시각 변환기와 주의 선택적 융합을 통한 얼굴 표정 인식

Fuyan Ma, Bin Sun, IEEE 회원, Shutao Li, IEEE 펠로우

초록—야외에서의 얼굴 표정 인식(FER)은 제약 없는 조건에서 폐색, 다양한 머리 자세, 얼굴 변형 및 모션 블러로 인해 매우 어렵습니다. 지난 수십 년 동안 자동 FER에서 상당한 진전이 있었지만 이전 연구는 주로 실험실 제어 FER을 위해 설계되었습니다. 실제 세계의 폐색, 다양한 머리 자세 및 기타 문제는 이러한 정보 부족 영역과 복잡한 배경으로 인해 FER의 어려움을 확실히 증가시킵니다. 이전의 순수한 CNN 기반 방법과 달리 얼굴 이미지를 시각적 단어 시퀀스로 변환하고 글로벌 관점에서 표정 인식을 수행하는 것이 가능하고 실용적이라고 주장합니다. 따라서 두 가지 주요 단계를 통해 야외에서 FER을 해결하기 위해 Visual Transformers with Feature Fusion(VTFF)을 제안합니다. 첫째, 2분기 CNN에서 생성된 두 가지 종류의 피처 맵을 활용하기 위한 주의 선택적 융합(ASF)을 제안합니다. ASF는 글로벌-로컬 주의와 여러 피처를 융합하여 차별적인 정보를 포착합니다. 융합된 특징 맵은 그런 다음 평평하게 만들어 시각적 단어 시퀀스로 투사합니다. 둘째, 자연어 처리에서 Transformers의 성공에 영감을 받아 글로벌 셀프 어텐션을 사용하여 이러한 시각적 단어 간의 관계를 모델링하는 것을 제안합니다. 제안된 방법은 세 가지 공개된 야생 얼굴 표정 데이터 세트(RAF-DB, FERPlus 및 AffectNet)에서 평가됩니다. 동일한 설정에서 광범위한 실험을 통해 우리 방법이 다른 방법보다 우수한 성능을 보이며 RAF-DB에서 88.14%, FERPlus에서 88.81%, AffectNet에서 61.85%로 새로운 최첨단 상태를 설정한다는 것을 보여줍니다. CK+에 대한 교차 데이터 세트 평가는 제안된 방법의 유망한 일반화 기능을 보여줍니다.

색인 용어—야생에서의 얼굴 표정 인식, 글로벌-로컬 주의, 트랜스포머, 글로벌 자기 주의

기음

1 서론

인간의 감정 상태를 이해하는 것은 심리학과 컴퓨터 과학과 같은 다양한 연구 분야에 걸친 학제간 분야인 감정 지능을 개발하는 기본적인 과제입니다. 얼굴 표정은 인간이 감정 상태와 의도를 전달하는 가장 자연스럽고 강력하며 보편적인 신호 중 하나입니다[1], [2]. 얼굴 표정 인식(FER) 시스템은 인간-로봇 상호작용(HRI), 정신 건강 평가 및 운전자 피로 모니터링을 포함한 다양한 응용 분야가 있습니다. 따라서 FER 개발을 촉진하기 위해 수많은 연구 노력이 투자되었습니다.

그림 1에서 보듯이, 실제 FER의 과제는 주로 폐색, 다양한 머리 포즈, 얼굴 변형, 모션 블러, 불충분한 정성적 데이터 등에서 비롯되며, 이는 얼굴 모양의 상당한 변화로 이어진다. 이러한 예상치 못한 문제는 확실히 표정 인식의 어려움을 증가시킨다. 머신 러닝, 특히 딥 러닝의 발전으로 연구자들은 지난 수십 년 동안 FER에서 큰 진전을 이루었다. 딥 러닝 시대 이전에는 기존 FER 방법에서 주로 수작업으로 만든 특징과 얕은 학습(예: 지향성 기울기 히스토그램(HOG)[3], 로컬 이진 패턴(LBP)[4],[5], 음이 아닌 행렬 분해(NMF)[6] 및 희소 표현[7])을 사용했다. 데이터 기반 기술의 인기로 인해 딥 러닝 기반 방법은 기존 방법을 크게 능가하여 최첨단 FER 성능을 달성했습니다(예:[8],[9],[10],[11]).

제한된 FER에 대한 학습된 표현의 상당한 성공에도 불구하고, 실제 환경에서의 FER 성능은 여전히 만족스럽지 못합니다. 제안된 알고리즘의 대부분은 CK +[12], MMI[13] 및 Oulu-CASIA [14]와 같은 실험실에서 수집한 데이터 세트에 구현됩니다. 이러한 알고리즘은 제어된 이미지가 최소한의 조명 변화와 제한된 폐색으로 정면이기 때문에 이러한 실험실에서 수집한 FER 데이터 세트에서 완벽하게 작동합니다. 그러나 RAF-DB[8], FERPlus[15] 및 AffectNet[16]과 같은 실제 FER 데이터 세트에서는 성능이 크게 저하됩니다. 실제 FER 데이터 세트와 비교할 때 실험실에서 수집한 FER 데이터 세트의 이미지 수는 비교적 적습니다. 특히, 실제 환경에서의 FER에 대한 합성곱 신경망(CNN) [17]은 일반적으로 실제 응용 프로그램에 대한 일반화를 보장하기 위해 충분한 훈련 얼굴 이미지가 필요합니다. FER에 대한 대부분의 공개 데이터 세트에는 충분한 양의 이미지가 없습니다.

- *Fuyan Ma and Bin Sun are with the College of Electrical and Information Engineering, Changsha 410082, China, and also with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha 410082, China. E-mail: {mafuyan, sunbin611}@hnu.edu.cn.*
- *Shutao Li is with the College of Electrical and Information Engineering, Changsha 410082, China, and with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Changsha 410082, China, and also with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha 410082, China. E-mail: shutao_li@hnu.edu.cn.*

Manuscript received 7 April 2021; revised 13 October 2021; accepted 18 October 2021. Date of publication 26 October 2021; date of current version 31 May 2023.
This work was supported by the National Key Research and Development Project under Grant 2018YFB1305200, the National Natural Science Fund of China under Grant 62171183, and the Key-Area Research and Development Plan of Guangdong Province under Grant 2018B010107001.
(Corresponding author: Bin Sun.)
Recommended for acceptance by M. Mahoor.
Digital Object Identifier no. 10.1109/TAFFC.2021.3122146

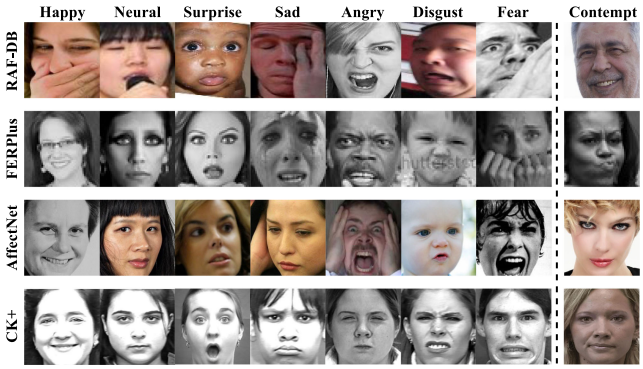


그림 1. RAF-DB, FERPlus, AffectNet 및 CK+의 샘플. 변형된 머리 포즈, 폐색 및 기타 제약 없는 조건은 위의 이미지에서 볼 수 있습니다. RAF-DB는 7개의 기본 표현식으로 주석이 달려 있고 FERPlus, AffectNet, CK+는 경멸 범주를 포함하여 8개의 표현식 레이블이 달려 있습니다.

훈련을 위해서입니다. 따라서 야생에서 FER의 성능은 제약이 없는 조건뿐만 아니라 사용 가능한 데이터 볼륨에 의해서도 제한을 받습니다.

제약이 없는 조건에서는 다양한 배경, 조명 및 머리 자세가 매우 흔하며, 이는 얼굴 표정과 관련이 없습니다[18]. 이러한 이미지에서 표정을 직접 인식하는 것은 큰 과제입니다. 복잡한 배경과 얼굴이 아닌 영역을 제거하려면 심층 신경망이 의미 있는 특징을 학습하도록 훈련하기 전에 얼굴을 감지하는 것이 필수적입니다. MTCNN[19] 및 Dlib[20]과 같은 여러 얼굴 감지기가 복잡한 시나리오에서 얼굴을 감지하는 데 사용됩니다. 비교적 정확한 얼굴 영역을 얻은 후, FER 정확도를 개선하고 표정 인식 알고리즘의 일반화 능력을 향상시키기 위한 다양한 방법이 제안되었습니다. 대부분의 연구자들은 실제 시나리오에서 FER에 대한 두 가지 주요 장애물인 폐색 인식[21], [22] 및 다중 뷰 FER에 대한 방법을 설계합니다. 아주 최근에, 주의 모델이 FER에 성공적으로 적용되어 의미 있는 영역을 탐색했습니다. Li et al.[23]은 폐색과 함께 표정 인식을 위한 경로 수준 주의를 통합하는 패치 게이트 CNN을 제안했습니다. 또한, 주의 모델은 FER에 성공적으로 적용되어 의미 있는 영역을 탐색했습니다. [23]과 유사하게 [10], [24], [25]와 같은 여러 방법도 주의와 유사한 메커니즘을 사용하여 가장 차별적인 특징에 초점을 맞춰 FER 정확도를 개선했습니다. 정면이 아닌 얼굴 이미지는 실제 시나리오에서 압도적으로 흔합니다. 이전 방법은 종종 정면이 아닌 얼굴 이미지의 표정 인식을 다중 뷰 FER 문제로 처리합니다. Zheng[26]은 그룹 희소 감소 순위 회귀를 기반으로 표정 인식에 가장 기여하는 얼굴의 최적 하위 영역을 선택하는 것을 제안했습니다. Liu et al.[27]은 다중 채널 특징 추출, 다중 스케일 특징 융합 및 포즈 인식 인식의 세 부분으로 다중 뷰 FER을 해결하는 것을 제안했습니다. 또한 생성적 적대 네트워크(GAN)도 다중 뷰 FER에 적용되었습니다. GAN 기반 방법[28][29]은 FER에 대한 훈련 세트를 확대하기 위해 큰 머리 자세 변형을 가진 얼굴 이미지 크기를 합성할 수 있습니다. 특히 Zhang et al.[30]은 기하 정보를 활용하여 얼굴 랜드마크 세트와 포즈 불변 얼굴 표정 인식을 위한 GAN 기반 종단 간 모델을 제안했습니다.

GAN을 기반으로 한 비지도 교차형 얼굴 표정 인식을 위해.

예를 들어 얼굴 이미지에서 다양한 포즈와 폐색을 취하여 우리 방법의 동기를 설명합니다. 다양한 포즈와 크기 조정 작업은 정면 얼굴에 비해 얼굴 변형을 초래하며, 이는 순서가 없는 하위 영역으로 볼 수 있습니다. 얼굴 이미지의 폐색은 정보가 부족한 하위 영역을 초래합니다. 자연어와 유추하여 야생에서의 FER은 다른 방식으로 처리할 수 있습니다. 흥미로운 현상은 우리가 문장을 이해할 수 있다는 것입니다. 즉, "나는 내가 들은 것을 이해하지 못하는 것을 이해할 수 없었습니다. 인간의 마음의 힘입니다." 가장 가능한 이유 중 하나는 우리 인간의 인지 모드가 일반적으로 이 문장을 전체적으로 보고 생각하도록 강요하기 때문입니다. 또한 자연어 처리(NLP)의 이미지 설명 작업과 마찬가지로 여러 시각적 단어를 사용하여 이미지의 감정 상태를 설명할 수 있습니다. 이러한 인지 모드에서 영감을 얻어, 우리는 전역적 관점에서 시각적 단어 시퀀스로 얼굴 표정을 인식하는 것이 가능하고 효과적이라고 가정합니다. 따라서 우리는 야생에서 강력한 얼굴 표정 인식을 위해 특징 융합을 갖춘 시각적 변환기(VTFF)를 제안합니다. 우리의 방법은 하이브리드 특징 추출을 참조하여 시각적 단어의 표현을 더욱 풍부하게 하기 위해 LBP 특징과 CNN 특징을 결합합니다. 우리가 LBP 특징을 사용하는 이유는 얼굴의 작은 움직임을 포착하고 이미지 텍스처 정보를 추출할 수 있기 때문입니다. 우리는 주의 특징 융합(ASF)을 설계하여 LBP 특징과 CNN 특징을 적응적으로 통합합니다. ASF는 두 종류의 특징 간의 전역적 관계와 지역적 관계를 모두 집계하여 인식 성능을 효과적으로 개선할 수 있습니다. 우리는 특징 맵을 평탄화하고 투영하여 융합된 특징 맵을 시각적 단어 시퀀스로 간단히 변환합니다. 이러한 시각적 단어를 얻은 후, 우리는 다중 계층 Transformer 인코더를 활용하여 성능을 향상시킵니다. 다층 Transformer 인코더의 글로벌 셀프 어텐션을 통해 네트워크는 대표적인 시각적 단어의 맥락 정보를 모델링하고 가장 차별화되는 특징에 집중할 수 있습니다.

