<td>ASP-18</td>
<td>0.766</td>
<td>0.717</td>
<td>0.818</td>
<td>0.535</td>
<td>0.553</td>
<td>0.681</td>
<td>0.609</td>
<td>0.636</td>
<td>0.664</td>
</tr>
<tr>
<td>DV+ASP-18</td>
<td>0.726</td>
<td>0.721</td>
<td>0.898</td>
<td>0.569</td>
<td>0.796</td>
<td>0.685</td>
<td>0.715</td>
<td>0.700</td>
<td>0.726</td>
</tr>
</tbody>
</table>
when compared to DVFT. In general, the results suggest that the knowledge distillation method can perform better than the pre-trained and fine-tuning transfer learning strategy when the target domain data is insufficient.

**The impact of DVCT**

To validate the effectiveness of the teacher network based on DVCT, we develop a scheme to remedy the scarcity of ME databases. The DVCT scheme is applied to other knowledge distillation methods for fair evaluation. The comparison results are shown in Table 14. It can be seen that for all of the networks, the F1 score is improved with the teacher network based on DVCT. Specifically, DVCT enhances the ME-AU detection performances by 0.055, 0.061, and 0.061 for Fit-18, AT-18, and SP-18, respectively. The large improvements verify the effectiveness and generalization ability of DVCT. Moreover, our method DV+ASP-18 also outperforms DV+Fit-18, DV+AT-18, DV+SP-18 by a large margin (0.044, 0.032, and 0.038 on CASME II in terms of F1 score, respectively), indicating that DVASP provides a robust solution to distill dual-view AU knowledge for ME-AU detection which refers to domain shift.

**Visualization**

Figure 17 shows some example class activation maps. It can be seen that the DVASP can focus on the accurate region of most ME-AUs. For example, the baseline focuses on the wrong nose regions for AU14 (Dimpler). The ASP-18 only learns features from the left mouth corner. The proposed DVASP-18 focus on both mouth corners through dual views $v_1$ and $v_2$. This further verifies the effectiveness of DVASP for ME-AU detection. DVASP-18 achieves relatively lower F1 score on AU7 (Lid tighten) compared to other AUs, this may be caused by the small active region on lids and the similar appearance changes with eye blinking and gaze change, shown as the last row in Figure 17.

**3.5 Summary**

This chapter focuses on the ME-AU analysis. There is quite little research about AU detection in MEs. Thus, we firstly review the state-of-the-art progress about AU studies in macro-expressions. Then we introduce a ME-AU detection method with spatio-channel attention and a knowledge distillation method to leverage massive facial images for robust AU detection. Our contributions could be categorized into three aspects: (1) We are the first few researchers to study ME-AU detection and break the
ground and attract more researchers to ME-AU study; (2) We provide the baselines for ME-AU detection on the ME databases for future study; (3) The basic ME-AU detection framework is proposed and possible solutions (e.g., transfer learning and semi-supervised learning) for countering specific challenges in ME-AU analysis may inspire future work.

ME-AU detection analyzes facial movements at fine level. It is still at an early stage by now, but it is meaningful for understanding MEs. Currently, it attracts more attention and develops rapidly. In the future, we plan to continue to research ME-AU analysis and explore how to leverage the AU information for robust ME recognition. The detailed plans are introduced in Chapter 5.
4 Micro-expression recognition with action units

4.1 Introduction

The Facial Action Coding System (E. Friesen & Ekman, 1978) is a taxonomy of human facial expressions. With FACS, each facial expression can be objectively described as the combination of a set of AUs. However, AUs in FACS only describe facial muscle movements, and don’t provide the meaning of the behavior. W. V. Friesen et al. (1983) and (Ekman et al., 1998) proposed Emotional Facial Action Coding System (EMFACS) and Facial Action Coding System Affect Interpretation Dictionary (FACSAID) to translate the AUs into psychologically meaningful concepts. For instance, the sole inner brow raiser is sadness and the combination of inner and outer brow raiser is surprise. More examples of ordinary facial expressions with the their associated AUs are shown in Table 15. Considering the strong correlations between AUs and facial expressions, the AU information can be leveraged for accurate facial expression analysis (Sun et al., 2020). However, according to the annotations in current ME databases (Davison, Lansley, et al., 2018; X. Li et al., 2013; W. Yan et al., 2013; W. J. Yan, Li, et al., 2014), the relationship between AUs and MEs is relatively ambiguous. The contribution of AU detection to ME recognition is not very clear.

