

# **Đề tài: CFNet – Facial Expression Recognition via Constraint Fusion Under Multi- task Joint Learning Network**

**Lớp : D22CNPM02**



# Thành viên

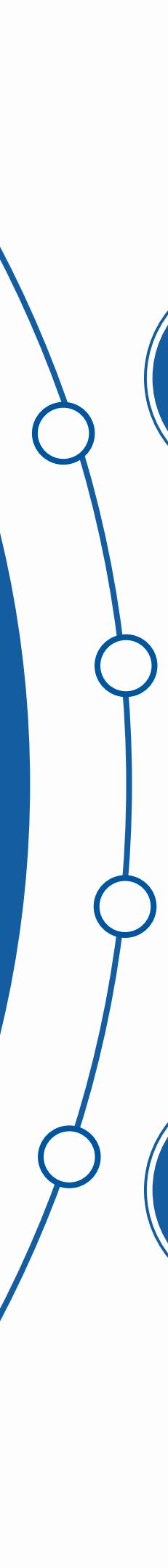


Đào Thị Huyền - B22DCCN399



Đặng Thị Huyền - B22DCCN400





01

**Giới thiệu chung**

02

**Trình bày bài toán**

03

**Phương pháp đề xuất**

04

**Kết quả thực nghiệm**

# 1. Giới thiệu chung

## 1.1. Thông tin bài báo

- **Tiêu đề:** CFNet: Facial Expression Recognition via Constraint Fusion under Multi-task Joint Learning
- **Tác giả:** Junhao Xiao, Chenquan Gan, Qingyi Zhu, Ye Zhu, Gang Liu
- **Nơi công bố:** Tạp chí Applied Soft Computing, Số 141, Trang 110312
- **Năm công bố:** 2023
- **Lĩnh vực nghiên cứu:** Xử lý ảnh & Thị giác máy tính

Applied Soft Computing 141 (2023) 110312  
Contents lists available at ScienceDirect  
 Applied Soft Computing  
journal homepage: [www.elsevier.com/locate/asoc](http://www.elsevier.com/locate/asoc)



CFNet: Facial expression recognition via constraint fusion under multi-task joint learning network

Junhao Xiao <sup>a</sup>, Chenquan Gan <sup>a,b</sup>, Qingyi Zhu <sup>b</sup>, Ye Zhu <sup>c,\*</sup>, Gang Liu <sup>d,\*</sup>

<sup>a</sup> School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

<sup>b</sup> School of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

<sup>c</sup> Centre for Cyber Resilience and Trust, Deakin University, Victoria 3125, Australia

<sup>d</sup> College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

### ARTICLE INFO

Article history:  
Received 25 August 2022  
Received in revised form 2 April 2023  
Accepted 9 April 2023  
Available online 19 April 2023

Dataset link: <https://github.com/robert181818/CFNet>

Keywords:  
Facial expression recognition  
Multi-task joint learning network  
Multi-loss mechanism  
Constraint fusion

### ABSTRACT

In facial expression recognition (FER), global and local features obtained from the same face may have different recognition accuracy, indicating that they have different advantages in recognition. Existing FER works usually focus on extracting and fusing global and local features to obtain better recognition results. Instead of evaluating the advantages of global and local features before fusion, these methods default to fusing them in the same proportion, which probably leads to mutual suppression of information representation between the two features, and then causes worse recognition ability and scene adaptability. To overcome this weakness, this paper proposes a multi-task joint learning network with a constraint fusion (called CFNet). To leverage the key features extracted from different tasks, CFNet adopts a multi-loss mechanism and a constraint fusion method to automatically assign corresponding weights based on the importance of global and local facial information. Compared with existing models that employ the direct fusion strategy, CFNet has better adaptability for FER in complex scenes. Extensive evaluations show the superior effectiveness of CFNet over state-of-the-art methods on real-world emotion datasets. Specifically, the accuracy scores of CFNet on CK+, MMI, and RAF-DB datasets are 99.07%, 84.62%, and 87.52%, respectively. The robustness of CFNet is also verified in noisy and blurred scenes.

© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

### 1. Introduction

The increasing richness of social scenes is accompanied by the diversity of human emotions [1]. In the process of human social communication, speech, text, and facial expression are common emotional carriers [2,3]. In social contact, facial expressions occupy about 55% proportion of all emotional transmission modes [4]. Generally, facial expressions are not only a natural and direct means for humans to communicate their emotions and intentions, but also the key means of non-verbal communication [5]. In the past, it was time-consuming to analyze facial emotions with the help of traditional image processing methods. However, deep learning techniques have made this task efficient [6–8]. Meanwhile, facial expression recognition (FER) is increasingly used in many scenarios in the era of rapid development of science and technology [9–10] e.g., mental health assessment, driver emotion

The FER method is designed to obtain stable emotional features via facial variation, and to complete the emotional classification after normalizing the features [11]. The existing FER methods recognize emotions based on extracted global features, local features, or their aggregations [12]. Global feature extraction methods consider the whole facial image as an object. These methods are good to use appearance texture [13] and geometric structure [14] as the transmission medium of facial emotions. Moreover, the attention mechanism is adopted to improve the capacity to collect vital information on global features [15,16]. However, the extracted global features usually contain redundant information that may suppress some key facial information and negatively affect the FER performance. Especially in real-world (wild) conditions, the background information of image samples introduces more redundancy, which is more unfriendly

# 1. Giới thiệu chung

## 1.2. Lý do chọn đề tài



### Giải quyết vấn đề thực tế

Các phương pháp FER hiện có **hợp nhất** trực tiếp các đặc trưng toàn cục và cục bộ theo tỷ lệ cố định, dẫn đến **ức chế** lẫn nhau giữa các thông tin quan trọng



### Tính mới và sáng tạo

CFNet cải thiện việc **kết hợp global-local** bằng một cơ chế fusion có ràng buộc và **multi-task learning**, giúp mô hình nhẹ hơn, chính xác hơn và ổn định hơn so với các phương pháp truyền thống

## 2. Trình bày bài toán

### 2.1. Mục tiêu bài toán

- **Mục tiêu tổng quát:** phát triển một mô hình mới kết hợp đặc trưng global-local theo cách tự điều chỉnh thông minh (constraint fusion) dưới kiến trúc multi-task để tăng độ chính xác, độ ổn định và khả năng tổng quát trong nhận dạng biểu cảm khuôn mặt.
- **Mục tiêu cụ thể:**
  - **Giải quyết hạn chế của việc kết hợp đặc trưng Global–Local:** Tìm cách kết hợp global và local một cách linh hoạt, có trọng số thích ứng, để không loại bỏ thông tin quan trọng
  - **Các feature (local feature, global feature) có biểu diễn tốt :** không học tách biệt local feature và global feature, mà học đồng thời và bổ trợ cho nhau
  - **Cải thiện độ chính xác và khả năng thích ứng:** Đạt kết quả cao trên các dataset chuẩn và tăng hiệu quả trong môi trường thực tế.
  - **Tăng độ ổn định:** Hoạt động tốt trong điều kiện nhiễu và duy trì hiệu suất với ảnh mờ.

## 2. Trình bày bài toán

### 2.2. Bài toán nghiên cứu

- **Định nghĩa bài toán:**

- **Input:** Ảnh khuôn mặt  $I \in \mathbb{R}^{nxnx3}$
- **Output:** Phân loại vào 7 cảm xúc cơ bản (怒 Anger, 冷漠 Contempt, 厌恶 Disgust, 恐惧 Fear, 高兴 Happiness, 悲伤 Sadness, 惊讶 Surprise)

- **Các thách thức:**

- **⚠ Trích xuất đặc trưng toàn cục:** Nhiều thông tin dư thừa, nhiễu từ nền, dễ che mất thông tin quan trọng → kém hiệu quả trong điều kiện thực tế.
- **⚠ Trích xuất đặc trưng cục bộ:** Chia khuôn mặt thành vùng nhỏ độc lập nên mất liên hệ giữa các vùng → hiệu suất không đạt yêu cầu.
- **⚠ Hợp nhất đặc trưng kém hiệu quả:** Hầu hết hợp nhất theo tỷ lệ cố định 50 - 50, không xem xét mức quan trọng → triệt tiêu lẫn nhau.
- **⚠ Sự khác biệt cá nhân:** Biểu cảm mỗi người khác nhau, các vùng khuôn mặt quan trọng thay đổi tùy người.