전반적으로 우리 작업의 주요 기여는 다음과 같이 요약될 수 있습니다.

1) 우리는 LBP 특징과 CNN 특징을 글로벌-로컬 어텐션과 글로벌 셀프 어텐션과 통합하여 표정 인식 정확도를 개선하는 FER의 야생용 특징 융합을 갖춘 Visual Transformers를 제안합니다. 2) 우리는 글로벌 및 로컬 얼굴 정보를 모두 집계하기 위해 ASF라는 간단하지만 효과적인 특징 융합 모듈을 설계합니다. 더욱이 ASF는 백본을 안내하여 필요한 정보를 추출하는 동시에 엔드투엔드 방식으로 쓸모없는 정보를 압축합니다. 3) 우리가 아는 한, 우리는 FER에 Transformers를 적용한 최초의 사례입니다. 글로벌 셀프 어텐션은 전체 네트워크가 시각적 특징 시퀀스의 요소 간의 관계를 학습하고 정보가 부족한 영역을 무시할 수 있도록 합니다. 4) 즉, RAF-DB, FERPlus 등 공개적으로 사용 가능한 세 가지 야생용 FER 데이터 세트에 대한 광범위한 실험 결과

. Sun *et al*. [31] proposed cyclic image generation and AffectNet, demonstrate that our VTFF는 달성한다

최첨단 표정 인식 성능. 특히, 우리는 또한 이 세 가지 데이터 세트의 폐색 및 변형 포즈 하위 세트와 CK+에 대한 교차 데이터 세트 평가에 대한 실험을 수행하여 우리 방법의 유망한 일반화 능력을 보여줍니다.

이 기사의 나머지 부분은 다음과 같이 구성되어 있습니다. 섹션 2에서는 야생에서 FER에 대한 관련 작업을 간략히 검토하고 FER의 최근 진전에 대한 포괄적인 검토를 제공합니다. 제안된 VTFF의 핵심 아이디어는 섹션 3에 제시되어 있습니다. 섹션 4에서는 제안된 방법과 최첨단 접근 방식에 대한 광범위한 성능 평가를 제시합니다. 섹션 5에서 작업을 마무리합니다.

2 관련 작업

2.1 야외에서의 얼굴 표정 인식

섹션 1에서 언급한 예상치 못한 상황에 직면했을 때 야생에서의 얼굴 표정은 상당한 도전이 됩니다. 연구자들은 수작업으로 만든 특징[3],[4],[5],[6],[7]에서 차별적 특징을 정확하게 추출하기 위한 딥 특징[32],[33]으로 주의를 돌리기 시작했습니다. 예를 들어, Tang[34]은 특징 추출에 CNN을 활용하고 소프트맥스 계층을 선형 SVM으로 대체하여 상당한 이득을 얻었고 ICML 2013 표현 학습 워크숍의 얼굴 표정 인식 챌린지에서 우승했습니다. 나중에 Li et al.[8]은 딥 지역성 보존 CNN(DLP-CNN) 방식으로 딥 특징의 차별적 힘을 강화하는 것을 제안했습니다. 이 두 가지 접근 방식[34],[8]은 주목할 만하며 FER에 대한 딥 특징의 효과성을 나타냅니다. 한편, 영역 기반 주의 네트워크는 특히 야생에서 FER에 적합합니다.이미지에 일부 폐색이나 클러터가 발생할 때 눈에 띄는 얼굴 특징이 동적으로 전면에 나오도록 허용하기 때문입니다.Wang 등은[10]에서 폐색 및 포즈 변형 FER에 대한 얼굴 영역의 중요성을 포착하기 위해 영역 주의 네트워크(RAN)를 제안했습니다.마찬가지로 Li[9] 등은 가장 차별적인 얼굴 영역에 초점을 맞춘 폐색 인식 FER에 대한 주의 메커니즘이 있는 CNN을 제안했습니다.[9], [10]과 유사하게, 우리의 방법은 글로벌 셀프 주의 메커니즘을 사용하여 글로벌 관점에서 얼굴 표정을 인식하는데, 이는 차별적인 영역을 강조하고 폐색이나 변형 포즈의 경우 잘 일반화합니다.이러한 속성과 무관한 방법 외에도 일부 FER 알고리즘은 표정 인식 성능을 개선하기 위해 인구 통계적 특징(예: 성별, 인종, 연령 등)을 통합했습니다. Fan[24]이 얼굴 속성을 고려하는 심층 감독 주의 네트워크를 처음 제안했다는 점을 언급할 가치가 있습니다.또한 Xu et al.[35]은 편향을 해결하기 위해 속성 정보를 입력으로 사용하고 FER에 대한 편향과 공정성에 대한 비교 연구를 수행했습니다.데이터 세트에 포함된 불확실성을 더욱 억제하기 위해 Wang et al.[11]은 야생에서 FER에 대한 자체 치료 네트워크(SCN)를 제안했는데, 이는 각 훈련 얼굴 샘플에 가중치를 부여하기 위해 주의 메커니즘을 최대한 활용했습니다.이러한 주류 심층 학습 기반 FER 작업은 FER에 대한 신중하게 설계된 기능(예: LBP)과 심층 기능 내의 상호 관계를 무시하고 심층 기능을 추출하기 위해 CNN을 직접 채택하는 데 중점을 둡니다.

2.2 FER의 기능 융합

대부분의 연구 작업은 얼굴 표정을 구별하기 위해 깊은 특징만 사용합니다. 섹션 1에서 언급했듯이, 영어: LBP[36] 및 HOG[37]와 같이 얼굴 표정 특징 표현을 위해 잘 개발된 다양한 수작업 특징이 있습니다.또한 특징 융합은 전체 네트워크의 대표 능력을 풍부하게 하여 일반화 능력과 인식 성능을 효과적으로 개선할 수 있습니다[38],[39],[40]. 따라서 다양한 특징 융합 기반 FER 방법이 표정 인식의 성능을 더욱 개선하기 위해 제안되었습니다.Chen et al.[41]은 야생에서 FER을 해결하기 위해 동적 텍스처, 기하학적 특징 및 음향 특징을 융합할 것을 제안했습니다.시각 정보의 동적 텍스처 기술자는 HOG의 확장으로, HOG-TOP(Histogram of Oriented Gradients from Three Orthogonal Planes)이라고 합니다.또한 Shao와 Qian[42]은 LBP 특징과 딥 특징을 병렬로 추출하기 위해 이중 분기 CNN을 제안했습니다.그들은 이러한 특징 맵을 융합하기 위해 연결 작업을 활용했습니다. 마찬가지로 Li et al. [25]은 LBP 특징과 CNN 특징을 밀집 연결로 결합하여 표현 인식 정확도를 개선했습니다. 그들은 보완 정보를 잘 활용하지 못했고 중복된 정보를 버렸습니다. 또한 이러한 특징 융합 방법은 주로 CK+ [12] 및 Oulu-CASIA [14]와 같은 실험실 제어 FER 데이터 세트에서 수행되었습니다. 이를 바탕으로 우리는 LBP 특징과 CNN 특징을 통합하기 위한 주의 선택적 융합(ASF)을 제안합니다. 이는 쓸모없는 정보를 압축하고 로컬 및 글로벌 관점에서 상관 관계가 있는 가중치 맵을 생성합니다.

2.3 컴퓨터 비전의 트랜스포머

CNN 기반 기능 표현 기술 외에도 Transformers[43]는 NLP에서 지배적인 성능을 보였습니다. Transformers의 성공에 영감을 받아 여러 연구자들이 객체 감지[44], 포즈 추정[45], 고해상도 이미지 합성[46], 비디오 인스턴스 분할[47], 궤적 예측[48] 등과 같은 컴퓨터 비전 작업에 Transformers를 투자하려고 시도했습니다. Transformer 기반 접근 방식은 대규모 데이터 세트에서 완전히 학습했을 때 CNN 기반 방법에 비해 우수한 성능을 보였습니다. Vision Transformer(ViT)[49]는 몇 가지 수정만으로 이미지에 바닐라 Transformer를 적용한 최초의 작업이었습니다. [49]에 따르면 ViT는 ImageNet[50]에서 학습했을 때 ResNet에 비해 정확도가 낮았습니다. ViT는 먼저 대규모 데이터 세트에서 학습한 다음 다운스트림 작업에 맞게 미세 조정되었습니다. Transformers는 컴퓨터 비전 작업에서 일반화하기 위해 많은 양의 데이터가 필요하기 때문입니다. CNN의 피처 피라미드 구조는 Transformers에도 적용되었습니다. Wang et al. [51]은 컨볼루션 없이 피처 추출 백본으로 작동할 수 있는 픽셀 수준 밀도 예측을 위한 Pyramid Vision Transformer(PVT)를 제안했습니다.컨볼루션 작업이 없는 이러한 Transformers 외에도 여러 연구 [52], [53], [54]에서 컨볼루션 레이어를 Transformers에 혼합하여 순수 Transformers의 성능을 더욱 개선하는 것을 제안했습니다.바닐라 Transformer와 이러한 뛰어난 Transformer 변형에서 영감을 얻어 먼저 FER에 Transformers를 직접 적용하는 것을 제안합니다.알다시피, 얼굴 표정 인식을 위해 심층적 피처 간의 관계를 포착하려는 연구는 없습니다.Transformers를 사용하여 전역 자기 주의 메커니즘을 통해 입력 시퀀스 간의 긴 종속성을 모델링합니다.이러한 전역 자기 주의는 모델이 정보가 부족한 영역을 무시하고

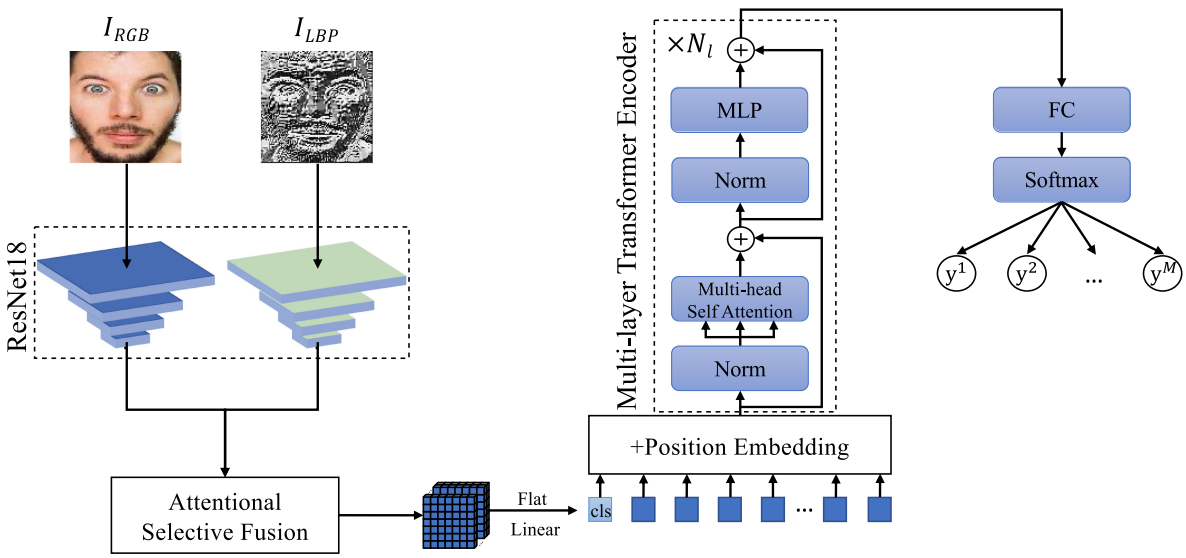


그림 2. 제안하는 VTFF의 개요. 시각적 단어 추출, 관계 모델링, 표현 분류의 세 부분으로 나눌 수 있습니다. 사전 훈련된 ResNet18을 백본으로 사용하여 피처 맵을 추출합니다. 추출된 모든 피처는 주의 선택적 융합을 통해 융합되어 대표적인 시각적 단어를 얻습니다. 입력 시각적 단어는 피처 맵의 공간적 차원을 간단히 평면화하고 특정 차원으로 투영하여 얻습니다. 그런 다음 다중 계층 Transformer 인코더를 적용하여 다양한 시각적 피처 구성 요소 간의 관계를 모델링합니다. 네트워크는 마지막으로 간단한 소프트맥스 함수로 표현 확률을 계산합니다.