In this chapter, we study the contribution of robust AU detection to ME recognition. The contents of this chapter are organized as follows: Section 4.2 reviews the previous works about emotion recognition with AU detection; A robust ME recognition framework with AU detection method based on contrastive learning is described in Section 4.3 together with experimental results; In Section 4.4, we summarize the work in the chapter.

Table 15. Examples of AUs involved in some facial expressions.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>AUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>6, 12, 25</td>
</tr>
<tr>
<td>Sadness</td>
<td>1, 4, 6, 11, 15, 17</td>
</tr>
<tr>
<td>Fear</td>
<td>1, 2, 5, 26, 27</td>
</tr>
<tr>
<td>Anger</td>
<td>4, 5, 7, 10, 17, 22-26</td>
</tr>
<tr>
<td>Disgust</td>
<td>9, 10, 16, 17, 25, 26</td>
</tr>
</tbody>
</table>
4.2 Related work

During the past decades, there are encouraging progresses on facial expression recognition and AU detection. These two tasks have strong correlations and could be benefited from each other by considering their relationships.

Tobias and Kemal (2011) developed a common framework for AU detection and emotion recognition using local appearance-based face representation and multiple one-versus-all support vector machine classifiers corresponding to the specific tasks. Then, the contribution of AU detection for emotion recognition is further investigated in the work of Thibaud, Kevin, and Lionel (2014). Thibaud et al. (2014) compared two emotion detectors: directly on a high-dimensional feature space and on projected facial images in the low-dimensional space of AU intensities. Their results demonstrate that the influence of each AU on emotion recognition varies according to the type of AU and AU detection accuracy.

Recently, with the development of deep learning, an Adaptively Weights Sharing Network (AWS-Net) was proposed to automatically learn how to borrow information from the emotion recognition and AU detection tasks (Chu, Jiabei, Shiguang, & Xilin, 2019). Dac, Manh, Hyung-Jeong, Kim, and Lee (2021) utilized the association between the emotions and AUs in the AffWild2 database (Dimitrios & Stefanos, 2018) based on the multi-task learning architecture to combine the knowledge of the correlated tasks. In this way, the emotion recognition and AU detection performance can be enhanced by a large margin. All above studies focused on macro-expressions. There is limited research which investigates AU detection for ME recognition.

4.3 ME recognition with AU detection based on contrastive learning

A robust AU detection is important for ME analysis. In this section, we first introduce a robust AU detection with intra- and inter-contrastive learning (IICL) and then develop it to recognize MEs through multi-task learning. The results demonstrate that our proposed method is able to efficiently identify subtle AUs and robust AU detection can benefit ME recognition.

4.3.1 Intra-contrastive learning

ME-AU detection suffers from low intensity, as shown in Figure 18. The frames around the apex frame are regarded as close-to-apex frames. From Figure 18, it can be seen that even the apex frame with the highest intensity does not have much difference compared
with the onset frame. To cope with the problem, an IntraCL module composed of three contrastive losses is constructed to make sure that the AU-related features of apex and close-to-apex are far away from the features of onset in the feature space, as illustrated in Figure 19.

Firstly, we locate the apex frame based on the frequency representation of facial muscle change in the frequency domain (Y. Li et al., 2020). Then, a contrastive loss (Hadsell, Chopra, & LeCun, 2006; B. Wu, Wei, Wu, & Lin, 2019) of the onset and apex frames $L_{OA}$ is developed to maximize the difference between the onset and apex frames in the feature space to obtain the discriminative representation of ME-AUs. Moreover, considering the limited number of apex frames may restrict the learning ability for ME-AUs, IntraCL further explores the relationship with weaker close-to-apex frames which are more commonly displayed in MEs. Specifically, $L_{OC}$ loss is designed to push the close-to-apex apart from the onset, as shown in Figure 19 (c).