## 2. Trình bày bài toán

### 2.3. Động cơ nghiên cứu

#### Quan sát

- Global và local features từ cùng một khuôn mặt có độ chính xác khác nhau → Chúng có lợi thế khác nhau
- Biểu cảm = phối hợp nhiều cơ mặt → Mỗi người thể hiện khác nhau tùy thuộc vào giới tính và thói quen cá tính của họ

#### Ý tưởng

- Hợp nhất không nên dùng tỷ trọng cố định
- Cần chiến lược hợp nhất thích ứng tự động gán trọng số theo tầm quan trọng của các đặc trưng
- Phải đánh giá mức quan trọng trước hợp nhất

#### Giải pháp

1. Học đồng thời global & local, duy trì mối tương quan
2. Đánh giá động **tầm quan trọng** của từng loại feature
3. Hợp nhất thích ứng dựa trên mức quan trọng
4. Hoạt động ổn định trong tình huống phức tạp

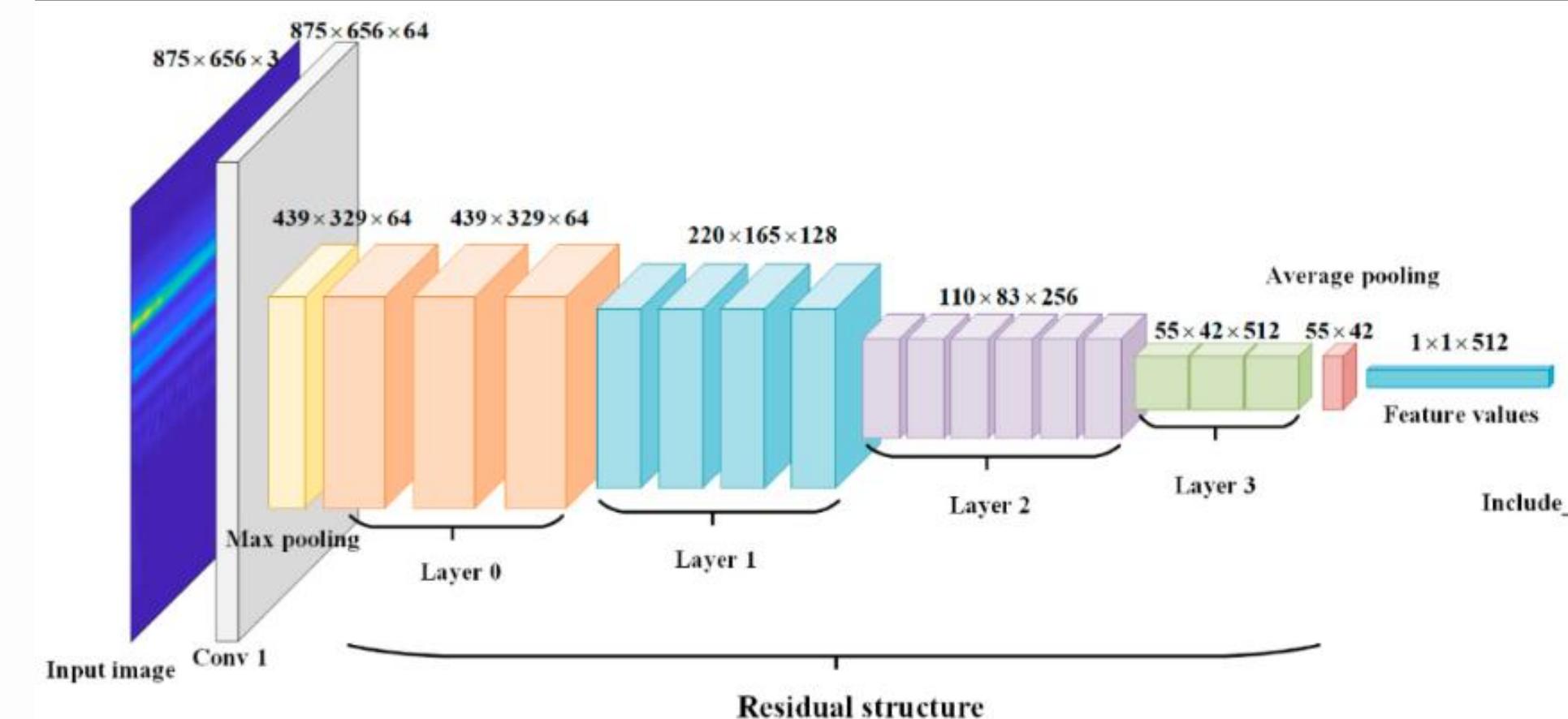
### 3. Phương pháp đề xuất

#### 3.1. Tổng quan

##### Các phương pháp hiện tại

###### 1. Phương pháp dựa trên Global features

- **Phương pháp:** Dùng toàn bộ khuôn mặt → CNN backbone (VGG, ResNet, CNN cơ bản)
- **Nhược điểm:**
  - feature extract có nhiều thông tin dư thừa → giảm tập trung vào vùng quan trọng



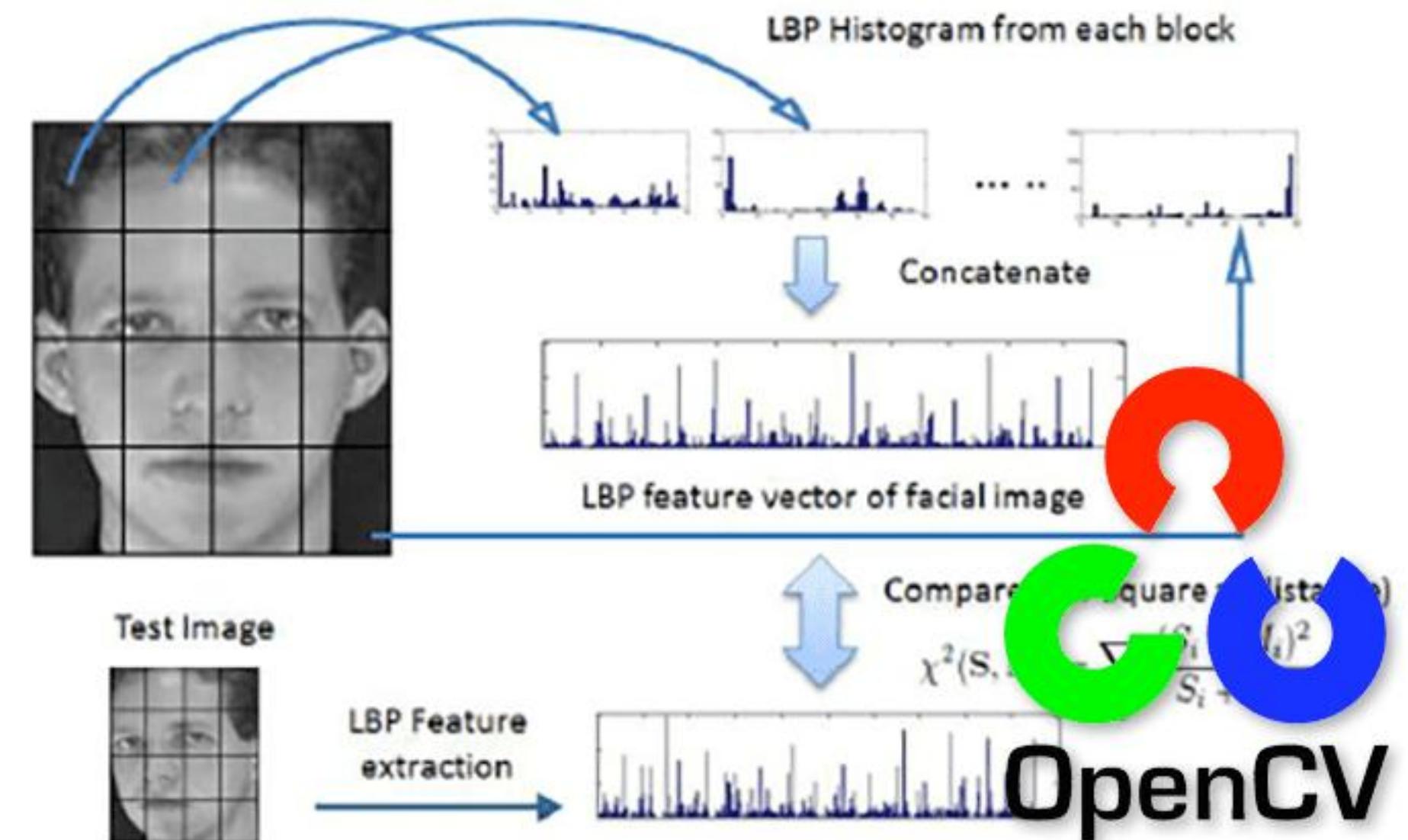
# 3. Phương pháp đề xuất

## 3.1. Tổng quan

### Các phương pháp hiện tại

#### 1. Phương pháp dựa trên local features

- **Phương pháp:** Chia mặt thành nhiều vùng → Trích xuất từng vùng với CNN
- **Nhược điểm:**
  - feature extract không biểu diễn tổng thể khuôn mặt
  - Mất mối liên hệ giữa các vùng mặt