FER에 대한 폐색이나 변형 포즈의 경우 글로벌 관점에서의 표현.

3 방법

3.1 개요

그림 2는 얼굴 표정 인식을 위한 특징 융합을 사용한 Visual Transformer의 전체 다이어그램을 보여줍니다. VTFF는 두 개의 사전 학습된 ResNet18[32] 네트워크를 기반으로 하며 두 가지 중요한 구성 요소로 구성됩니다. i) 주의 선택적 융합, ii) 다층 Transformer 인코더.

크기가 H × W × 3인 주어진 얼굴 이미지 I_{RGB}에 대해 먼저 크기가 H × W × 1인 LBP 특징 이미지를 구하고 이를 크기가 H × W × 3인 특징 이미지 I_{LBP}에 연결합니다. 특징 추출 백본은 두 개의 ResNet18 네트워크로 구성됩니다. 하나는 RGB 이미지용이고 다른 하나는 LBP 특징 이미지용입니다. 특히, 우리는 ResNet18의 처음 5단계를 백본으로 사용하여 $\frac{H}{R} \times \frac{W}{R} \times C_f$ 크기의 피처 맵 X_{LBP} 및 X_{RGB}를 추출합니다. 여기서 R은 ResNet18의 다운샘플링 속도이고 C_f은 5단계 출력의 채널 번호입니다. 단순화를 위해 $H_d = \frac{H}{R}$ 및 $W_d = \frac{W}{R}$로 표시합니다. 이 논문에서 R = 32 및 H = W = 224입니다. 우리는 MS-Celeb-1M 얼굴 인식 데이터 세트에 대한 사전 학습된 가중치로 전체 네트워크 가중치를 초기화합니다. 일반화를 잃지 않고 ASF는 RGB 이미지에서 추출한 특징과 섹션 3.2에서 자세히 소개할 LBP 특징 이미지에서 추출한 특징을 결합하는 데 활용됩니다. ASF 모듈은 이러한 특징의 가중치를 동적으로 조정하고 네트워크가 표정 인식을 개선하는 데 중요한 차별적 특징에 더 집중하도록 안내합니다. ASF의 융합 가중치는 글로벌-로컬 어텐션을 통해 생성되며, 이는 추가 표정 인식을 위해 글로벌 및 로컬 컨텍스트를 집계합니다. 융합된 특징 맵 X_{fused}의 크기도 $H_d \times W_d \times C_f$입니다. 그런 다음 평탄화된 특징을 선형 투영과 학습 가능한 분류에 공급하여 크기는 $\delta H_d W_d \times 1 \times C_p$이고, 여기서 C_p은 투영 후 평탄화된 특징의 채널입니다. 또한 [43] 및 [49]에서와 같이 위치 정보를 유지하기 위해 임베딩에 위치 임베딩을 추가합니다. 입력 임베딩은 N_l 인코더 블록으로 구성된 Transformer 인코더에 추가로 공급됩니다. 마지막으로 얼굴 표정의 확률은 완전히 연결된 레이어와 소프트맥스 함수에 의해 생성됩니다.

3.2 주의 선택적 융합

그림 3에서 볼 수 있듯이, 우리의 주의 선택적 융합은 전역적 주의와 지역적 주의로 구성되어 있으며, 이는 다양한 유형의 정보를 융합하는 데 추가적인 유연성을 제공할 수 있습니다. 섹션 3.1에서 언급했듯이, 백본에서 추출된 두 개의 피처 맵 X_{LBP}; $X_{RGB} \in R^{H_d \times W_d \times C_f}$이 주어지면, 우리는 먼저 LBP 피처 X_{LBP}와 CNN 피처 X_{RGB}를 융합하여 후속 정보 상호 작용을 포착합니다.

$$U = W_L X_{LBP} + W_C X_{RGB}, \tag{1}$$

여기서 U는 X_{LBP}과 X_{RGB} 간 합산 후의 통합된 특징 맵이고, þ는 요소별 합산을 나타냅니다. W_L과 W_C은 초기 통합을 위한 가중치이며 간단히 두 개의 1×1 합성곱으로 구현됩니다.

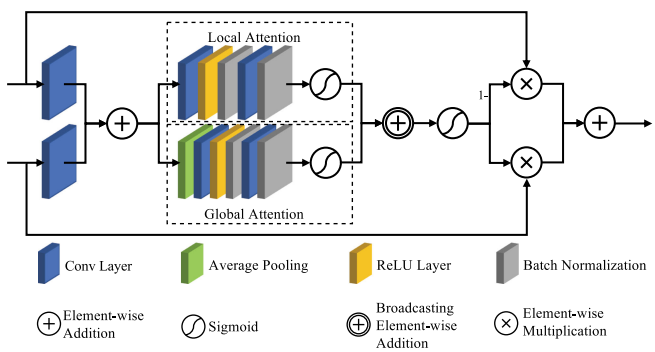


Fig. 3. The attentional selective fusion module.

ken is added. We get embedded visual words

전역 및 지역 선택적 융합을 모두 수행하기 위해 전역 평균 풀링과 픽셀별 합성을 각각 전역 컨텍스트 및 지역 컨텍스트 집계자로 선택합니다. 전역 컨텍스트는 크기가 $H_d \times W_d$인 각 피처 맵을 점진적으로 스칼라로 압축하고 피처의 채널 간 관계를 활용합니다. 전역 컨텍스트와 달리 지역 컨텍스트는 입력 피처의 미묘한 세부 사항을 보존하고 강조하여 전역 컨텍스트를 보완합니다. 지역 및 전역 컨텍스트를 집계하면 네트워크가 다양한 유형의 피처의 이점을 얻고 모호한 표정을 보다 정확하게 인식할 수 있습니다. 전역 컨텍스트와 지역 컨텍스트는 다음과 같이 계산됩니다.

$$G(U) = \sigma(\mathbb{BN}(Conv_G^2(\delta(\mathbb{BN}(Conv_G^1(\mathbb{AP}(U))))))), \quad (2)$$

$$L(U) = \sigma(\mathbb{BN}(Conv_L^2(\delta(\mathbb{BN}(Conv_L^1(U)))))). \quad (3)$$

Eq. (2)에서 보여지는 것처럼, 전역 적응 평균 풀링 AP는 먼저 $U \in R^{H_d \times W_d \times C_f}$의 각 피처 맵을 공간 차원의 스칼라로 압축하는 데 적용되며 이는 Eq. (4)를 사용하여 수행됩니다.

$$\mathbb{AP}(U) = \frac{1}{H_d W_d} \sum_{i=1}^{H_d} \sum_{i=1}^{W_d} x_c(i,j), c = 1, 2, \ldots, C_f, \quad (4)$$

여기서 H_d와 W_d는 입력 피처 맵 U의 높이와 너비이고 C_f은 U의 채널 수입니다. BN은 배치 정규화입니다. d는 ReLU 함수를 나타내고 s는 시그모이드 함수를 나타냅니다. $Conv_G^1$과 $Conv_G^2$의 커널 크기는 $\frac{C_f}{r} \times C_f \times 1 \times 1$, $C_f \times \frac{C_f}{r} \times 1 \times 1$입니다. 그리고 $Conv_L^1$, $Conv_L^2$의 커널 크기는 $\frac{C_f}{r} \times C_f \times 1 \times 1$, $1 \times \frac{C_f}{r} \times 1 \times 1$입니다. $Conv_G^1$와 $Conv_L^1$은 동일한 커널 크기를 갖지만 입력에 대한 함수는 상당히 다르며, 특히 $Conv_G^2$와 $Conv_L^2$와 결합할 때 더욱 그렇습니다. 구체적으로, $\frac{C_f}{r} \times C_f \times 1 \times 1$ 크기의 $Conv_G^1$은 감소 비율 r로 입력을 압축합니다. 이 논문에서는 r을 8로 설정했습니다. 그리고 $Conv_G^2$은 입력의 채널 차원을 증가시키고 다중에 전역 융합 가중치 $G \in R^{1 \times 1 \times C_f}$을 생성합니다. 따라서 $Conv_G^1$와 $Conv_G^2$은 채널 차원을 따라 이러한 피처의 전역 관계를 캡처하는 데 적용됩니다. 한편, $Conv_L^1$는 $U \in R^{H_d \times W_d \times C_f}$를 $H_d \times W_d \times \frac{C_f}{r}$ 모양으로 압축합니다. 그리고 $Conv_L^2$도 입력의 공간 차원을 일정하게 유지하면서 채널 차원을 따라 입력을 압축합니다. $Conv_L^2$의 출력 모양은 $H_d \times W_d \times 1$입니다. 즉, $Conv_L^1$와 $Conv_L^2$은 피처 맵의 채널 차원을 점진적으로 줄여 로컬하게 분산된 차별 정보에 집중합니다.

전역 융합 가중치 G와 지역 융합 가중치 L가 주어지면, 정제된 전역-지역 어텐션 가중치는 Eq. (5)를 통해 얻을 수 있습니다.

$$GL(U) = G(U) \oplus L(U), \quad (5)$$

여기서 $\oplus$는 브로드캐스팅 추가를 나타냅니다. 그런 다음 융합된 피처 맵 X_{fused}은 다음과 같이 계산됩니다.

$$X_{fused} = X_{LBP} \otimes \sigma(GL(U)) + X_{RGB} \otimes \sigma(1 - GL(U)), \quad (6)$$

여기서 $\otimes$은 요소별 곱셈입니다.

3.3 다층 변압기 인코더

융합된 2D 피처 맵 X_{fused}는 1D 시각적 임베딩 시퀀스로 평면화되어야 하며, 더 나아가 다중 계층 Transformer 인코더에 입력으로 공급될 수 있습니다. 따라서 $X_{fused} \in R^{H_d \times W_d \times C_f}$을 평면화된 시퀀스로 재구성하고 선형 투영에 공급하여 $X_f \in R^{H_d W_d \times C_p}$을 얻습니다. 여기서 $H_d W_d$는 시퀀스 길이이고 C_p, C_f은 각각 512와 768로 설정됩니다. [49]에서와 같이 분류 토큰 [cls]가 입력 시퀀스 X_f의 시작 부분에 추가됩니다. Transformer 인코더의 출력에서 [cls] 토큰의 학습 가능한 상태는 전체 피처 시퀀스를 나타내는 데 활용되며, 이는 최종 예측에 사용됩니다. 다중 레이어 변환기 인코더에 위치 정보를 통합하기 위해 1D 학습 가능한 위치 임베딩이 피처 임베딩에 추가되었습니다.

$$Z^0 = [x_{cls}; x_f^1; x_f^2; x_f^3; \ldots; x_f^{H_d W_d}] + PE(H_d W_d + 1; C_p), \quad (7)$$

여기서 $PE(H_d W_d + 1; C_p) \in R^{(H_d W_d + 1) \times C_p}$은 각 위치 인덱스에 대한 임베딩을 학습하고 [cls] 토큰을 포함하며 Z^0는 결과 위치 인식 기능 시퀀스를 나타냅니다.