The $L_{OA}$ and $L_{OC}$ are defined as

$$L_{OA} = \frac{1}{N} \sum_{i=1}^{N} \max \{0, \delta - \| f(I_{o_i}) - f(I_{a_i}) \|_2^2 \}, \quad (45)$$

$$L_{OC} = \frac{1}{N} \sum_{i=1}^{N} \max \{0, \delta - \| f(I_{o_i}) - f(I_{c_i}) \|_2^2 \}, \quad (46)$$

where $f(I_{o_i})$, $f(I_{a_i})$, and $f(I_{c_i})$ represent the normalized features of the onset, apex, and close-to-apex frames. The objectives of $L_{OC}$ and $L_{OA}$ are learning representations with a greater distance for onset and apex frames, and onset and close-to-apex frames, respectively. In this way, the AUs in apex and close-to-apex frames can be differentiated from onset frames. Specifically, when the distance is not bigger than $\delta$, the loss will be positive and the net parameters will be updated to generate more discriminative features for subtle ME-AU representation. $\delta$ is a margin and set to 1 in our experiment following (Schroff, Kalenichenko, & Philbin, 2015). $N$ is the training batch size.

Furthermore, recent research (X. Zhao et al., 2016) demonstrated that considering the intrinsic correlations between weak and strong expressions can achieve better results on
Fig. 19. Illustration of the training and testing stages of IICL for ME-AU detection. During training, IICL takes the processed onset, apex, and close-to-apex images as input. After passing the images through several convolutional layers, the features can be obtained for onset, apex, and close-to-apex images, respectively. The IntraCL drives the AU-related features of apex and close-to-apex away from the features of onset. The InterCL module automatically selects the negative pairs (a pair of apex images without the same AU presence) in a batch and enlarges the distance between different AUs. During testing, the IICL takes the apex as input, outputting the predicted probabilities for all MEAUs. Reprinted with permission, Paper V © 2021 ACM.

weak expressions. Inspired by X. Zhao et al. (2016), AUs in both apex and close-to-apex frames are classified during training, as shown in Figure 19 (e). Moreover, a loss termed as $L_{AC}$ is designed to pull apex and close-to-apex frames towards each other for robust ME-AU detection. The $L_{AC}$ is formulated as

$$L_{AC} = \frac{1}{N} \sum_{i=1}^{N} \| f(I_{a_i}) - f(I_{c_i}) \|_2^2,$$

where $f(I_{a_i})$ and $f(I_{c_i})$ are the normalized apex and close-to-apex features, respectively.

4.3.2 Inter-contrastive learning

AUs in ME apex frames still have low intensity. This makes it difficult to distinguish different AUs in MEs. InterCL designs a strategy automatically choosing the negative ME AU pairs in a batch and enlarges their difference during training to improve the AU detection robustness, as shown in Figure 19 (d).

As AUs can co-exist in MEs, only the pairs without the same AU presence can be treated as the negative pairs. Thus, the negative AU pairs should be orthogonal. In practical, the negative ME pairs are decided following the equation below:
\[ C_{negative} = \text{Aus}_i \cdot \text{Aus}_j, \]  

where \( \text{Aus}_i \) and \( \text{Aus}_j \) are the i-th and j-th ME AU labels in a batch and \( 1 \leq i < j \leq N \). \( N \) represents the batch size. If \( C_{negative} = 0 \), it is the negative pair.

Then the contrastive loss \( L_{NA} \) is employed to maximize the apex feature distance of negative AU pairs for the improvement of AU detection robustness.

\[
L_{NA} = \frac{1}{K} \sum_{k=0}^{K} \max \{ 0, \delta - \| f(I_{a_i}) - f(I_{a_j}) \|_2^2 \},
\]

where \( f(I_{a_i}) \) and \( f(I_{a_j}) \) are i-th and j-th normalized apex features belonging to a negative pair. \( K \) represents the number of negative AU pairs in a batch. Similar to \( L_{OC} \) and \( L_{OA} \), \( L_{NA} \) will be positive when the distance is smaller than the margin value \( \delta \) and network will be updated for negative AU pairs, so that the training can focus on more difficult AU pairs. The margin \( \delta \) is set to 1 following the work of Schroff et al. (2015).