### **3. Phương pháp đề xuất**

#### **3.1. Tổng quan**

##### **Các phương pháp hiện tại**

###### **1. Phương pháp song song Global + Local + Direct Fusion**

- **Phương pháp:**
  - Mô hình trích xuất: global vector, local vector
  - Nối (concatenate) lại trực tiếp → đưa vào Dense → softmax
- **Nhược điểm:**
  - Khi nối trực tiếp, mô hình không học được mức độ quan trọng của từng feature
  - 1 loại feature sẽ chiếm ưu thế làm giảm hiệu quả tổng hợp

### 3. Phương pháp đề xuất

#### 3.1. Tổng quan

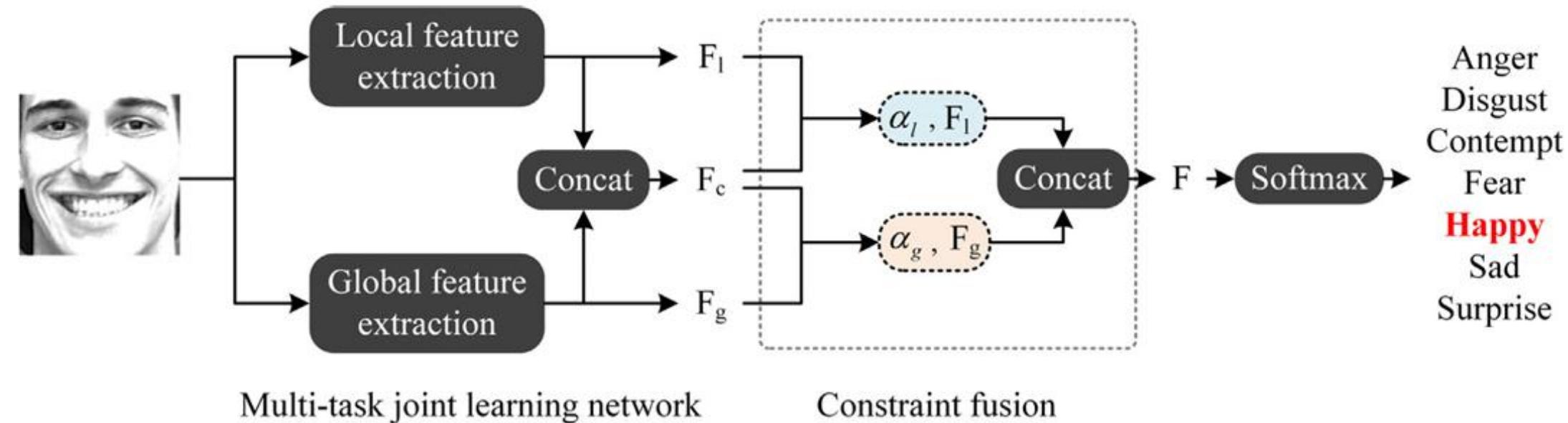


Fig. 1. The overview of CFNet.

- **Mạng chia làm 2 phần:**
  - **Multi-task join learning network:** trích xuất 3 feature của mặt bằng cách học song song 3 nhiệm vụ:
    - phân loại từ local feature
    - phân loại từ global feature
    - phân loại từ combined feature
  - **Constraint fusion:** học trọng số để kết hợp các feature động dựa trên cosine similarity
- **Cách trainning:** Tiến hàng training độc lập 2 mạng
  - GD1 : training mạng Multi-task
  - GD2: training mạng constraint fusion

# 3. Phương pháp đề xuất

## 3.2. Multi-task joint learning network

- **Phương pháp:**

- 3 nhánh học song song
- Mỗi nhánh có 1 loss

- **3 đặc điểm chính của mạng:**

- **Global branch và local branch :** Cách mạng trích xuất feature
- **Weight share :** Cách tăng sự liên kết giữa local feature và global feature
- **Multi-loss:** Cách họ dạy cho mạng học

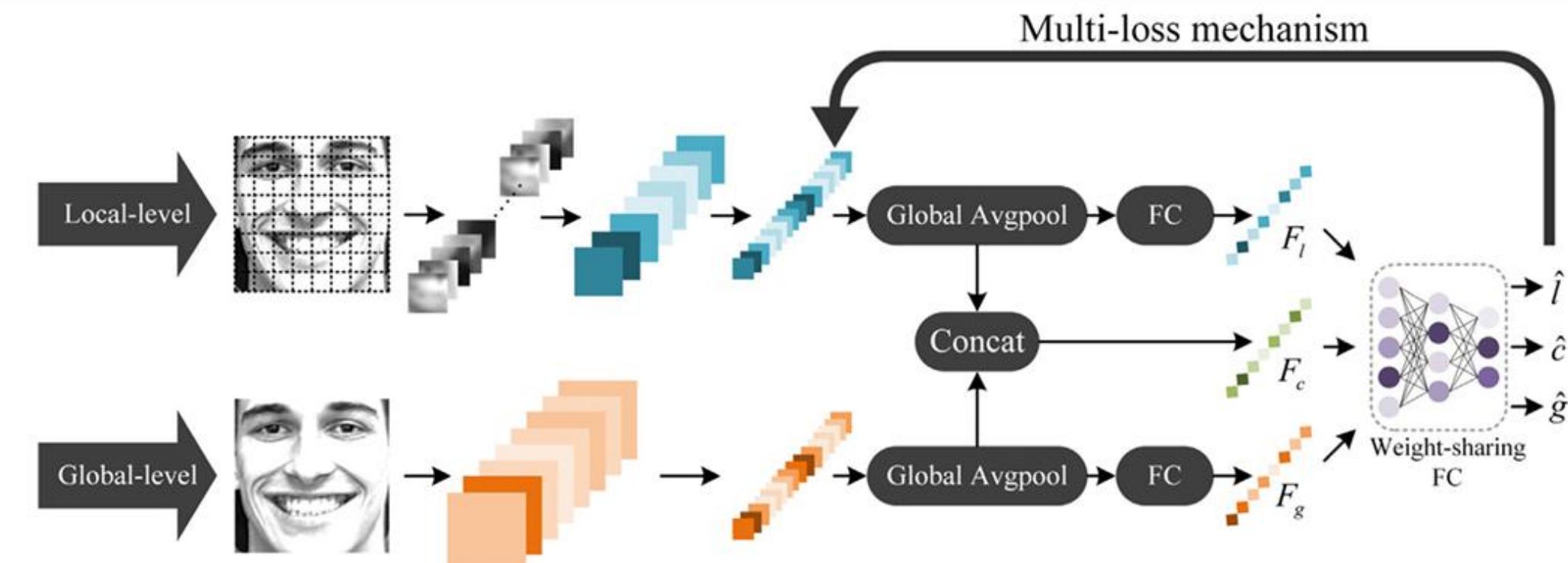
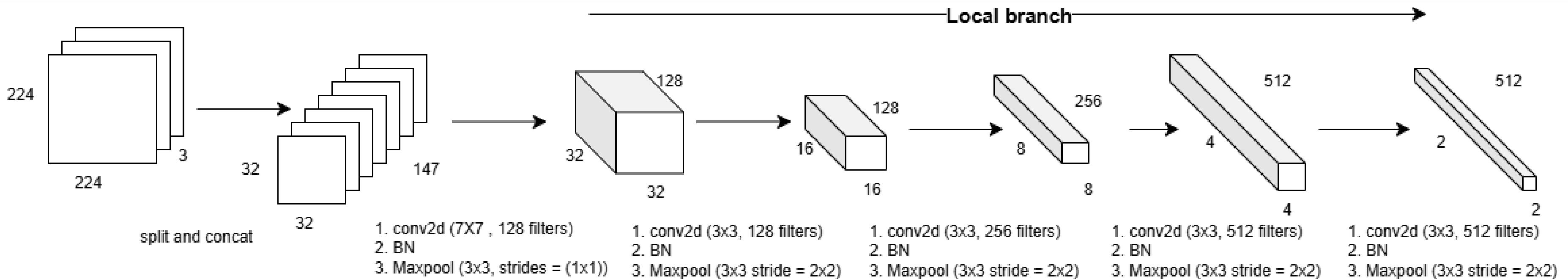


Fig. 2. The workflow of the multi-task joint learning network.