얼굴 특징 임베딩의 모든 요소 간의 복잡한 상호 작용을 모델링하기 위해 표준 다중 레이어 Transformer 인코더에 Z^0를 입력합니다. Transformer 인코더는 다중 헤드 자기 주의(MHSA)를 통해 임베딩 Z^0의 가중치를 계산합니다. 이는 학습 가능한 쿼리 Q, 키 K 및 값 V에 의해 수행됩니다. Eq. (8)을 사용하여 단일 헤드 글로벌 자기 주의(SHSA)를 계산합니다. 첫 번째 레이어의 SHSA 세부 정보는 다음과 같이 공식화할 수 있습니다.

$$\begin{aligned} head_j &= Attention(Q_j, K_j, V_j) \\ &= softmax\left(\frac{Q_j K_j^T}{\sqrt{d}}\right) V_j \\ &= softmax\left(\frac{Z^0 W_j^Q (Z^0 W_j^K)^T}{\sqrt{d}}\right) Z^0 W_j^V, \end{aligned} \quad (8)$$

여기서 $Q_j = Z^0 W_j^Q$, $K_j = Z^0 W_j^K$, $V_j = Z^0 W_j^V$ 및 $W_j^Q \in R^{C_p \times d}$, $W_j^K \in R^{C_p \times d}$, $W_j^V \in R^{C_p \times d}$는 이러한 선형 투영의 매개변수입니다. 구체적으로, 여러 쿼리, 키 및 값은 Z^0을 N_h개의 다른 표현 부분 공간으로 투영합니다. 다중 헤드 자기 주의(MHSA)는 다음과 같이 설명할 수 있습니다.

$$MHSA(Z^0) = concat(head_1, \ldots, head_{N_h}) W^O, \quad (9)$$

여기서 N_h은 다른 헤드의 수이고, concat은 연결 연산을 나타냅니다. $W^O \in R^{h_1 \times d}$은 선형 투영의 매개변수이며, 각 헤드 d의 차원은 $\frac{C_p}{N_h}$와 같고 h_1은 첫 번째 레이어의 숨겨진 크기입니다. 각 Transformer 인코더는 N_l개의 MHSA 블록 레이어로 구성됩니다. 형식적으로 표준 다중 레이어 Transformer 인코더는 $i = 1; \ldots; N_l$개의 레이어에 대해 전방으로 계산합니다.

$$\hat{Z}^i = MHSA(LN(Z^{i-1})) + Z^{i-1} \quad (10)$$

$$Z^i = MLP(LN(\hat{Z}^i)) + \hat{Z}^i \quad (11)$$

여기서 $\hat{Z}^i$와 Z^i은 레이어 i에서의 중간 출력과 최종 출력입니다. MLP는 두 개의 위치별 피드포워드 레이어와 GELU 비선형 활성화 함수로 구성됩니다. LN은 레이어 정규화를 나타내며, 이는 다음과 같이 적용됩니다.

모든 어텐션 블록과 MLP 이전에. 이 논문에서 MLP의 숨겨진 차원은 3,072로 설정됩니다. 계층 N의 출력 $Z_0^{N_l}$도 LN에 의해 정규화됩니다. 최종 [cls] 토큰에 대해 완전 연결 계층만 적용한다는 점이 주목할 만합니다. 이는 표현 확률 점수를 계산하기 위한 것입니다. 수학적으로 확률 점수는 다음과 같이 생성됩니다.

$$Y = LN(Z_0^{N_l}), \tag{12}$$

$$y^j = \frac{e^{\theta_i^T Y}}{\sum_{i=1}^{M} e^{\theta_i^T Y}}, \tag{13}$$

여기서 $Z_0^{N_l}$는 전체 시퀀스 Z_l^N의 첫 번째 [cls] 토큰이고 Y는 다중 계층 변압기 인코더의 출력입니다. u는 완전 연결 계층의 매개변수를 나타내고 u_i는 u의 i번째 열입니다. 표정 클래스의 수는 M입니다. y^j는 i번째 얼굴 표정의 확률 점수이고 최종 예측 표정은 추론 중에 arg max 함수로 쉽게 얻을 수 있습니다.

4가지 실험

제안된 방법의 효과를 입증하기 위해, 우리는 세 가지 야생 FER 데이터 세트(즉, RAF-DB, FERPlus 및 AffectNet)에 대한 광범위한 실험과 CK+에 대한 교차 데이터 세트 평가를 수행합니다. 이 섹션에서는 먼저 실험에 사용된 FER 데이터 세트와 구현 세부 정보를 소개합니다. 그런 다음 제안된 방법을 여러 최신 접근 방식과 비교합니다. 그런 다음 제안된 VTFF 모델의 각 구성 요소의 영향을 이러한 데이터 세트에 대한 실험을 통해 조사합니다.

4.1 데이터 세트

우리는 자주 사용되는 세 가지 얼굴 표정 데이터 세트(RAF-DB, FERPlus 및 AffectNet)에서 우리의 접근 방식을 평가합니다. 이러한 데이터 세트는 모두 야외에서 수집되었으며, 서로 다른 조명과 폐색으로 인해 어려움을 겪을 수 있습니다. 실제 환경에서 폐색 및 변형 포즈 문제를 처리할 때 우리 방법의 효과를 보여주기 위해 RAF-DB, FERPlus 및 AffectNet의 하위 세트인 Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, Pose-AffectNet에 대한 실험도 수행합니다. 교차 데이터 세트 평가 실험은 또한 CK+에서 수행되어 우리 방법의 우수한 일반화 능력을 검증합니다. 실험에 사용된 데이터 세트의 세부 사항은 다음과 같습니다.

RAF-DB에는 Flickr에서 수집한 29,672개의 실제 얼굴 이미지가 포함되어 있습니다. RAF-DB의 전체 이미지는 315명의 잘 훈련된 주석자에 의해 레이블이 지정되고 각 이미지는 약 40명의 독립적인 주석자에 의해 레이블이 지정됩니다. RAF-DB에는 단일 레이블 하위 집합과 다중 레이블 하위 집합의 두 가지 다른 하위 집합이 있습니다. 실험에서는 7가지 기본 감정(중립, 행복, 놀람, 슬픔, 화남, 혐오, 두려움)을 포함하여 단일 레이블 하위 집합만 사용합니다. 단일 레이블 하위 집합의 이미지는 12,271개의 훈련 샘플과 3,068개의 테스트 샘플로 나뉩니다. 훈련 이미지와 테스트 이미지의 표정은 불균형 분포를 갖습니다. 전체 샘플 정확도는 성능 측정에 사용됩니다.

FERPlus는 ICML 2013 표현 과제를 위한 원래 FER2013 데이터 세트에서 확장되었습니다. 학습. FERPlus에는 28,709개의 훈련 이미지와 3,589개의 테스트 이미지가 포함되어 있으며, 모두 Google 검색 엔진에서 수집합니다. FERPlus의 이미지 원래 크기는 48 × 48입니다. FERPlus의 각 이미지는 10명의 주석자가 주석을 달았으며 FERPlus는 원래 FER2013 레이블보다 더 나은 품질의 레이블을 제공합니다. RAF-DB와 같은 7가지 기본 감정 외에도 경멸 범주가 레이블에 포함됩니다. 우리는 주로 성과 측정을 위한 다수결 투표의 감독 하에 전체 샘플 정확도를 보고합니다.

AffectNet은 인터넷에서 수집한 1,000,000개 이상의 얼굴 이미지가 있는 가장 큰 얼굴 표정 데이터 세트입니다. AffectNet은 불연속적인 범주형 및 연속적인 차원(즉, 가치와 각성) 주석을 모두 제공합니다. AffectNet은 불균형한 훈련, 검증 및 테스트 세트를 가지고 있으며, 그 중 450,000개 이미지에 수동으로 주석이 달려 있다는 점에 유의해야 합니다. 실험에서는 FERPlus로 8가지 기본 표정으로 주석이 달린 이미지를 활용하고, 287,652개 이미지는 훈련에, 4,000개 이미지는 테스트에 활용합니다. 테스트 세트는 대중에게 공개되지 않으므로 주로 성능 측정 및 다른 방법과의 공정한 비교를 위해 검증 세트에 대한 평균 클래스 정확도를 보고합니다.

Occlusion and Pose Variant Datasets(즉, Occlusion-RAF-DB, Pose-RAF-DB, Occlusion-FERPlus, Pose-FERPlus, Occlusion-AffectNet, Pose-AffectNet)는 RAF-DB, FERPlus, AffectNet의 폐색 및 포즈 하위 집합이며 [10]에서 수동으로 수집했습니다. 이러한 데이터 세트의 샘플에서 발생하는 다양한 폐색 유형은 마스크 착용, 안경 착용, 상체/하체 얼굴의 물체 등입니다. 또한 변형 포즈 문제는 30도보다 큰 포즈와 45도보다 큰 포즈의 두 가지 범주로 나눌 수 있습니다. 샘플 분포는 [10]에서 찾을 수 있습니다. 해당 원본 데이터 세트에 따라 전체 샘플 정확도 또는 평균 클래스 정확도를 보고합니다.

CK+는 실험실에서 수집한 환경에서 수집한 얼굴 동작 단위 및 표정 인식을 위한 확장된 Cohn-Kandade(CK) 데이터 세트입니다. CK+의 원본 데이터에는 123명의 피험자로부터 얻은 593개의 비디오 시퀀스가 있으며, 그 중 327개에는 7가지 기본 감정과 경멸이 주석으로 달려 있습니다. 각 비디오 시퀀스는 시작(첫 번째 프레임)부터 최대 표정(마지막 프레임)까지의 이미지로 구성됩니다. 이전 연구에 따라 비디오 시퀀스의 첫 번째 프레임을 중립 샘플로 사용하고 마지막 프레임을 대상 표정으로 사용합니다. 총 7가지 감정이 있는 이미지 618개와 8가지 감정이 있는 이미지 654개를 테스트용으로 얻습니다. 전체 샘플 정확도를 사용하여 방법의 일반화 용량을 평가합니다.

4.2 구현 세부 사항

실험에서 얼굴 이미지는 MTCNN[19]에 의해 감지되고 224×224의 크기로 추가로 조정됩니다. 이전 최신 방법과 공정하게 비교하기 위해 MS-Celeb-1M 얼굴 인식 데이터 세트에서 사전 학습된 동일한 백본 ResNet18을 사용합니다. 얼굴 특징은 ResNet18의 마지막 합성곱 단계에서 추출됩니다. 방법의 학습률은 0.005로 초기화됩니다. 1,000단계의 선형 학습률 워밍업과 코사인 학습률 감소를 사용합니다. Adam 최적화기[61]를 사용하여 배치 크기 32로 전체 네트워크를 최적화하고 RAF 및 FERPlus에서 20,000단계, AffectNet에서 각각 40,000단계로 모델을 학습합니다. 표준 교차 엔트로피 손실을 사용하여 표현을 위해 모델을 잘 일반화하도록 감독합니다.

표 1 RAF-DB에 대한 최신 방법과의 비교

method	Year	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Accuracy
VGG [8]	2018	66.05	25.00	37.84	73.08	51.46	53.49	47.21	69.34
baseDCNN [8]	2018	70.99	52.50	50.00	92.91	77.82	79.64	83.09	82.66
Center Loss [8]	2018	68.52	53.13	54.05	93.08	78.45	79.63	83.24	82.86
DLP-CNN [8]	2018	71.60	52.15	62.16	92.83	80.13	81.16	80.29	82.74
FSN [55]	2018	72.80	46.90	56.80	90.50	81.60	81.80	76.90	81.14
gACNN [9]	2018	-	-	-	-	-	-	-	85.07
RAN [10]	2020	-	-	-	-	-	-	-	86.90
SCN [11]	2020	-	-	-	-	-	-	-	87.03
DSAN-VGG-RACE [24]	2020	82.71	56.25	58.11	94.01	83.89	**89.06**	80.00	85.37
SPWFA-SE [56]	2020	80.00	59.00	59.00	93.00	84.00	88.00	86.00	86.31
Ours	2021	**85.80**	**68.12**	**64.86**	**94.09**	**87.24**	85.41	**87.50**	**88.14**

가장 좋은 결과는 굵은 글씨로 표시되어 있습니다.