### 4.3.3 ME-AU detection objective

For multi-label ME-AUs, each AU can be treated as a specific task, as shown in Figure 19 (e). The loss for each task is defined as a binary cross-entropy loss. Thus, the loss of \( M \) AUs is formulated as

\[
L_{AUs} = \frac{1}{M} \sum_{m=1}^{M} y_m \log(\hat{y}_m) + (1 - y_m) \log(1 - \hat{y}_m),
\]

where \( y_m \) is the ground truth for the \( m \)-th AU in the ME, with 1 denoting occurrence of the AU and 0 denoting absence. \( \hat{y}_m \) is the predicted probability of the occurrence of \( m \)-th AU. \( M \) is number of AU categories.

The overall loss of the IICL framework is defined as

\[
L_{AUtotal} = L_{AU_i} + L_{AUc} + \lambda (L_{OA} + L_{OC} + L_{AC} + L_{NA}),
\]

where \( L_{AU_i} \) and \( L_{AUc} \) are the losses for AUs in apex and close-to-apex frames, respectively. \( \lambda \) is the hyper-parameter that balances the influence of contrastive loss.

### 4.3.4 A framework for ME recognition with AU detection

To leverage the AUs for ME analysis, we design an end-to-end ME recognition and AU detection framework based on IICL with multi-task learning. The classifier module is shown in Figure 20.
Fig. 20. Classifier module for ME recognition with AU detection based on multi-task learning.

The loss of $L_{MEAUs}$ is the combination of the overall loss of the IICL framework for AU detection and ME recognition loss $L_{ME}$, that is

$$L_{MEAUs} = \alpha L_{AU total} + L_{ME}, \quad (52)$$

where $\alpha$ is the weight balancing the losses.

4.3.5 Model training

In the experiments, the aligned face images provided by the databases are employed. The three frames before and after the apex frame are regarded as the close-to-apex frames. The popular SEnet-50 (Hu, Shen, & Sun, 2018) is employed as the backbone. All models were pre-trained on VGG-FACE database (Parkhi et al., 2015). The images are resized to $256 \times 256$ and then randomly cropped to $224 \times 224$ for training purposes (Niu, Han, Yang, et al., 2019). In the pre-processing step, the MEs are magnified with ratio 10, according to the research of (X. Li et al., 2018; H. Wu et al., 2012). All of the methods are evaluated with magnified MEs. During training, the networks are optimized using SGD. The initial learning rate is set to 0.01, divided by 10 every 40 epochs until 80 epochs. The weight of the contrastive losses $\lambda$ is set to 0.1 for balancing the losses. Following the experimental settings in Paper III, the subject independent four-fold cross validation is used for ME-AU detection in our experiments. For ME recognition, the leave-one-subject-out protocol is employed.

4.3.6 Experiments

Ablation study

In this section, we provide ablation study on the CASME II database to investigate the effectiveness of each part in the IICL network. To verify the effectiveness of IntraCL and InterCL modules, we add IntraCL and InterCL modules on the baseline, separately.
Table 16. Abalation study on the CASME II database. The baseline is SEnet (Hu et al., 2018). Reprinted with permission, Paper V © 2021 ACM.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AU14</th>
<th>AU15</th>
<th>AU17</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.43</td>
<td>0.48</td>
<td>0.89</td>
<td>0.20</td>
<td>0.60</td>
<td>0.59</td>
<td>0.20</td>
<td>0.17</td>
<td>0.45</td>
</tr>
<tr>
<td>IntraCL</td>
<td>0.85</td>
<td>0.66</td>
<td>0.90</td>
<td>0.26</td>
<td>0.63</td>
<td>0.49</td>
<td>0.20</td>
<td>0.25</td>
<td>0.53</td>
</tr>
<tr>
<td>InterCL</td>
<td>0.70</td>
<td>0.58</td>
<td>0.90</td>
<td>0.22</td>
<td>0.48</td>
<td>0.48</td>
<td>0.15</td>
<td>0.43</td>
<td>0.49</td>
</tr>
<tr>
<td>IICL</td>
<td>0.78</td>
<td>0.67</td>
<td>0.89</td>
<td>0.30</td>
<td>0.56</td>
<td>0.53</td>
<td>0.33</td>
<td>0.33</td>
<td>0.55</td>
</tr>
</tbody>
</table>