### 3. Phương pháp đề xuất

#### 3.2. Multi-task join learning network

##### Local branch

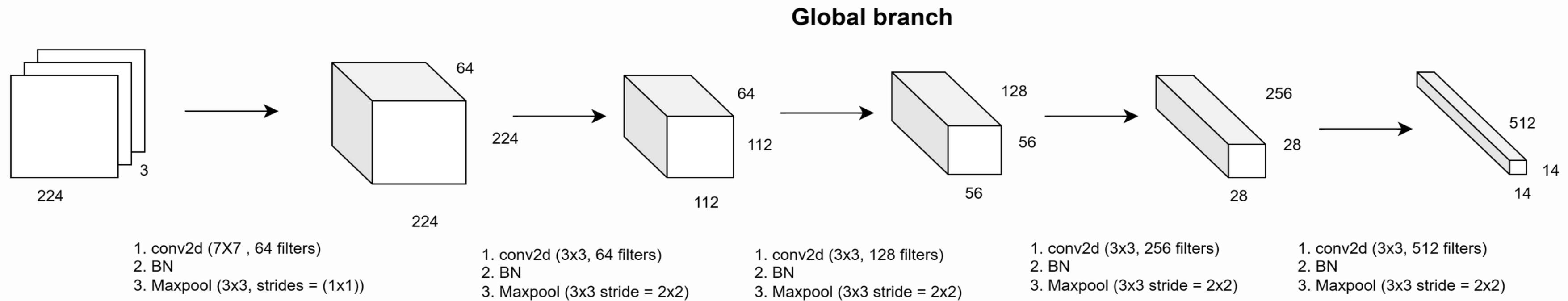


- Chia ảnh thành 49 patch nhỏ (mỗi patch 32x32) làm nổi bật biểu đạt cảm xúc ở các vùng quan trọng của khuôn mặt
  - Sau đó dùng conv 3x3 để trích rút đặc trưng cục bộ
  - Đi qua lớp Batch Norm để cải thiện hiệu quả và độ ổn định khi huấn luyện
  - Dùng max pooling để làm nổi bật thông tin quan trọng
- => Kết quả cuối cùng ta trích rút được feature quan trọng từ các vùng nhỏ trên khuôn mặt

# 3. Phương pháp đề xuất

## 3.2. Multi-task joint learning network

### Global branch



- Ảnh không chia nhỏ thành từng patch như nhánh local mà đưa thẳng vào các block conv + BN + maxpool để trích xuất đặc trưng
  - Mạng global có kiến trúc tương tự nhánh local để feature local và global nằm trong cùng 1 không gian học
    - global feature không học theo cách hoàn toàn khác với local feature
    - Chỉ khác ở phạm vi nhìn ảnh : Local nhìn các vùng nhỏ, Global nhìn toàn bộ khuôn mặt
- => Kết quả feature trích xuất ra chứa thông tin cảm xúc tổng thể của toàn bộ khuôn mặt

### 3. Phương pháp đề xuất

#### 3.2. Multi-task joint learning network

- Để giảm độ phức tạp, CFNet sử dụng Global Average Pooling (GAP) để nén đầu ra của nhánh local  $l \in 2 \times 2 \times 512$  và global  $g (14, 14, 512)$  thành vector: (512,)
- $L = GAP(l), G = GAP(g)$

=> GAP giống như tóm tắt đặc trưng

- Sau khi GAP tóm tắt đặc trưng thì cần phải có lớp FC để đưa ra quyết định “đặc trưng đó quan trọng thế nào cho từng cảm xúc”

=> cần 1 lớp FC để chiếu sang không gian phân loại : Dense(1024)

- $F_l = FC(L)$
- $F_g = FC(g)$
- Trong quá trình trích xuất,  $G$  và  $L$  được nén bởi GAP. So với  $F_g$  và  $F_l$ ,  $G$  và  $L$  chưa qua FC nên giữ được thông tin cảm xúc nguyên bản tốt hơn. CFNet gộp trực tiếp chúng bằng:  $F_c = concat(L, G)$

Global average pooling

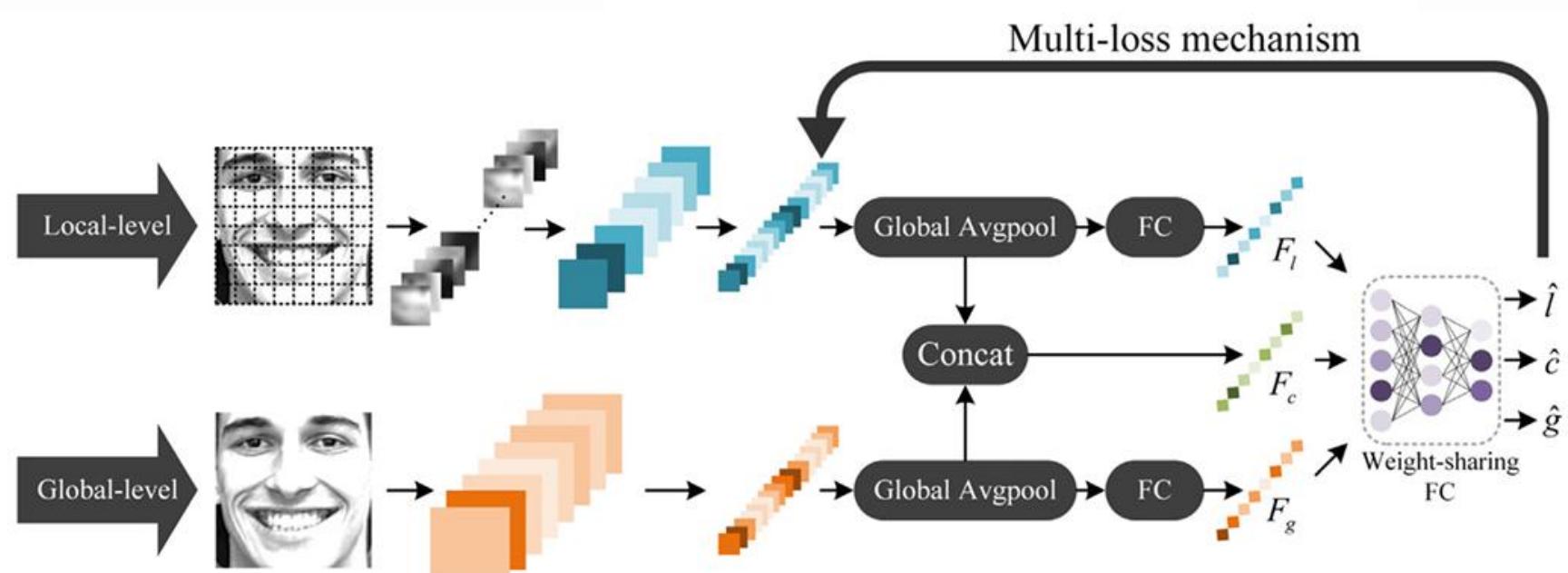
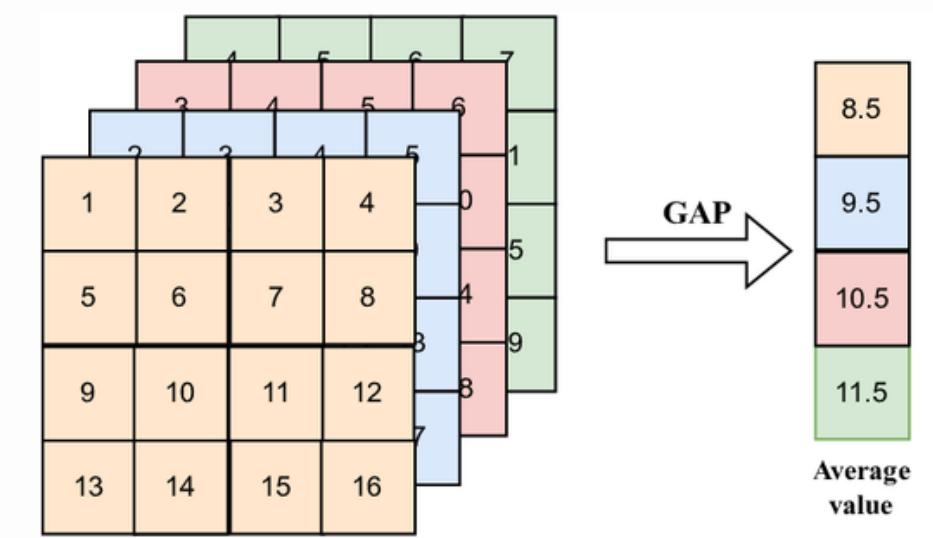


Fig. 2. The workflow of the multi-task joint learning network.