인식. 우리는 Pytorch[62] 툴박스를 사용하여 방법을 구현하고 모든 실험을 단일 NVIDIA GTX 1080Ti GPU 카드에서 수행합니다.

4.3 최신 방법과의 비교

제안된 VTFF 모델은 RAF-DB, FERPlus 및 AffectNet에서 여러 최신 방법과 비교됩니다. 우리는 이러한 데이터 세트에서 우리의 지식에 대한 새로운 최신 결과를 얻습니다. 더 나은 성능은 제안된 방법의 우수성을 보여줍니다.

RAF-DB에 대한 결과. 다른 최신 방법과의 비교는 표 4에서 확인할 수 있습니다. [8]의 방법은 평균 정확도를 사용하여 성능을 제시했습니다. 공정한 비교를 위해 [56]에 따라 결과를 정확도로 변환했습니다. gACNN[9], RAN[10] 및 SCN[11]은 특정 표현 인식 정확도나 혼동 행렬을 보고하지 않았으므로 표 4에서 해당 위치를 '-'로 표시했습니다. SPWFA-SE[56]도 특정 표현 인식 정확도를 보고하지 않았지만 RAF-DB에 대한 혼동 행렬을 제공했습니다. 따라서 비교를 위해 혼동 행렬에서 정확도 결과를 빌렸습니다. 전반적으로 제안하는 방법은 RAF-DB에서 88.14%를 달성합니다. 표 4에서 볼 수 있듯이, 우리 방법은 놀라움 범주를 제외한 모든 방법 중에서 가장 좋은 결과를 달성합니다. 자세히 말하면, 우리의 VTFF는 각각 기준 방법과 이전 최첨단 방법인 VGG와 SCN에 비해 18.8%와 1.11%의 이득을 얻었습니다.DSAN-VGG-RACE는 순수하게 사용된 표정 레이블과 비교했을 때 추가 데이터인 인종 레이블이 있는 심층 감독 블록과 주의 블록을 통합했습니다. RAF-DB도 극도로 불균형한 분포를 가지고 있기 때문에 놀라움 범주의 약간의 성능 저하는 합리적이고 수용 가능합니다. 우리 방법의 역겨움 표정 인식의 정확도는 [56]의 이전 최고 결과에 비해 9.12%의 증가를 기록했으며, 이는 특징 학습에서 우리 방법의 효과성과 우수성을 보여줍니다.

FERPlus에 대한 결과. 표 4는 FERPlus에 대한 비교 결과를 보여줍니다. 우리는 CSLD [15], ResNet+VGG [57], SHCNN [58], LDR [59]을 포함한 CNN 기반 방법과 두 가지 최신 최첨단 방법(RAN [10], SCN [11])과 모델을 비교합니다. 표 4에서 볼 수 있듯이 제안하는 VTFF는 FERPlus에서 88.81%를 달성합니다. 동일한 경험에서

iment settings, total improvements of our VTFF FERPlus에서 RAN 및 SCN과 비교했을 때 0.96% 및 0.80%입니다. 특히 RAN°는 추가 얼굴 정렬이 있는 개선된 RAN을 의미합니다. 얼굴 정렬은 얼굴 인식 및 얼굴 표정 인식에 필수적이지만 사전 처리 단계이며 CNN 기반 방법은 종단 간 방식으로 표정을 인식하는 경향이 있습니다. 실험 설정에서 얼굴 정렬 없이도 VTFF는 RAN°보다 훨씬 더 나은 결과를 달성합니다.

AffectNet에 대한 결과. 우리는 우리의 방법을 AffectNet에 대한 여러 방법과 비교하고 표 4에 결과를 보고합니다. 우리는 장식 없이 AffectNet에서 오버샘플링으로 61.85%를 얻습니다. IPA2LT[60], gACNN[9] 및 SPWFA-SE[56]는 경멸 범주 없이 AffectNet에서 7개 클래스에 대해 훈련되었습니다. 위에서 언급했듯이 AffectNet은 불균형 분포를 갖습니다. SPWFA-SE는 초점 손실 함수를 사용하여 불균형 문제를 처리합니다. RAN[10] 및 SCN[11]과 마찬가지로 불균형 문제를 처리하기 위해 우리는 실험에서 오버샘플링 전략을 채택합니다. 특히,

표 2 FERPlus 및 AffectNet의 최신 방법과의 비교

(a) Results on FERPlus.

Method	Year	Accuracy
CSLD [15]	2016	83.85
ResNet+VGG [57]	2017	87.4
SHCNN [58]	2019	86.54
LDR [59]	2020	87.6
RAN° [10]	2020	88.55
RAN [10]	2020	87.85
SCN [11]	2020	88.01
Ours	2021	**88.81**

(b) Results on AffectNet.

Method	Year	Accuracy
IPA2LT [60]	2018	55.11
gACNN [9]	2018	58.78
SPWFA-SE [56]	2020	59.23
RAN† [10]	2020	52.97
RAN [10]	2020	59.50
SCN [11]	2020	60.23
Ours†	2021	56.13
Ours	2021	**61.85**

The best results are in bold.

표 3 AffectNet-7에 대한 최근 발표된 방법과의 비교

method	Year	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Accuracy
MFMP+ [63]	2021	55.00	46.00	53.00	88.00	55.00	55.00	64.00	58.86
IDFL [64]	2021	31.00	65.00	49.00	95.00	59.00	43.00	73.00	59.20
WSFER [65]	2021	58.54	30.28	50.42	88.01	**67.56**	51.90	73.55	60.04
T21DST [66]	2021	18.00	40.00	53.00	**96.00**	62.00	62.00	**79.00**	60.12
SDW [67]	2021	53.00	**56.00**	61.00	86.00	58.00	53.00	59.00	61.11
ReCNN [68]	2021	59.00	54.40	**65.60**	87.60	59.40	60.00	62.40	64.06
Ours	2021	**61.20**	53.00	60.40	88.40	60.80	**64.80**	65.00	**64.80**

정확도 결과는 혼동 행렬에서 차용한 것입니다. 가장 좋은 결과는 굵은 글씨로 표시되어 있습니다.

†로 표시된 방법은 오버 샘플링 전략 없이 학습되었음을 나타냅니다. 오버 샘플링을 사용하지 않는 저희 방법은 RAN†을 3.16% 초과합니다. 또한 오버 샘플링 기술은 RAN†와 저희 방법을 각각 6.53%와 5.72% 향상시켜 오버 샘플링 기술이 불균형 클래스 분포의 부정적 효과를 크게 제거함을 보여줍니다. 학습을 위해 AffectNet만 사용한 저희 VTFF는 사전 학습을 위해 추가 데이터 세트 WebEmotion을 적용한 다음 AffectNet에서 SCN을 미세 조정한 SCN보다 1.62% 더 우수한 성능을 보였습니다. 이전 방법에 비해 저희 VTFF가 개선된 점은 VTFF가 AffectNet과 같은 대규모 표현 인식 데이터 세트에서도 실제로 더 나은 일반화 능력을 가지고 있음을 시사합니다.

영어: AffectNet에서 제안한 VTFF를 더 자세히 조사하기 위해 VTFF의 성능을 [63], [64], [65], [66], [67], [68]을 포함한 최근에 발표된 접근 방식과 비교합니다.공정한 비교를 위해 경멸 범주를 제외한 모델도 훈련합니다.표 3은 AffectNet-7에서 세부 비교 결과를 보여줍니다.제안하는 방법이 다른 방법에 비해 평균 클래스 정확도 측면에서 5.94%~0.74%의 개선으로 뛰어난 성능을 달성한 것을 알 수 있습니다.또한 제안하는 방법은 이러한 방법 중 두 가지 얼굴 표정(분노와 놀람)에 대해 각각 61.2%와 64.80%로 가장 높은 정확도를 달성합니다.Xia et al. [68]은 가장 차별적인 얼굴 영역에 초점을 맞추는 Relation Convolutional Neural Network를 제안하여 평균 64.6%의 정확도와 Fear(64.6%)에서 가장 높은 정확도를 달성했습니다. IDFL[64]과 T21DST[66]는 Neural(79.00%, 73.00%)과 Happy(96.00%, 95.00%)에서 더 높은 정확도를 달성했지만 Angry(31.00%, 18.00%)와 Fear(49.00%, 53.00%)에서의 성능은 만족스럽지 못했습니다.이는 불균형 클래스 분포를 고려하지 않았기 때문입니다.[64], [66]과 달리 Hayale et al.[67]은 훈련 세트의 비율을 기반으로 클래스에 대해 다른 페널티 가중치를 설정하여 Disgust(56.00%)에서 가장 높은 정확도를 달성하는 것을 제안합니다. Zhang et al. [65]은 노이즈가 있는 데이터를 활용하여 FER의 성능을 높이고 Sad에서 최상의 결과(67.56%)를 얻었습니다. 표 3에서 제안하는 VTFF가 경쟁력 있는 결과를 달성하고 최근에 발표된 방법보다 성능이 더 뛰어나다는 결론을 내릴 수 있습니다. 이는 VTFF의 효과성을 보여줍니다.

오클루전 및 포즈 변형 데이터 세트에 대한 결과. 실제 시나리오에서 오클루전 및 변형 포즈의 경우 방법을 검토하기 위해 오클루전-RAF-DB, 포즈-RAF-DB, 오클루전-FERPlus, 포즈-FERPlus, 오클루전-AffectNet 및 포즈-AffectNet에 대한 여러 실험도 수행합니다. 표 4는 정확도를 보여줍니다.

the experimental results under corresponding

하위 집합. RAN [10]은 얼굴 이미지를 하위 영역으로 나누고 폐색 및 포즈 변형 표현 인식을 위해 다른 영역의 중요성을 포착하기 위해 영역 편향 손실을 도입했습니다. 우리의 VTFF는 폐색 및 변형 포즈 FER 문제를 위해 특별히 설계되지 않았지만, 우리의 방법은 각 경우에서 큰 마진으로 RAN보다 성능이 뛰어나서 우리 방법의 우수성을 보여줍니다. 구체적으로, 우리의 방법은 Occlusion-RAF-DB, Occlusion-FERPlus 및 Occlusion-AffectNet에서 RAN보다 각각 1.23%, 1.16% 및 4.48% 더 뛰어납니다. 우리의 방법은 또한 Pose-RAF-DB, Pose-FERPlus 및 Pose-AffectNet에서도 RAN보다 성능이 뛰어납니다. 포즈가 30도보다 클 때 이득은 각각 1.23%, 6.06% 및 6.71%입니다. 45도보다 큰 포즈를 가진 Pose-FERPlus, Pose-AffectNet 및 Pose-RAF-DB에서 우리 방법은 각각 3.15%, 6.8% 및 7.81%의 이득으로 RAN보다 상당히 우수한 성능을 보였습니다. 전반적으로 이러한 결과는 폐색 및 변형 포즈 문제에 대한 우리 방법의 효과를 확실하게 검증합니다. 또한, 우수한 성능은 글로벌 관점에서 일련의 시각적 단어로 얼굴 표정을 인식하는 것이 가능하고 효과적이라는 가설과 일치합니다.

CK+에 대한 결과. 또한 우리 방법의 우수한 일반화 능력을 검증하기 위해 교차 데이터 세트 평가를 수행합니다. 구체적으로, 우리는 먼저 RAF-DB, FERPlus 및 AffectNet에서 개별적으로 네트워크를 훈련한 다음

표 4 폐색 및 포즈 변형 데이터 세트에 대한 다른 방법과의 비교

(a) Results on Occlusion-RAF-DB, Pose-RAF-DB.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	80.19	84.04	83.15
RAN [10]	82.72	86.74	85.20
Ours	83.95	87.97	88.35

(b) Results on Occlusion-FERPlus, Pose-FERPlus.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	73.33	78.11	75.50
RAN [10]	83.63	82.23	80.40
Ours	84.79	88.29	87.20

(c) Results on Occlusion-AffectNet, Pose-AffectNet.