As shown in Table 16, the framework with InterCL outperforms the baseline by 0.04 in terms of the average F1 score on CASME II. The framework with IntraCL reaches higher F1 scores in six out of eight AUs on CASME II, compared with the baseline. The results demonstrate the effectiveness of InterCL and IntraCL modules. Furthermore, IICL consisted of InterCL and IntraCL modules achieves the best performance and improves the average F1 score by 22.22%, in comparison with the baseline. The results indicate that contrastive learning can explore discriminative representation for subtle ME-AUs. From Table 16, it can be seen that the F1 scores decline on AU12 (Lip corner puller) and AU14 (Dimpler). This may caused by the similar appearance changes on the same region lip corner. It is hard to distinguish them. In general, the results suggest that enlarging the between-frame and between-AU differences can improve the discriminative ability of subtle AUs and is useful for most ME-AUs.

Comparisons of methods of AU detection

Table 17 shows the AU detection results of different methods on the CASME II database. The proposed IICL and the baseline methods based on SEnet (Hu et al., 2018) are tested on apex images in MEs. Compared to the baseline (SEnet), the IICL reaches higher F1 scores in five out of eight AUs on the CASME II database and all AUs on the SAMM database. Furthermore, IICL outperforms SEseq which employs temporal information through aggregating the onset, close-to-apex, and apex frame features by 0.13 in terms of the average F1 score on CASME II. Moreover, IICL achieves large improvements on the challenging AUs containing unclear motions and very few samples, compared with the baseline on CASME II (0.43 vs. 0.78 on AU1, 0.20 vs. 0.33 on AU15, and 0.17 vs 0.33 on AU17 in terms of F1 score). The results demonstrate that learning the between-frame and between-AU contrastive information can improve the discriminative ability for ME-AU detection. In order to further verify the IICL effectiveness of ME-AU detection, the IICL is compared to the methods contrastive learning based on Contrastive
Table 17. F1 scores on the CASME II database. The baseline is SEnet (Hu et al., 2018). 
Reprinted with permission, Paper V © 2021 ACM.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AU1</th>
<th>AU2</th>
<th>AU4</th>
<th>AU7</th>
<th>AU12</th>
<th>AU14</th>
<th>AU15</th>
<th>AU17</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.43</td>
<td>0.48</td>
<td>0.89</td>
<td>0.20</td>
<td>0.60</td>
<td>0.59</td>
<td>0.20</td>
<td>0.17</td>
<td>0.45</td>
</tr>
<tr>
<td>SEseq</td>
<td>0.47</td>
<td>0.41</td>
<td>0.91</td>
<td>0.20</td>
<td>0.36</td>
<td>0.60</td>
<td>0.26</td>
<td>0.13</td>
<td>0.42</td>
</tr>
<tr>
<td>RESnet</td>
<td>0.51</td>
<td>0.35</td>
<td>0.90</td>
<td>0.11</td>
<td>0.62</td>
<td>0.51</td>
<td>0.22</td>
<td>0.00</td>
<td>0.40</td>
</tr>
<tr>
<td>RESseq</td>
<td>0.24</td>
<td>0.21</td>
<td>0.86</td>
<td>0.09</td>
<td>0.43</td>
<td>0.49</td>
<td>0.20</td>
<td>0.27</td>
<td>0.35</td>
</tr>
<tr>
<td>SCA</td>
<td>0.29</td>
<td>0.45</td>
<td>0.89</td>
<td>0.25</td>
<td>0.48</td>
<td>0.33</td>
<td>0.40</td>
<td>0.52</td>
<td>0.45</td>
</tr>
<tr>
<td>CL</td>
<td>0.65</td>
<td>0.58</td>
<td>0.90</td>
<td>0.30</td>
<td>0.56</td>
<td>0.50</td>
<td>0.20</td>
<td>0.25</td>
<td>0.49</td>
</tr>
<tr>
<td>TL</td>
<td>0.71</td>
<td>0.62</td>
<td>0.88</td>
<td>0.28</td>
<td>0.50</td>
<td>0.54</td>
<td>0.32</td>
<td>0.32</td>
<td>0.52</td>
</tr>
<tr>
<td>IICL</td>
<td>0.78</td>
<td>0.67</td>
<td>0.89</td>
<td>0.30</td>
<td>0.56</td>
<td>0.53</td>
<td>0.33</td>
<td>0.33</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 18. Comparisons of ME recognition on CASEM II. The baseline framework is SEnet (Hu et al., 2018).