# 3. Phương pháp đề xuất

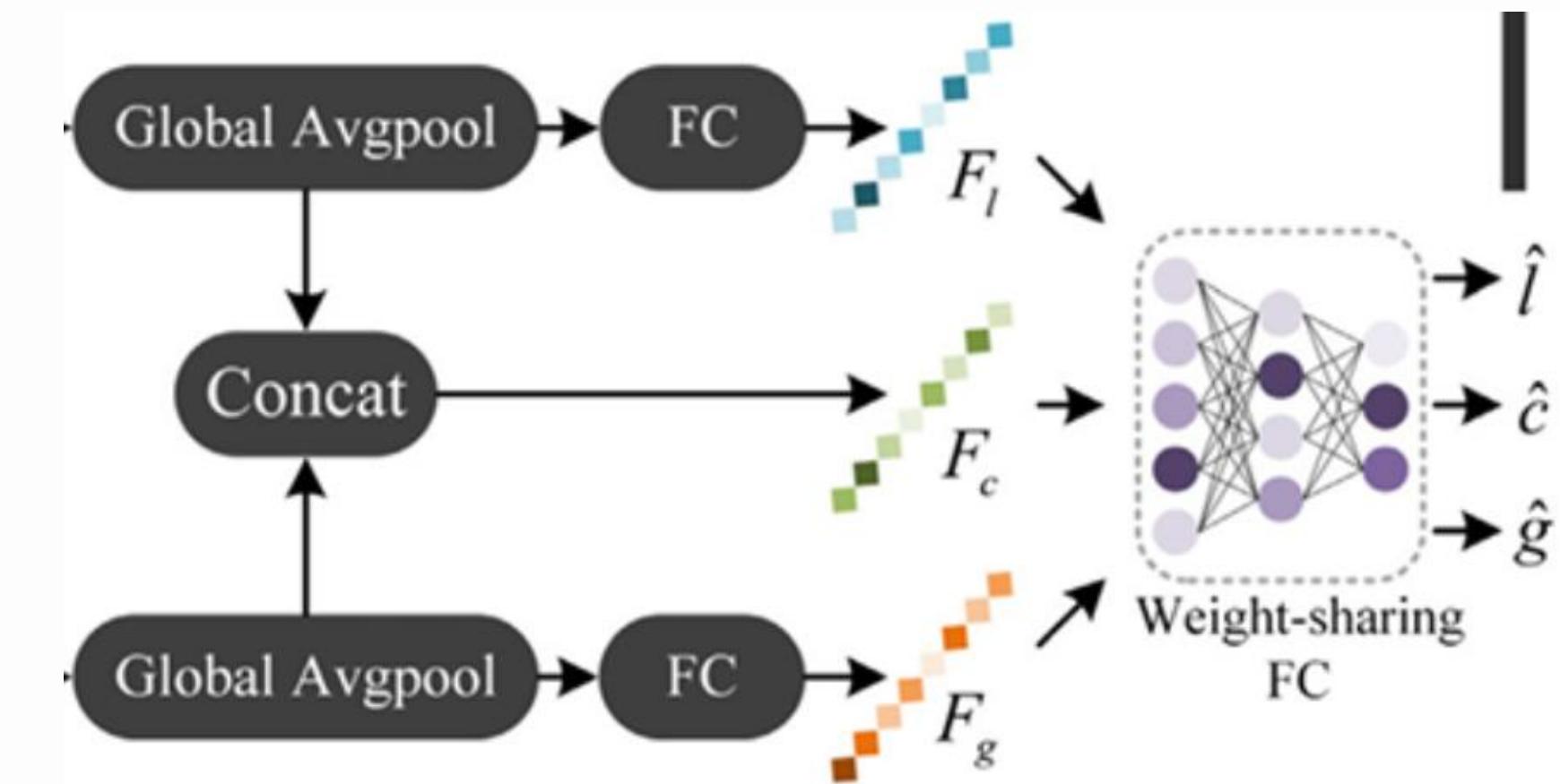
## 3.2. Multi-task join learning network

### Weight share : Mạng phân loại

- Một mạng FC dùng chung để phân loại cho cả
  - $F_g$ : global
  - $F_l$ : local
  - $F_c$ : direct fusion
- Nhờ chia sẻ trọng số, mạng học đồng thời 3 loại đặc trưng

→ **Tăng khả năng khái quát hóa, và cho phép đặc trưng local-global liên kết và hỗ trợ lẫn nhau để phân loại cảm xúc**

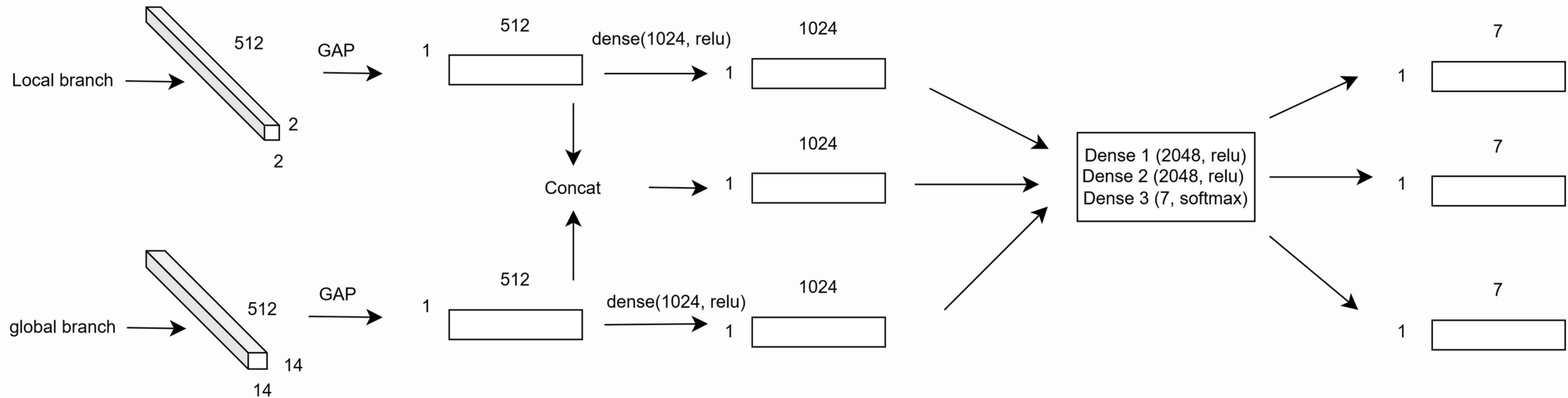
- Mạng weight share giúp : “Happy” phải được hiểu giống nhau dù nhìn:
  - Toàn mặt (global)
  - Từng vùng (local)
  - Ghép lại (fusion)



### 3. Phương pháp đề xuất

#### 3.2. Multi-task join learning network

Weight share : Mạng phân loại



### 3. Phương pháp đề xuất

#### 3.2. Multi-task join learning network

##### Multi-loss

- Vì huấn luyện song song cả 3 nhiệm vụ:
  - Phân loại local feature
  - Phân loại global feature
  - Phân loại direct fusion feature

=> Cần có 3 loss cho từng nhiệm vụ đồng thời có cơ chế kết hợp cân bằng 3 loss để ra loss tổng cho mạng weight share:

- Loss cho từng nhánh : cross entropy
- Loss tổng :  $Loss_M = -\frac{1}{N} \sum_{i \in N} y_i \ln \hat{l}_i + y_i \ln \hat{g}_i + y_i \ln \hat{c}_i,$

=> Ba loss có trọng số bằng nhau → đảm bảo global và local đều được chú trọng.