Method	Occlusion	Pose(30)	Pose(45)
Baseline [10]	49.48	50.10	48.50
RAN [10]	58.50	53.90	53.19
Ours	62.98	60.61	61.00

표 5 CK+에 대한 교차 데이터 세트 평가 결과

Method	Train	Test	Accuracy
gACNN [9]	RAF-DB	CK+	81.07
SPWFA-SE [56]	RAF-DB	CK+	81.72
SPWFA-SE [56]	AffectNet-7	CK+	85.44
Ours	RAF-DB	CK+	81.88
Ours	FERPlus	CK+	83.79
Ours	AffectNet-8	CK+	86.24

모델은 CK+에서 직접 수행됩니다. 표 5는 우리의 방법이 이전 접근 방식보다 더 나은 성능을 달성한다는 것을 보여줍니다. gACNN과 SPWFA-SE는 7가지 기본 감정을 예측하도록 훈련되었습니다. 우리 모델은 일반적으로 표 3에 표시된 대로 최종 결과를 낮추는 한 가지 표정(경멸)을 더 예측하지만 AffectNet에서 훈련된 모델은 CK+에서 86.54%의 정확도를 얻습니다. gACNN과 SPWFA-SE와 비교할 때 우리 방법은 더 나은 성능을 보이며 각각 0.81%와 0.16% 증가합니다. AffectNet에서 훈련된 모델은 RAF-DB와 FERPlus에서 훈련된 모델보다 더 높은 정확도를 달성합니다. AffectNet은 비교적 대규모 데이터 세트이고 더 많은 얼굴 이미지를 포함하고 있기 때문입니다. 표 5는 우리 방법이 더 나은 일반화 용량을 가지고 있으며 예외 없이 더 나은 성능을 달성한다는 것을 보여줍니다.

4.4 절제 연구

그림 2에서 보듯이, 제안하는 방법 VTFF는 주의 선택적 융합(ASF) 모듈과 다층 트랜스포머 인코더(MTE)를 갖춘 LBP 특징으로 구성됩니다. 이러한 모듈의 효과를 검증하기 위해 VTFF의 일부를 버림으로써 RAF, FERPlus 및 AffectNet에 대한 비교 실험을 수행합니다. 이러한 실험의 세부 설정은 표 6에서 찾을 수 있으며, 여기서 설정(a)는 기준선 방법을 나타냅니다. ASF는 LBP 특징과 CNN 특징을 통합하는 데 사용되므로 LBP 모듈 없이는 유지할 수 없습니다.

FER에 대한 LBP 특징의 효과성. 저희 방법은 LBP 특징 추출로 시작하므로, 저희는 FER에 대한 LBP 특징의 영향을 조사하기 위해 절제 연구를 설계합니다. 예를 들어 RAF-DB에 대한 결과를 살펴보면, RAF-DB에 아무런 수정도 하지 않은 기준선은 평균 정확도 측면에서 86.37%를 달성합니다. RAF-DB에 LBP 특징을 추가하면 기준선이 0.16% 향상됩니다. 설정(a, b)와 설정(d, e)는 LBP 특징을 통합하면 FERPlus와 AffectNet의 기준선이 각각 0.47%와 0.20% 향상됨을 보여주며, 이는 LBP 특징이 표정 인식 성능 개선. 이는 LBP 특징이 질감 정보를 추출하고 미묘한 표정 차이를 보여주는 미세한 얼굴 변화를 반영할 수 있기 때문에 설명할 수 있습니다. 그럼에도 불구하고 FER에 대한 추가 LBP 특징을 직접 사용하는 것은 LBP 특징과 CNN 특징을 결합하는 간단한 추가 융합 전략이 만족스럽지 않기 때문에 제한적입니다.

주의 선택적 융합(ASF) 평가. LBP 특징과 CNN 특징을 융합하는 ASF의 효과를 검증하기 위해 ASF를 간단한 요소별 추가로 대체하여 실험을 수행했습니다. 설정(b, c)와 설정(e, f)에 따르면, ASF는 LBP 특징과 CNN 특징을 융합할 때 인식 정확도가 증가하여 제안된 ASF의 효과를 보여줍니다. 구체적으로, 설정(b, c)에서 설계된 ASF가 성능을 0.61%, 0.38%, 1.23% 더 향상시킨다는 것을 알 수 있습니다. 설정(a, c)에 따르면, ASF는 추가 LBP 특징으로 기준 설정(a)를 각각 0.76%, 0.85%, 1.43% 더 향상시킵니다. ASF는 LBP 특징과 CNN 특징을 융합하기 위해 전역 및 지역 컨텍스트를 집계하여 인식 성능을 더욱 향상시킵니다.

또한 ASF를 연결 연산과 비교합니다. L_{LBP}와 L_{RGB}을 연결하는 것은 실제로 데이터 수준 융합 방법인 반면, 주의 선택적 융합은 기능 수준에서 작동합니다. 주의 선택적 융합이 연결 연산을 능가하는지 확인하기 위해 해당 실험을 수행합니다. 구체적으로, 사전 학습된 ResNet의 입력에 맞게 $H \times W \times 4$ 모양을 $H \times W \times 3$으로 줄이기 위해 추가 레이어를 추가합니다. 전체 모델을 미세 조정하고 모델을 처음부터 학습합니다. 표 7에서 ASF를 사용하는 방법이 간단한 연결 연산을 사용한 방법보다 모든 데이터 세트에서 성능을 크게 향상시킨다는 결론을 내릴 수 있으며, 이는 이 조건에서 기능 수준 융합 기술(즉, ASF)이 데이터 수준 융합 기술(즉, 연결)을 능가함을 나타냅니다.

다층 변압기 인코더(MTE)의 효과성. MTE의 영향을 알아보기 위해 이 부분에서 MTE의 성능을 평가합니다. 또한 RAF-DB, FERPlus 및 AffectNet에 대한 MTE의 효과를 비교합니다. 표 6에서 다른 모듈의 효과를 알아보고 공정한 절제 연구를 수행하기 위해 계층 수 $N_l = 4$와 헤드 수 $N_h = 8$을 설정했다는 점을 설명하는 것이 좋습니다. 설정(a, d), 설정(b, e) 및 설정(c, f)에서 MTE가 성능을 크게 향상시킨다는 것을 분명히 알 수 있습니다. 특히 기준 설정(a)과 비교했을 때 설정(d)는 MTE를 통합하면 기준보다 성능이 우수하다는 것을 보여줍니다.

표 6 RAF-DB, FERPlus 및 AffectNet에서 수행된 LBP 기능, ASF 및 MTE에 대한 절제 연구

Setting	LBP	ASF	MTE	RAF-DB	FERPlus	AffectNet
a	✗	✗	✗	86.37 ± 0.35	86.70 ± 0.24	58.45 ± 0.10
b	✓	✗	✗	86.53 ± 0.27	87.17 ± 0.30	58.65 ± 0.14
c	✓	✓	✗	87.14 ± 0.39	87.55 ± 0.18	59.88 ± 0.21
d	✗	✗	✓	87.60 ± 0.12	87.65 ± 0.15	60.50 ± 0.09
e	✓	✗	✓	87.83 ± 0.04	88.15 ± 0.23	60.90 ± 0.12
f	✓	✓	✓	88.19 ± 0.21	88.70 ± 0.17	61.52 ± 0.07

그것은우리는다른 모듈의 효과를 탐색하고 공정한 절제 연구를 수행하기 위해 $N_l = 4$와 $N_h = 8$을 설정했습니다. 우리는 여러 번 실험을 세 번 실행합니다. ple random 사용한 다음 인식 정확도의 평균과 표준 편차를 계산합니다.

표 7 ASF와 RAF-DB, FERPlus 및 AffectNet에 대한 연결 융합에 관한 절제 연구(평균 ± 표준 편차)

Method	RAF-DB	FERPlus	AffectNet
Ours w/ Concat	81.12 ± 0.25	83.97 ± 0.37	58.07 ± 0.15
Ours* w/ Concat	87.35 ± 0.39	86.86 ± 0.16	59.85 ± 0.10
Ours w/ ASF	82.30 ± 0.31	84.71 ± 0.20	58.77 ± 0.09
Ours* w/ ASF	87.83 ± 0.04	88.15 ± 0.23	60.90 ± 0.12

* 는 사전 훈련된 가중치를 사용하는 방법을 나타내며, 그렇지 않은 경우 방법은 처음부터 훈련됩니다.

표 8 RAF-DB, FERPlus 및 AffectNet에서 수행된 헤드 수, 레이어 수에 대한 절제 연구

Setting	N_l	N_h	Params(M)	RAF-DB	FERPlus	AffectNet
i	4	4	51.8	87.45	88.46	61.08
ii	4	8	51.8	**88.14**	88.69	61.55
iii	4	12	51.8	87.52	88.65	**61.85**
iv	8	4	80.1	87.22	88.14	60.82
v	8	8	80.1	87.61	**88.81**	61.23
vi	8	12	80.1	87.48	88.52	61.30
vii	12	4	108.5	87.23	88.20	60.85
viii	12	8	108.5	87.29	88.52	60.45
ix	12	12	108.5	87.09	88.21	61.50

굵은 글씨로 표시된 값은 가장 좋은 성과를 나타냅니다.

1.23%, 0.85%, 2.05%가 RAF-DB, FERPlus 및 AffectNet에 적용되었습니다. 설정(a, b)와 설정(b, c)는 LBP 기능과 ASF를 통합하면 평균 정확도 측면에서 각각 0.26%와 0.61%만큼 RAF의 기준선이 개선됨을 보여줍니다. 설정(a, d)에 따르면 MTE를 사용하면 기준선 방법 ResNet18보다 1.23% 더 우수합니다. 표 6에서 N_l과 N_h을 일정하게 유지할 때 MTE가 LBP 기능과 ASF보다 정확도 개선에 가장 크게 기여한다는 결론을 내릴 수 있습니다. 또한 MTE를 사용하면 판별적 기능을 학습하는 능력이 증가하여 해당 기준선보다 성능이 우수하다고 추론합니다.

MTE의 계층 및 헤드 수의 영향.다중 계층 Transformer 인코더는 N_l개의 동일한 계층으로 구성됩니다.각 계층의 다중 헤드 셀프 어텐션을 통해 모델은 정보를 N_h개의 표현 부분 공간으로 분해하고 다른 위치에서 차별적인 정보를 공동으로 캡처할 수 있습니다.MTE 모듈의 효과는 하이퍼 매개변수, 즉 N_l 및 N_h에 따라 달라집니다.따라서 표 8에서 RAF-DB, FERPlus 및 AffectNet에 대해 4에서 12까지의 다른 하이퍼 매개변수 설정을 사용한 MTE의 효과를 조사합니다.표 8은 정확도/평균 정확도 및 매개변수 측면에서 방법의 성능을 비교합니다.계층 수 N_l를 늘리면 전체 네트워크의 매개변수에 큰 부담이 된다는 것을 관찰했습니다.일반적으로 작은 N_l 값은 더 나은 인식 성능을 얻는 경향이 있으며 큰 N_l 값은 과대적합의 위험이 있는 과도한 매개변수를 초래합니다. N_h의 값이 작을수록 잠재 표현을 학습할 수 있는 부분 공간이 충분하지 않기 때문에 성능이 더 나빠질 수 있습니다. 특히, 이 모델은 N_l ¼ 4 및 N_h ¼ 8인 RAF 데이터 세트에서 가장 좋은 성능을 생성하는 것을 볼 수 있습니다. 그러나 N_l ¼ 12 및 N_h ¼ 12인 경우 MTE가 부정적인 효과를 가져올 수도 있으며 이 모델은 더 나쁜 결과를 생성합니다.

표 9 ViT와 우리의 VTFF의 비교

Setting	N_l	N_h	Params(M)	RAF-DB	FERPlus	AffectNet
ViT	12	12	85.8	47.55	47.72	27.87
ViT*	12	12	85.8	85.14	88.07	58.77
Ours	4	8	51.8	82.27	84.80	58.75
Ours*	4	8	51.8	88.14	88.69	61.55
p-values	-	-	-	$3.47 \times e^{-6}$	$7.30 \times e^{-6}$	$3.39 \times e^{-6}$
p-values*	-	-	-	$6.68 \times e^{-5}$	$1.18 \times e^{-4}$	$2.57 \times e^{-5}$

RAF-DB, FERPlus 및 AffectNet에 대한 이러한 방법의 p-값은 마지막 두 행에 제공됩니다. *으로 표시된 방법은 사전 학습된 가중치를 상속함을 나타내며, 그렇지 않으면 방법은 처음부터 학습됩니다.