<table>
<thead>
<tr>
<th>Methods</th>
<th>ACC (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>64.37</td>
<td>0.61</td>
</tr>
<tr>
<td>IICL_ME</td>
<td>66.58</td>
<td>0.64</td>
</tr>
<tr>
<td>IICL_MEAU</td>
<td>66.39</td>
<td>0.65</td>
</tr>
</tbody>
</table>

loss (CL) (Hadsell et al., 2006) and Triplet loss (TL) (Schroff et al., 2015) and SCA (Y. Li et al., 2021). In Table 17, it can be seen that IICL improves the average F1 score by around 12% and 6% on CASME II, in comparison with CL and TL, respectively. Furthermore, IICL outperforms SCA by 0.10 in terms of average F1-score on CASME II. The results indicate that the IICL can learn discriminative AU representation from the subtle movements in MEs effectively.

Comparisons of ME recognition with AU detection

In this subsection, we verify AU contribution for ME recognition. Table 18 demonstrates the results of ME recognition performance with AU detection which is denoted as $IICL_{MEAU}$. Compared to the baseline, the IICL framework with single ME recognition task (denoted as $IICL_{ME}$) improves the accuracy by 1.88% which further verifies the effectiveness of our proposed IICL framework. Moreover, the $IICL_{MEAU}$ outperforms the $IICL_{ME}$ by 1.23% in terms of accuracy. The results demonstrate that AU detection contributes to improving ME recognition performance.
4.4 Summary

In this chapter, a ME recognition framework with AU detection is introduced. We first review the studies of AU analysis for facial expression recognition. Then our proposed ME recognition method with AU detection is described including: (1) the intra- and inter-contrastive learning method for ME-AU detection; (2) a framework for ME recognition with AU detection; and (3) the experimental results which demonstrate the AUs’ contribution to ME recognition.

Encoding expressions through facial AUs coded by the FACS is effective for describing MEs and improves the MEs’ recognition performance. In the future, we plan to continue to explore more advanced methods to enhance ME recognition performance by leveraging AU information. The detailed plans are introduced in Chapter 5.
5 Summary

Emotion analysis is important in human’s daily life. Currently, most research focuses on perceiving emotion based on common facial expressions, audios, and gestures. However, the emotion can be suppressed by people’s will, lead to the concealment of true feelings. Recent research demonstrates that a fast and involuntary expression would occur when people try to hide their true feelings, called micro-expressions. MEs can reveal people’s hidden feelings in high-stake situations and have many potential applications in different fields, such as clinical diagnosis, national security, and interrogation. This thesis presents works of using computer vision methodologies to analyse MEs from facial videos.

5.1 Contributions

The contributions of this thesis are from three aspects of ME analysis: ME recognition with deep learning, ME-AU detection, and ME recognition with AU detection.