### 3. Phương pháp đề xuất

#### 3.1. Contrain fusion

- **Global và local có lợi thế khác nhau cho từng biểu cảm :**

- Có những biểu cảm được hình thành chủ yếu từ 1 vùng duy nhất
- Có biểu cảm lại hình thành từ sự phối hợp giữa nhiều vùng khuôn mặt

- **Việc kết hợp 2 đặc trưng này có thể giúp cải thiện độ chính xác nhận dạng:**

- Với 1 cảm xúc, mỗi người thì cách kích hoạt biểu cảm khác nhau → Trong cùng 1 cảm xúc độ chính xác của local feature và global feature có thể khác nhau.
- Nếu gộp trực tiếp local feature và global feature theo tỷ lệ cố định → Đặc trưng kém hiệu quả hơn có thể làm nhiễu đặc trưng tốt



=> **Đề xuất cơ chế hợp nhất có ràng buộc (constraint fusion)**

### 3. Phương pháp đề xuất

#### 3.1. Contrain fusion

- Ý tưởng chính:
  - Đánh giá động mức độ hiệu quả của global và local features trên từng mẫu
  - Sử dụng cosine similarity giữa local feature và global feature với feature fusion → chuyển độ tương đồng thành trọng số đánh giá độ hiệu quả của feature vào nhận diện biểu cảm

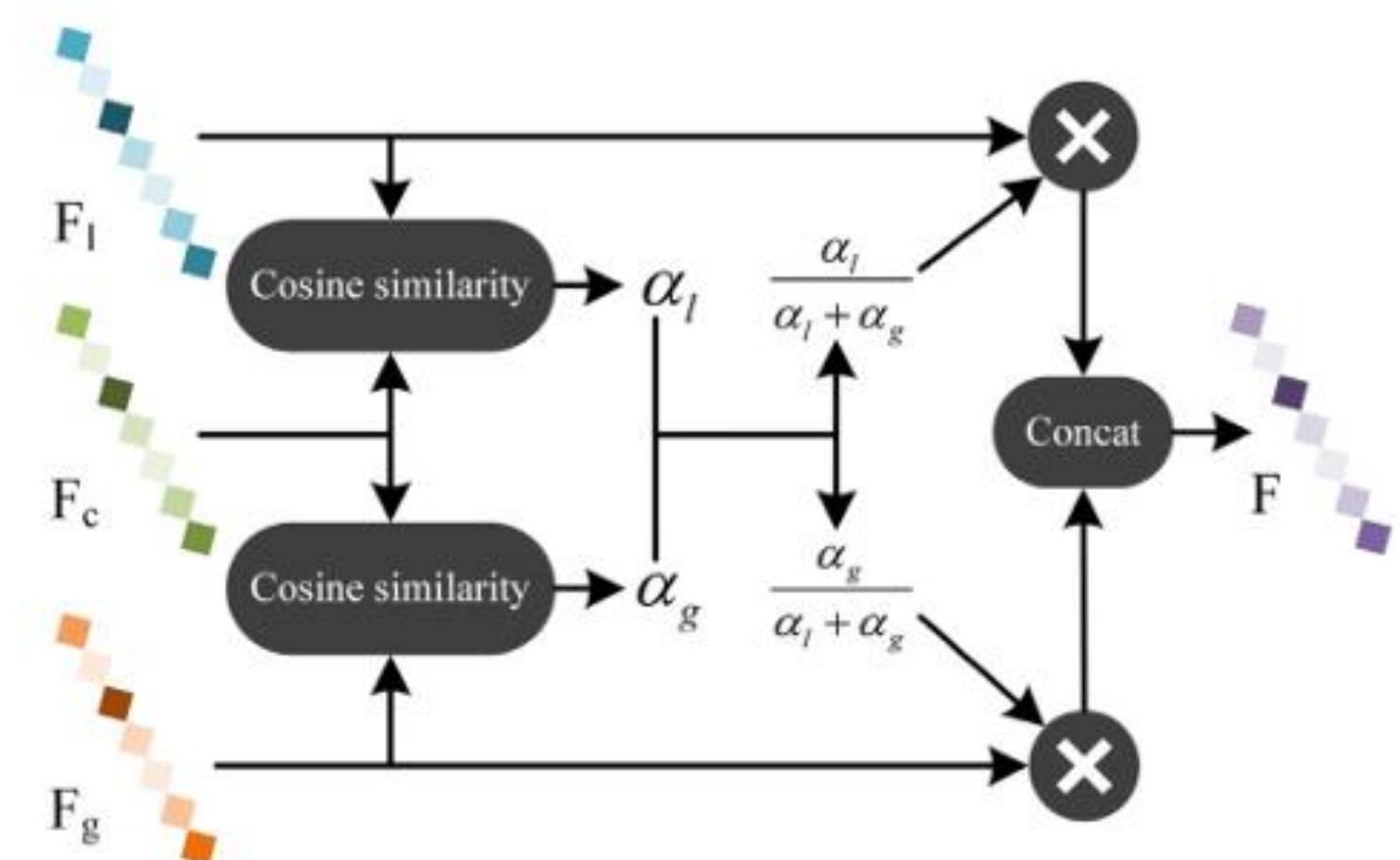


Fig. 3. The principle of constraint fusion.

## 4. Kết quả thực nghiệm

### 4.1. Độ chính xác trên các dataset

#### CK+

- **Accuracy:** 99.07%
- **Samples:** 327 images in-the-lab
- **Subjects:** 118
- 7 basic emotions
- **Best:** Contempt, Disgust, Fear, Happiness (100%)
- **Confusion:** Anger ↔ Sadness

#### MMI

- **Accuracy:** 84.62%
- **Samples:** 208 videos in-the-lab
- **Subjects:** 31
- 6 basic emotions
- **Best:** Happiness
- **Worst:** Sadness (10 trường hợp bị nhầm)
- **Confusion:** Sadness ↔ Anger ↔ Disgust

#### RAF-DB

- **Accuracy:** 87.52%
- **Samples:** 15,338 images in-the-wild
- 7 basic emotions
- Hiệu suất đồng đều trên các loại cảm xúc
- Khả năng thích ứng cao trong các cảnh phức tạp

## 4. Kết quả thực nghiệm

### 4.2. Hiệu quả của hợp nhất có ràng buộc

Phương pháp	CK+	RAF-DB
<b>Chỉ đặc trưng cục bộ</b>	96.61%	81.29%
<b>Chỉ đặc trưng toàn cục</b>	97.23%	85.07%
<b>Hợp nhất trực tiếp</b>	98.46%	86.70%
<b>Hợp nhất có ràng buộc</b>	99.07%	87.52%

→ **Hợp nhất có ràng buộc** giúp tăng vượt trội 0.6–2.5% so với mọi kiểu đặc trưng khác.

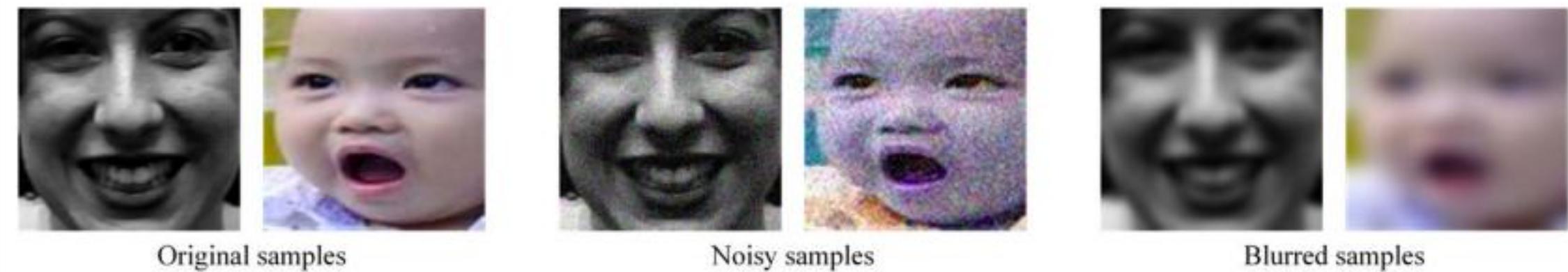
## 4. Kết quả thực nghiệm

### 4.3. Đánh giá tính ổn định - ảnh nhiễu và mờ

→ CFNet chỉ giảm nhẹ so với ảnh gốc ( $99.07\% \rightarrow 96.92\%$ ), trong khi các mạng pre-trained sụt giảm đáng kể ( $90.20\% \rightarrow 46.83\%$ )

→ CFNet duy trì độ chính xác ổn định 80-83% trên RAF-DB ngay cả khi kernel size tăng