표 10 RAF-DB, FERPlus 및 Affectnet에 대한 우리 방법의 교육 비용(GPU-일)

Method	RAF-DB	FERPlus	AffectNet
Ours	0.06	0.07	0.36
Ours*	0.06	0.07	0.35

표 6에 표시된 대로 MTE가 없는 모델보다 RAF 데이터 세트에서 인식 정확도가 더 높습니다. 그 이유는 모델 매개변수의 수가 N_l와 N_h의 증가에 따라 극적으로 증가하고 N_l ¼ 12와 N_h ¼ 12인 모델의 108.5M 매개변수가 RAF와 같은 작은 데이터 세트에서 학습하여 효과적으로 최적화하기에는 너무 많을 수 있기 때문입니다. AffectNet과 같은 더 큰 데이터 세트에서는 MTE의 이러한 부정적인 영향이 관찰되지 않습니다.

사전 훈련된 가중치의 영향. 사전 훈련된 가중치가 우리 방법에 미치는 영향을 알아내기 위해 우리는 모델을 처음부터 훈련하거나 사전 훈련된 ResNet18 가중치에서 모델을 미세 조정합니다. 또한 표 9에서 ViT와 VTFF를 비교합니다. 이 세 가지 FER 데이터 세트에 ViT-Base[49]를 구현하고 [49]에서 설명한 대로 기본 설정에 따라 실험을 수행합니다. ViT는 처음부터 훈련되었음을 나타내고 ViT*은 ImageNet-21k 및 ImageNet에서 사전 훈련된 가중치로 미세 조정되었음을 나타냅니다. ViT* 및 Ours* 설정에서 우리 방법이 더 적은 매개변수로 더 나은 성능을 달성한다는 것을 알 수 있습니다. 미세 조정 대신 처음부터 훈련할 경우 ViT의 성능이 크게 떨어집니다. ViT의 특징 추출 용량은 대규모 데이터 세트의 지침 없이는 비교적 제한되기 때문입니다. 표 9는 사전 훈련된 가중치에서 모델을 미세 조정하면 일반적으로 더 나은 성능을 얻을 수 있음을 보여줍니다. 각 데이터 세트에 대한 방법의 훈련 비용은 표 10에서 볼 수 있습니다. 이론적으로 동일한 방법에 대한 훈련 비용은 동일해야 합니다. 그러나 실제로는 일부 다른 프로그램이 GPU 리소스를 할당하고 데이터 I/O 장치가 다양한 조건에서 다르게 작동할 수 있기 때문에 동일한 방법에 대한 실험 간에 약간의 차이가 있습니다.

또한 우리는 두 모델 간의 차이가 통계적으로 유의미하다는 것을 검증하기 위해 결과에 대한 통계적 증거를 제공합니다. McNemar의 검정은 제안된 방법과 ViT의 인식 정확도를 비교하는 데 사용됩니다. 우리는 두 모델의 성능이 동일하다는 귀무 가설을 수립합니다. 따라서 대안 가설은 다음과 같이 정의됩니다. 두 모델 중 하나가 다른 모델보다 성능이 더 좋습니다. 통계적 유의성을 평가하기 위해 우리는 유의성을 다음과 같이 설정합니다.

표 11 ViT에 대한 ASF의 영향 및 우리의 방법

Method	RAF-DB	FERPlus	AffectNet
ViT* w/o ASF	85.07 ± 0.30	87.93 ± 0.52	58.70 ± 0.09
ViT* w/ ASF	65.51 ± 0.39	69.03 ± 0.46	27.31 ± 0.18
Ours* w/o ASF	87.83 ± 0.04	88.15 ± 0.23	60.90 ± 0.12
Ours* w/ ASF	88.19 ± 0.21	88.70 ± 0.17	61.52 ± 0.07

우리는 각 데이터 세트에서 난수 시드를 사용하여 실험을 세 번씩 실행하고 인식 정확도의 평균과 표준 편차를 계산합니다.

모든 실험에 대한 p-값을 계산하기 위해 임계값 a ¼ 0:05를 사용했습니다. 모든 데이터 세트에 대한 결과 p-값은 표 9에 보고되었습니다. 이러한 결과에 따르면 모든 데이터 세트에서 p -값이 유의 수준 0.05보다 작았기 때문에 귀무 가설이 기각되었으며, 따라서 ViT와 우리 방법 간의 차이가 통계적으로 유의미하다는 결론을 내렸습니다. 또한 표 9에서 우리 모델이 얻은 개선 사항이 우연이 아니라는 것을 추론할 수 있습니다.

ASF가 ViT와 VTFF에 미치는 영향. ASF가 ViT와 VTFF에 미치는 영향을 조사하기 위해 ASF가 있는 방법과 없는 방법을 평가했습니다. 실험 결과는 표 11에 나와 있습니다. 동일한 데이터 설정에서 바닐라 ViT(ASF가 없는 ViT)는 모든 데이터 세트에서 우리 방법(ASF가 있는 우리 방법)에 비해 잘 작동하지 않았습니다. 그 이유는 다음과 같이 요약할 수 있다고 생각합니다. a) ViT의 매개변수가 많으면 미세 조정 시 비교적 작은 FER 데이터 세트에 큰 부담이 됩니다. 어느 정도 이는 다른 레이어 값 N_l과 헤드 수 N_h에 대한 절제 연구와 일치합니다. N_l와 N_h이 클수록 성능이 저하될 수 있습니다. b) vanilla ViT는 FER 불변 작업을 위해 설계되었으며 FER에 대한 어떤 것도 고려하지 않는 반면, 우리 모델은 MS-Celeb-1M 얼굴 인식 데이터 세트에서 사전 학습된 ResNet을 통합하고 기능 융합을 위한 ASF 모듈을 설계합니다. 우리 방법은 Transformers의 전역 표현 용량을 상속할 뿐만 아니라 FER을 위한 작업별 모듈로 구성됩니다.

그러나 ASF를 직접 적용하여 ViT의 입력을 융합하는 것은 ASF가 일종의 피처 수준 융합 방법이기 때문에 실행 불가능해 보입니다. 구체적으로, ASF는 채널 차원을 따라 전역 및 로컬 피처 융합을 수행합니다. 우리 방법의 채널 번호는 512인 반면 ViT의 입력 채널 번호는 3입니다. ASF는 이렇게 작은 채널 번호로 피처 맵을 융합할 때 효과적이지 않을 수 있습니다. 우리는 가설을 검증하기 위해 입력에 대한 주의 선택적 융합으로 ViT를 훈련합니다. 표 11의 ViT w/o ASF 및 ViT w/ ASF 방법에서 볼 수 있듯이 ASF가 있는 ViT의 성능은 ASF가 없는 ViT에 비해 극적으로 떨어집니다. 또한 표 11의 마지막 두 행에서 ASF가 포함된 모델이 인식 성능을 더욱 향상시킨다는 결론을 내릴 수 있으며, 이는 피처 수준 융합에 대한 ASF의 효과를 보여줍니다.

4.5 시각화

우리의 방법은 시각적 단어 간의 관계를 계산하고 표현 인식을 위한 차별적 특징을 포착합니다. 표 4는 우리 방법의 효과를 뒷받침하는 경험적 증거를 제공합니다. 우리 방법의 뛰어난 용량을 시각적으로 보여주기 위해 원시 이미지와 그

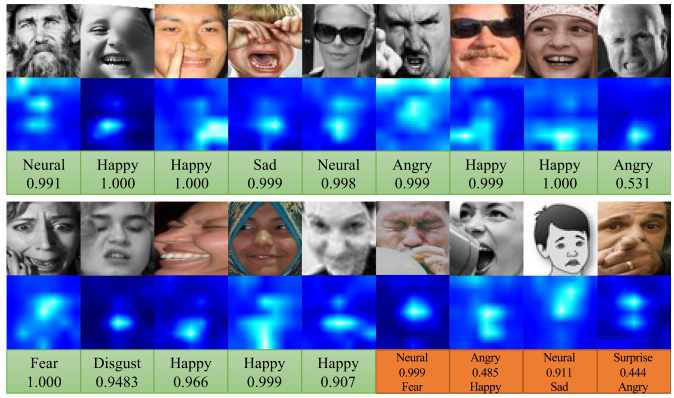


그림 4. 학습된 주의 맵의 그림. 첫 번째 행은 원시 이미지를 보여주고, 두 번째 행은 MTE의 주의 맵이며, 세 번째 행에는 해당 예측 또는 레이블이 표시됩니다. 녹색으로 표시된 주석은 예측된 표현과 신뢰도 점수이고, 주황색으로 표시된 다른 주석은 기준 진실 레이블이 주석된 거짓 예측입니다.

해당 주의 가중치 맵은 그림 4에 시각화되어 있습니다. 우리는 RAF-DB, FERPlus 및 AffectNet에서 여러 이미지를 무작위로 선택하고 모든 헤드에 걸쳐 주의 가중치 맵을 계산합니다. 그런 다음 모든 레이어의 가중치를 재귀적으로 곱하고 입력 이미지 공간으로 투영합니다.

그림 4에서 보듯이, 가중치 맵은 우리 방법이 폐색 영역을 식별하고 차별적 특징을 강조할 수 있음을 보여줍니다. 예를 들어, 세 번째 원시 이미지는 왼쪽 부분에 손가락으로 폐색되어 있는 반면, 우리 방법은 오른쪽 부분에 높은 주의 가중치를 제공합니다. 여덟 번째 원시 이미지의 가중치 맵은 하단에 높은 주의가 포함되어 있어 입 영역을 덮고 있습니다. 또한 그림 4에서 몇 가지 잘못된 예측을 제공합니다. 첫 번째 거짓 양성 샘플은 한 남자가 터질 풍선을 불고 있고 두려움에 눈을 감고 있음을 나타냅니다. 이는 탐지 프로세스가 양날의 검이며 귀중한 얼굴 위치를 제공하지만 배경 지식을 제외한다는 것을 보여줍니다. 만화 얼굴은 불행히도 이러한 데이터 세트에 만화 얼굴이 제한되어 잘못 예측됩니다. 다른 두 개의 거짓 양성 샘플도 사회적 규범과 같은 외부 지식을 향후 연구 작업에 고려해야 함을 나타냅니다. 그림 4에서 알 수 있듯이, 우리 방법은 이러한 시각적 단어들 간의 관계를 동적으로 모델링하고 차별 영역을 강조하여 인식 성능을 높일 수 있다는 결론을 내릴 수 있습니다.

5 결론

이 논문에서 우리는 야생에서 얼굴 표정 인식을 위한 Visual Transformers with Feature Fusion(VTFF)을 제시합니다. 우리는 얼굴 이미지를 시각적 단어 시퀀스로 변환하고 글로벌 관점에서 인식을 수행하여 표정 인식 문제를 해결하는 것을 제안합니다. 이러한 목표를 달성하기 위해 우리는 인식 정확도를 향상시키기 위해 LBP 특징과 CNN 특징을 동적으로 적응적으로 결합하는 주의 선택적 융합을 설계합니다. 시각적 단어는 융합된 특징 맵을 평면화하고 투사하여 생성됩니다. 다층 Transformer 인코더는 글로벌 자기 주의 메커니즘을 활용하여 글로벌 관점에서 차별적인 시각적 단어로 주의를 전환합니다. 실험 결과는 VTFF가 자주 사용되는 세 가지 얼굴 인식에서 다른 최첨단 방법을 능가한다는 것을 보여줍니다.

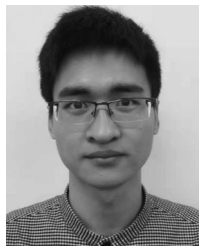
표현 데이터 세트, 즉 RAF-DB, FERPlus, AffectNet. CK+에 대한 교차 데이터 세트 평가는 또한 우리 방법의 유망한 일반화 능력을 보여줍니다.