5.1.1 ME recognition with deep learning

The first aspect of contribution concerns ME recognition based on deep learning. Recently, the deep learning has achieved considerable performance in various fields. However, the deep learning-based ME recognition is still challenging, due to the fact that MEs have small-scale databases and subtle facial movements. Most recent works have attempted to recognize MEs with spatial and temporal information from video clips. According to psychological studies, the apex frame conveys the most emotional information expressed in facial expressions. However, it is not clear how the single apex frame contributes to ME recognition. Paper I studied the contribution of apex frame in MEs with deep learning. A novel method was proposed to detect the apex frame by estimating pixel-level change rates in the frequency domain. With frequency information, it performed more effectively on apex frame spotting than the currently existing apex frame spotting methods based on the spatio-temporal change information. After the apex frame was obtained, the MEs were recognized with deep model only based on the magnified apex frame. The experimental results demonstrated that our proposed method was effective compared to the state-of-the-art methods and the only apex frame can work well for ME recognition.

To further improve the ME recognition performance, Paper II proposed a joint feature learning architecture coupling local and global information to recognize MEs, because
not all regions made the same contribution to ME recognition and some regions did not even contain any emotional information. More specifically, the proposed model involved the local information learned from the facial regions which contribute major emotion information and the global information learned from the whole face. Leveraging the local and global information enabled our model to learn discriminative ME representations and suppressed the negative influence of unrelated regions to MEs. The proposed method was extensively evaluated using CASME, CASME II, SAMM, SMIC, and composite databases. Experimental results demonstrated that our method with the detected apex frame achieved considerably promising ME recognition performance, compared to the state-of-the-art methods employing the whole ME sequence. Moreover, the results further indicated that the apex frame could significantly contribute to ME recognition.

5.1.2 ME-AU detection

The second aspect of contribution of this thesis concerns ME-AU analysis. The FACS indicates that successful AU detection greatly facilitates the analysis of the complicated facial actions or expressions (W. Li et al., 2017). In other words, it is very essential to explore AUs for deeply interpreting the facial behavior of expressions. To the best of our knowledge, few work is conducted on analyzing AUs for MEs. Compared to macro-expression AU detection, ME-AU detection becomes more difficult due to the low intensity, short duration, very small number of samples, and low AU correlation in MEs. To address some of the above-mentioned problems, in Paper III, we proposed a spatio-channel attention mechanism exploring the second-order correlations of spatio-wise and channel-wise features with multi-task learning to better represent subtle ME-AUs. To the best of our knowledge, this was the first work to detect ME-AUs. Intensive experiments were conducted on three publicly ME data-bases with AU labels. The results demonstrated the effectiveness of our method. Moreover, we provided the baseline results of ME-AU detection to the research community.

To address the small-scale ME database issue, Paper IV proposed a novel dual-view attentive similarity-preserving distillation method for robust ME-AU detection by leveraging massive facial expressions in the wild. Through such an attentive similarity-preserving distillation method, we broke the domain shift problem and essential AU knowledge from common facial AUs was efficiently distilled. Furthermore, considering that the generalization ability of teacher network is important for knowledge distillation, a semi-supervised co-training approach was developed to construct a generalized teacher network for learning discriminative AU representation. Extensive experiments illustrated
that our proposed knowledge distillation method was able to effectively distill and transfer the cross-domain knowledge for robust ME-AU detection.

5.1.3 ME recognition with AU detection

The last aspect of contribution concerns ME recognition with AU analysis. The survey of facial expression recognition (S. Li & Deng, 2020) introduced that simultaneously conducting facial expression recognition with AU detection could jointly improve the performance of facial expression recognition (Pons & Masip, 2018). To further verify the contribution of AUs for ME analysis, we designed a multi-task learning framework for ME recognition and AU detection based on contrastive learning. The experimental results demonstrated that the AU detection made important contribution to ME recognition.

5.2 Limitations and future work

ME recognition has drawn increasing interest recently due to its practical importance in various fields. Although the ME analysis has achieved improvements with the development of deep learning, the ME recognition performance is still limited and current ME recognition methods are hard to be utilized in the realistic situations.