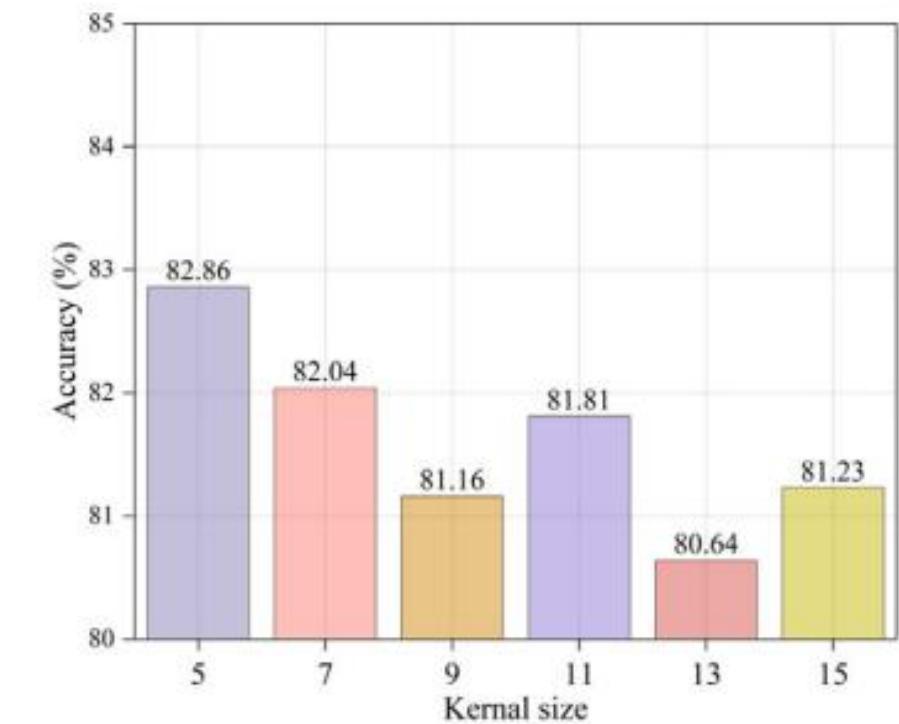
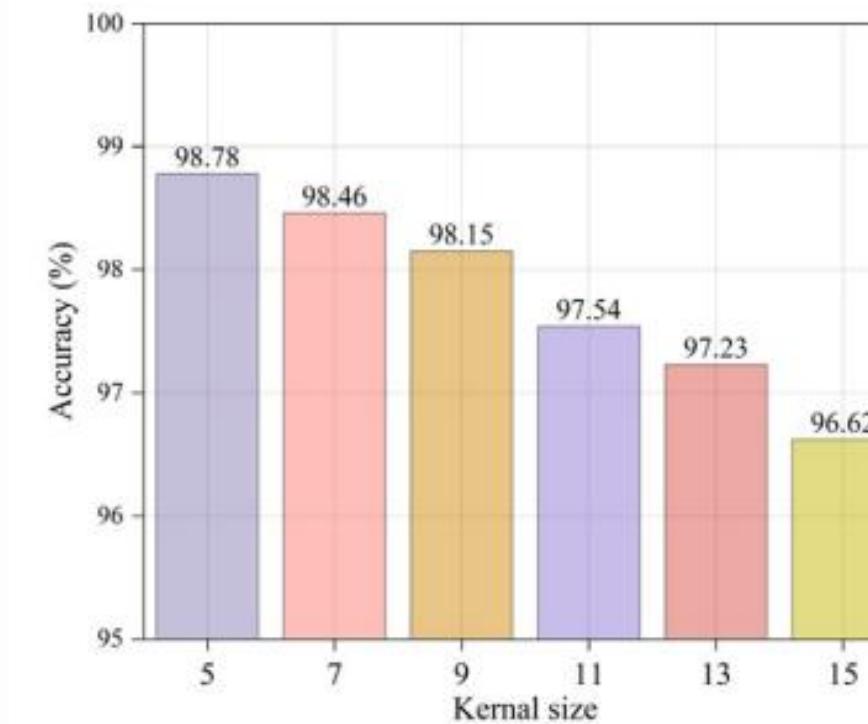
→ Thể hiện độ ổn định và khả năng thích ứng cao.



Original samples

Noisy samples

Blurred samples



(a) Performance on CK+.

(b) Performance on RAF-DB.

Methods	Accuracy on CK+ (%)		Accuracy on RAF-DB (%)	
	Blurred	Noisy	Blurred	Noisy
VGG16 [57]	83.29%	84.83%	46.67%	44.06%
VGG19 [57]	83.94%	85.75%	45.90%	40.97%
ResNet50 [58]	80.01%	85.03%	46.32%	46.56%
DenseNet121 [59]	85.07%	87.78%	47.04%	48.10%
DenseNet169 [59]	88.13%	89.25%	47.83%	47.02%
DenseNet201 [59]	88.01%	90.20%	49.40%	46.83%
<b>CFNet (Ours)</b>	<b>96.62%</b>	<b>96.92%</b>	<b>81.23%</b>	<b>81.16%</b>

Blurred: the kernel size is  $15 \times 15$ .

## 4. Kết quả thực nghiệm

### 4.4. Hiệu quả tính toán

Methods	Input size	FLOPs ( $\times 10^8$ )	# Param (M)
DMA-CNN [15]	$224 \times 224$	4.47	237.33
CFNet (Ours)	<b><math>224 \times 224</math></b>	<b>0.28</b>	<b>13.91</b>

→ **FLOPs của CFNet** ( $0.28 \times 10^8$ ) thấp hơn nhiều so với DMA-CNN ( $4.47 \times 10^8$ ), cho thấy CFNet dễ triển khai hơn với thời gian training ít hơn

→ **Global average pooling layer** giảm **parameters** từ 222.58M xuống 13.91M

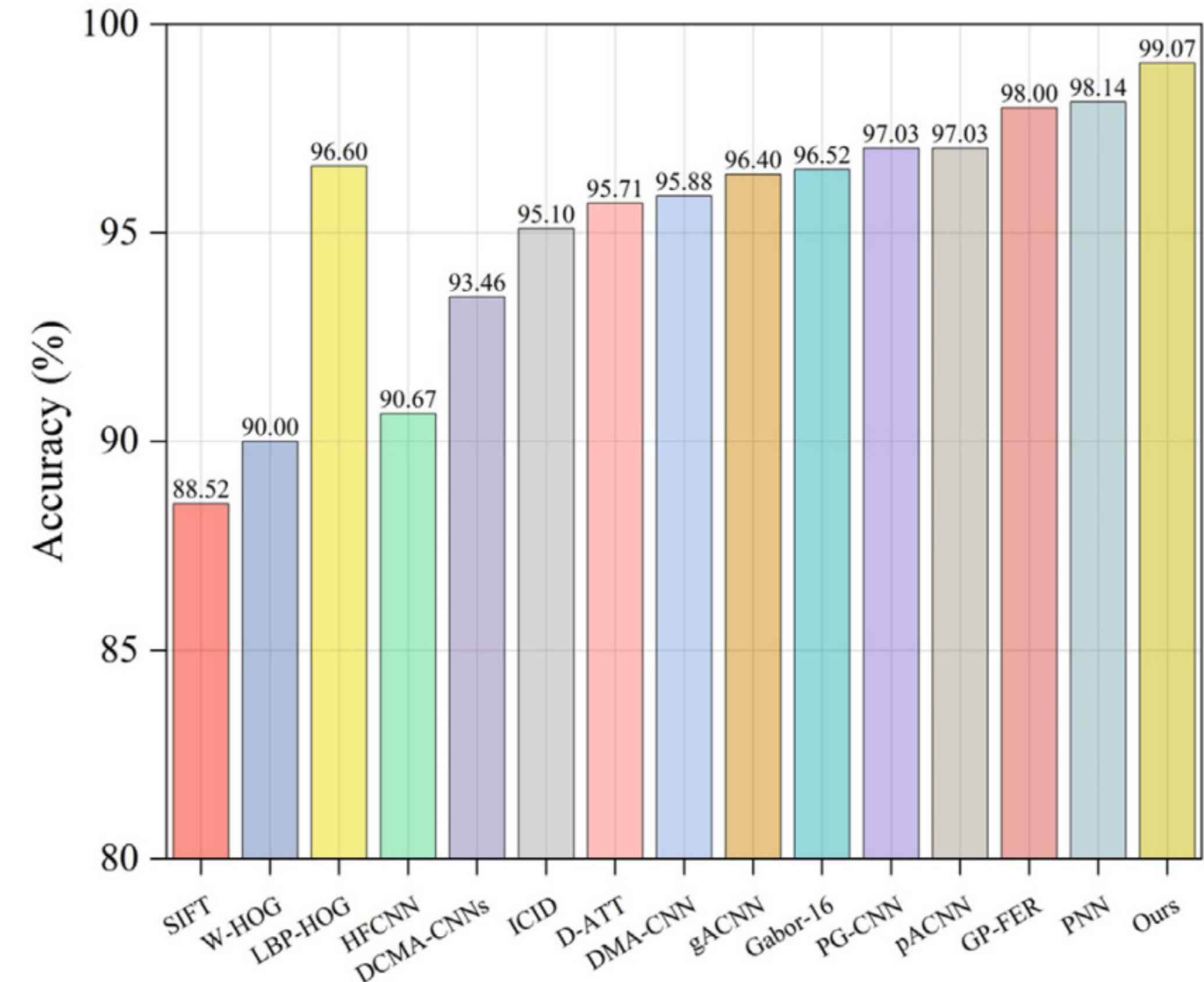
## 4. Kết quả thực nghiệm

### 4.5. So sánh với CK+

→ CFNet có độ chính xác cao nhất

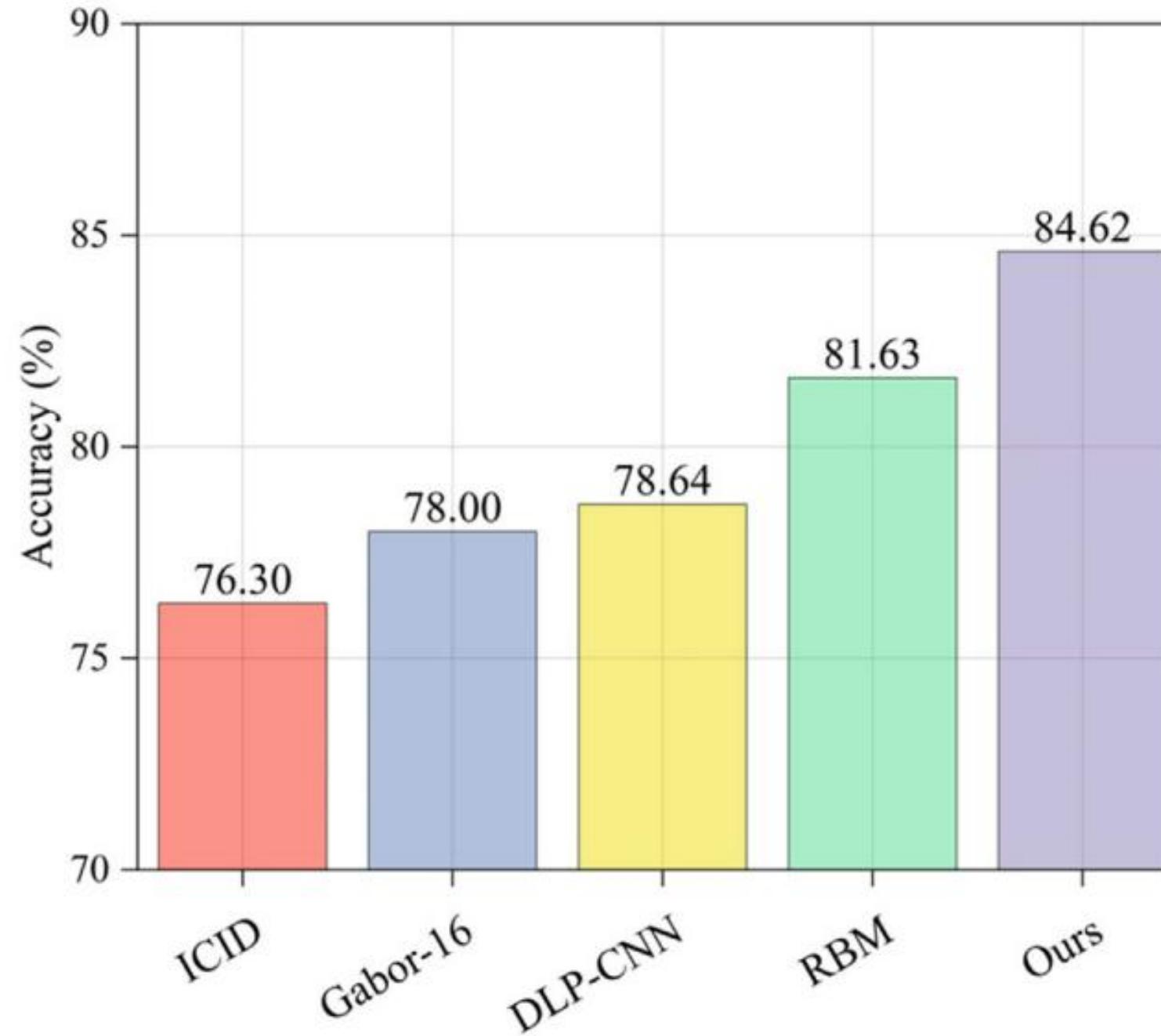
99.07%:

- Cải thiện ~0.93% so với PNN (tốt nhất trước đó) nhờ tự động gán trọng số thích ứng
- Cải thiện ~2.55% so với Gabor-16 (vì phương pháp global thường bị ảnh hưởng bởi thông tin dư thừa)
- Cải thiện ~2.04% so với PG-CNN và pACNN (vì phương pháp local bỏ qua mối liên hệ giữa vùng)

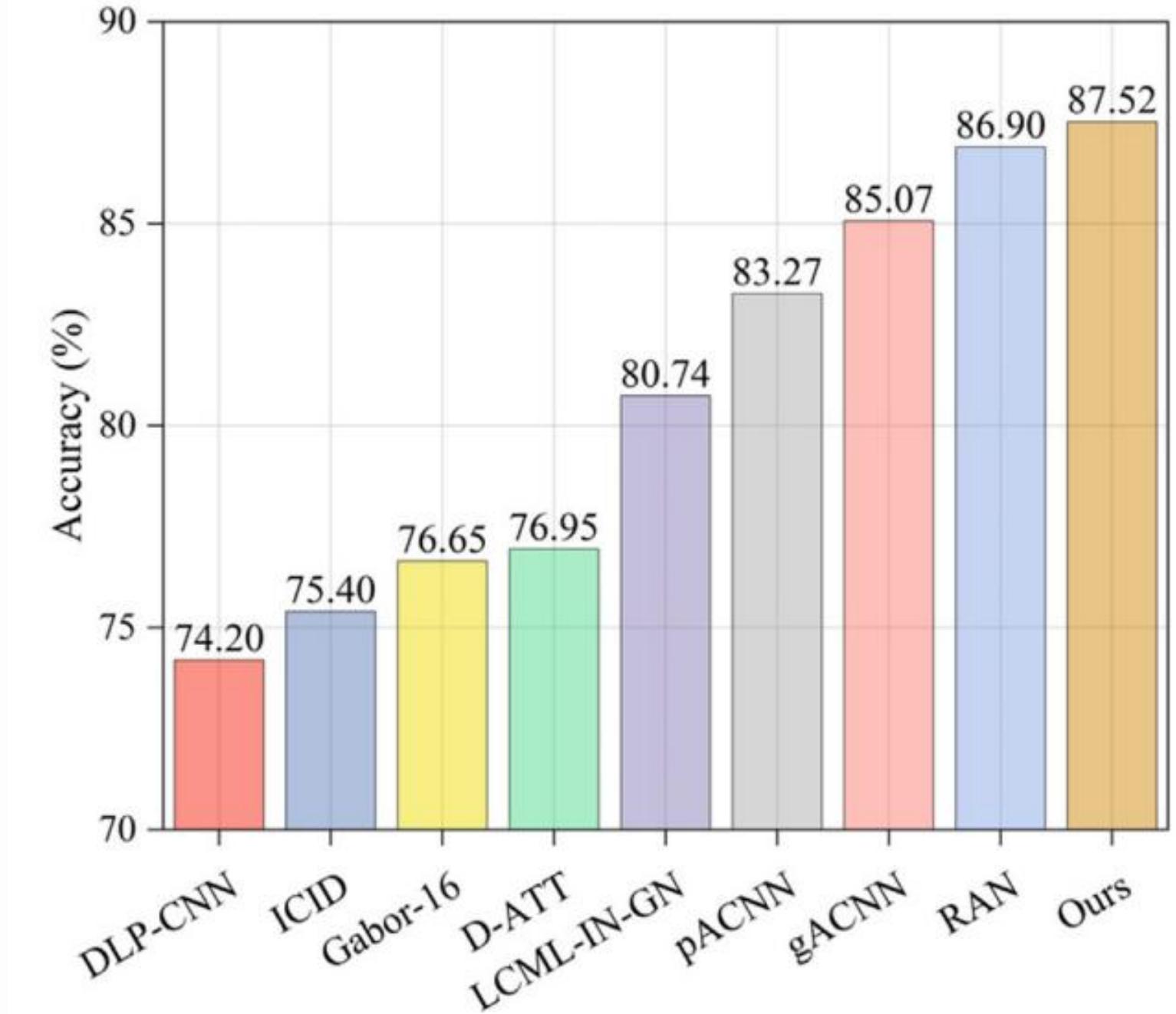


## 4. Kết quả thực nghiệm

### 4.5. So sánh trên MMI và RAF-DB



**Fig. 9.** Performance comparison on MMI.



**Fig. 12.** Performance comparison on RAF-DB.