The ME studies can be continued and improved from three aspects in future work. First, about the ME databases: Although significant progress has been made for ME recognition in recent years, most existing ME recognition algorithms are based on 2D facial images and sequences. The data in 2D domain can not solve the challenging problems of illumination and pose variations in real-world applications. Recent research illustrates that the lighting and pose variation issues can be addressed through 3D facial data (Sandbach, Zafeiriou, Pantic, & Yin, 2012). Moreover, 3D geometry information may include important features for facial expression recognition and provide more data for better training. Thanks to the benefits of 3D faces and technological development of 3D scanning, establishing a 3D ME database and ME recognition based on 3D sequence could be future research directions.

Second, about the ME spotting: The ME apex frame spotting method based on frequency information analysis is described in Section 2.3, which can effectively locate the apex frame in ME clips. However, utilizing the frequency based apex frame spotting method can not achieve the end-to-end ME recognition based on apex frames and it aimed to locate the apex frame in ME clips. In the future, more deep learning-based apex frame spotting methods should be explored. Moreover, the frequency-based ME
apex frame spotting method could be extended to spot ME clips in long videos. One challenge of current spotting methods is the existence of the brief but non-emotional facial movements, e.g., blinking of eyes. Fine-grained ME spotting methods based on AU analysis should be studied. Besides, MEs and macro-expressions may occur at the same time in long videos Future studies should explore methods to detect and distinguish the micro- and macro-expressions when they occur simultaneously.

Third, about the ME recognition: One of the major challenge in ME recognition is the small-scale ME databases. To alleviate the small-scale database issue, Paper II fine-tuned the MEs on a pre-trained VGG model to avoid over-fitting. The results suggest that leveraging the relevant task can contribute important information for ME recognition and effective transfer learning methods for ME recognition should be explored. However, the current methods only leveraged the facial images with labels which are also limited. Considering that there are massive unlabeled facial videos on the internet, semi-supervised learning and un-supervised learning can be further studied to take advantage of the unlabeled facial images for further ME recognition improvement. Besides, encoding AUs has been verified as effective for enhancing ME recognition performance. The relationship between AUs and MEs can be further explored through graph learning. In the future, we could utilize the graph to model the complex relationship between AUs and emotions for robust ME recognition.
References


(FG) (Vol. 6, pp. 1–6).


Original publications


IV Yante Li & Guoying Zhao (2021). Micro-expression Action Unit Detection with Dual-view Attentive Similarity-Preserving Knowledge Distillation. IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 01-08. IEEE.


Reprinted with permission from IEEE, ACM and Elsevier.

Original publications are not included in the electronic version of the dissertation.


813. Hultgren, Matias (2021) Control design for CFB boilers integrated with process design

814. Safarpour, Mehdi (2021) Energy efficient solutions for computing and sensing

815. Kupila, Riikka (2021) Catalytic conversion of furfural and glucose over activated carbon-supported metal catalysts

816. Okwube, Jude (2021) Software-defined resource management for industrial internet of things

817. Luttinen, Esko (2022) Spectrum as a resource, spectrum as an asset: a techno-economic study


819. Tervo, Nuutti (2022) Concepts for radiated nonlinear distortion and spatial linearization in millimeter-wave phased arrays

820. Nissilä, Tuukka (2022) Ice-templated cellulose nanofiber structures as reinforcement material in composites

821. Yu, Zizong (2022) Physiological signals measurement and spoofing detection from face video

822. Chen, Haoyu (2022) Human gesture and micro-gesture analysis: datasets, methods, and applications

823. Khan, Iqra Sadaf (2022) Exploring Industry 4.0 and its impact on sustainability and collaborative innovation


825. Zhang, Ruichi (2022) Vanadium removal and recovery from liquid waste streams

826. Väätäjä, Maria (2022) Prospects of the room temperature fabrication method for electroceramics: feasibility for printing techniques and integration with temperature-sensitive materials

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi
Yante Li

MACHINE LEARNING FOR PERCEIVING FACIAL MICRO-EXPRESSION