참고문헌

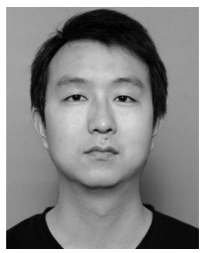
[1] C. Darwin 및 P. Prodger, 인간과 동물의 감정 표현.영국 런던: 미국 옥스포드 대학 출판부, 1998년. [2] Y.-I. Tian, T. Kanade 및 J. F. Cohn, "얼굴 표정 분석을 위한 동작 단위 인식", IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 2, pp. 97 – 115, 2001년 2월. [3] N. Dalal 및 B. Triggs, "인간 감지를 위한 지향성 기울기 히스토그램", Proc. IEEE Conf. Comput. Vis. Pattern Recog- nit., 2005, pp. 886 – 893. [4] C. Shan, S. Gong 및 P. W. McOwan, "로컬 이진 패턴을 사용한 강력한 얼굴 표정 인식", Proc. IEEE Int. Conf. Image Process., 2005, pp. II – 370. [5] X. Feng, M. Pietikainen, 및 A. Hadid, "로컬 이진 패턴 및 선형 프로그래밍을 사용한 얼굴 표정 인식", 패턴 인식 이미지 분석., vol. 15, no. 2, 2005, Art. no. 546. [6] I. Buciu 및 I. Pitas, "얼굴 표정 인식에 대한 음이 아닌 및 로컬 음이 아닌 행렬 인수 분해의 응용", Proc. 17th Int. Conf. Pattern Recognit., 2004, pp. 288 – 291. [7] S. H. Lee, K. N. Plataniotis, 및 Y. M. Ro, "희소 표현 기반 얼굴 표정 인식을 위한 훈련 표정 이미지를 사용한 클래스 내 변화 감소", IEEE Trans. Affective Comput., vol. 5, no. 3, pp. 340 – 351, 2014년 3분기. [8] S. Li 및 W. Deng, "제약 없는 얼굴 표정 인식을 위한 신뢰할 수 있는 크라우드소싱 및 심층적 지역성 보존 학습", IEEE Trans. Image Process., vol. 28, no. 1, pp. 356 – 370, 2019년 1월. [9] Y. Li, J. Zeng, S. Shan 및 X. Chen, "주목 메커니즘을 갖춘 CNN을 사용한 폐색 인식 얼굴 표정 인식", IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439 – 2450, 2019년 5월. [10] K. Wang, X. Peng, J. Yang, D. Meng 및 Y. Qiao, "포즈 및 폐색에 강한 얼굴 표정 인식을 위한 영역 주의 네트워크", IEEE Trans. Image Process., vol. 29, pp. 4057 – 4069, 2020년 1월. [11] K. Wang, X. Peng, J. Yang, S. Lu, 및 Y. Qiao, "대규모 얼굴 표정 인식을 위한 불확실성 억제", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6897 – 6906. [12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, 및 I. Matthews, "확장된 cohn-kanade 데이터 세트(CK+): 동작 단위 및 감정 지정 표현을 위한 완전한 데이터 세트", Proc. IEEE Conf. Comput. Vis. Pattern Recognit.-Workshops, 2010, pp. 94 – 101. [13] M. Valstar 및 M. Pantic, "유도된 혐오감, 행복감 및 놀라움: mmi 얼굴 표정 데이터베이스에 추가", Proc. 3rd Intern. Workshop EMOTION, Corpora Res. Emotion Affect, 2010, pp. 65 – 70. [14] G. Zhao, X. Huang, M. Taini, S. Z. Li 및 M. Pietik € a Inen, "근적외선 비디오에서 얼굴 표정 인식", Image Vis. Comput., vol. 29, no. 9, pp. 607 – 619, 2011. [15] E. Barsoum, C. Zhang, C. C. Ferrer 및 Z. Zhang, "군중 소싱 레이블 분포를 사용한 얼굴 표정 인식을 위한 딥 네트워크 훈련", Proc. 18th ACM Int. Conf. :Multimodal Inter-act., 2016, pp. 279 – 283. [16] A. Mollahosseini, B. Hasani, 및 M. H. Mahoor, "AffectNet: 야생에서의 얼굴 표정, 가치 및 각성 컴퓨팅을 위한 데이터베이스", IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 18 – 31, 2019년 1분기. [17] Y. LeCun 외, "필기 우편번호 인식에 적용된 역전파", Neural Comput., vol. 1, no. 4, pp. 541 – 551, 1989. [18] S. Li 및 W. Deng, "심층 얼굴 표정 인식: 조사", IEEE Trans. 영어: 출판 예정인 Affective Comput., doi: 10.1109/TAFFC.2020.2981446. [19] K. Zhang, Z. Zhang, Z. Li, 및 Y. Qiao, "멀티태스크 계단식 합성곱 네트워크를 사용한 공동 얼굴 감지 및 정렬", IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499 – 1503, 2016년 10월. [20] B. Amos, B. Ludwiczuk, 및 M. Satyanarayanan, "OpenFace: 모바일 애플리케이션을 갖춘 범용 얼굴 인식 라이브러리", CMU School Comput. Sci., 2016년. [21] F. Bourel, C. C. Chibelushi, 및 A. A. Low, "폐쇄가 있는 경우의 얼굴 표정 인식", Proc. Brit. Mach. Vis. Conf., 2001, pp. 1 – 10. [22] S. Happy 및 A. Routray, "눈에 띄는 얼굴 패치의 특징을 사용한 자동 얼굴 표정 인식", IEEE Trans. Affective Comput., vol. 6, no. 1, pp. 1 – 12, 2014년 1분기.

[23] Y. Li, J. Zeng, S. Shan, and X. Chen, "가림 인식 얼굴 표정 인식을 위한 패치 게이트 CNN", Proc. 24th Int. Conf. Pattern Recognit., 2018, pp. 2209 – 2214. [24] Y. Fan, V. Li, and J. C. Lam, "심층 감독 주의 네트워크를 통한 얼굴 표정 인식", IEEE Trans. Affective Comp- Put., 출판 예정, doi: 10.1109/TAFFC.2020.2988264. [25] J. Li, K. Jin, D. Zhou, N. Kubota, and Z. Ju, "얼굴 표정 인식을 위한 주의 메커니즘 기반 CNN", Neurocomputing, vol. 411, pp. 340 – 350, 2020. [26] W. Zheng, "그룹 희소 감소 랭크 회귀에 기반한 다중 뷰 얼굴 표정 인식", IEEE Trans. Affective Comp- Put., vol. 5, no. 1, pp. 71 – 85, 2014년 1분기. [27] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "다중 뷰 얼굴 표정 인식을 위한 다중 채널 포즈 인식 합성 신경망", Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recogn- nit., 2018, pp. 458 – 465. [28] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "얼굴 표정 인식을 위한 조인트 포즈 및 표정 모델링", Proc. IEEE Conf. Comput. Vis. 한국어: Pattern Recognit., pp. 3359 – 3368, 2018년 6월. [29] Y.-H. Lai 및 S.-H. Lai, "다중 뷰 얼굴 표정 인식을 위한 생성적 적대 네트워크를 통한 감정 보존 표현 학습", Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit., 2018, pp. 263 – 270. [30] F. Zhang, T. Zhang, Q. Mao 및 C. Xu, "기하학적으로 유도되는 포즈 불변 얼굴 표정 인식", IEEE Trans. Image Process., vol. 29, pp. 4445 – 4460, 2020년 2월. [31] N. Sun, Q. Lu, W. Zheng, J. Liu 및 G. Han, "비지도 교차 뷰 얼굴 표정 이미지 생성 및 인식", IEEE Trans. 영어: Affective Comput., 출판 예정, doi: 10.1109/TAFFC.2020.3029531. [32] K. He, X. Zhang, S. Ren, 및 J. Sun, "이미지 인식을 위한 딥 레지듀얼 러닝", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770-778. [33] G. Huang, Z. Liu, L. Van Der Maaten, 및 K. Q. Weinberger, "밀집 연결 합성곱 네트워크", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4700-4708. [34] Y. Tang, "선형 지원 벡터 머신을 사용한 딥 러닝", 2013, arXiv:1306.0239. [35] T. Xu, J. White, S. Kalkan, 및 H. Gunes, "얼굴 표정 인식의 편향과 공정성 조사", Proc. Eur. Conf. Comp-put. Vis., 2020, 506-523쪽. [36] S. Moore 및 R. Bowden, "다중 뷰 얼굴 표정 인식을 위한 로컬 바이너리 패턴", Comput. Vis. Image Understanding, vol. 115, no. 4, 541-558쪽, 2011. [37] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, 및 T. S. Huang, "다중 뷰 얼굴 표정 인식", Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit., 2008, 1-6쪽. [38] K.-H. Pong 및 K.-M. Lam, "얼굴 인식을 위한 다중 해상도 기능 융합", Pattern Recognit., vol. 47, no. 2, pp. 556 – 567, 2014. [39] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "객체 감지를 위한 기능 피라미드 네트워크", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2117 – 2125. [40] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: 고해상도 의미 분할을 위한 다중 경로 정제 네트워크", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1925 – 1934. [41] J. Chen, Z. Chen, Z. Chi 및 H. Fu, "다중 특징 융합을 통한 비디오에서의 얼굴 표정 인식", IEEE Trans. Affective Comp- Put., vol. 9, no. 1, pp. 38-50, 2016년 1분기. [42] J. Shao 및 Y. Qian, "야생에서의 얼굴 표정 인식을 위한 3가지 합성곱 신경망 모델", Neurocomputing, vol. 355, pp. 82-92, 2019. [43] A. Vaswani et al., "주의만 있으면 됩니다", 2017, arXiv: 1706.03762. [44] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, 및 D. Kislyuk, "변압기 기반 객체 감지를 향하여," 2020, arXiv: 2012.09958. [45] S. Yang, Z. Quan, M. Nie, 및 W. Yang, "TransPose: 변압기를 통한 설명 가능한 인간 포즈 추정을 향하여," 2020, arXiv: 2012.14214. [46] P. Esser, R. Rombach, 및 B. Ommer, "고해상도 이미지 합성을 위한 변압기 길들이기," 2020, arXiv:2012.09841. [47] Y. Wang et al., "변압기를 사용한 종단 간 비디오 인스턴스 분할," 2020, arXiv:2011.14503. [48] M. Bhat, J. Francis 및 J. Oh, “ Trajformer: 자율 주행을 위한 로컬 자기 주의 맥락을 갖춘 궤적 예측 ”, 2020, arXiv:2011.14910.

[49] A. Dosovitskiy et al., "이미지는 16x16 단어의 가치가 있습니다: 대규모 이미지 인식을 위한 변환기", 2020, arXiv:2010.11929. [50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, 및 L. Fei-Fei, "ImageNet: 대규모 계층적 이미지 데이터베이스", IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248 – 255. [51] W. Wang et al., "피라미드 비전 변환기: 합성곱 없이 밀도 예측을 위한 다재다능한 백본", 2021, arXiv:2102.12122. [52] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, Q. Tian, "Visformer: 시각 친화적 변환기," 2021, arXiv:2104.12533. [53] H. Wu 외, "CvT: 시각 변환기에 합성곱 도입," 2021, arXiv:2103.15808. [54] Z. Liu 외, "ConvTransformer: 비디오 프레임 합성을 위한 합성곱 변환기 네트워크," 2020, arXiv: 2011.10185. [55] S. Zhao, H. Cai, H. Liu, J. Zhang, S. Chen, "얼굴 표정 인식을 위한 CNN의 특징 선택 메커니즘," Proc. Brit. Mach. Vis. Conf., 2018, Art. no. 317. [56] Y. Li, G. Lu, J. Li, Z. Zhang, D. Zhang, "다중 레벨 특징 및 주의 메커니즘을 사용한 야생에서의 얼굴 표정 인식", IEEE Trans. Affective Comput., 출판 예정, doi: 10.1109/TAFFC.2020.3031602. [57] C. Huang, "감정 인식을 위한 합성곱 신경망 결합", IEEE MIT 학부 연구 기술 컨퍼런스 논문집, 2017, 1-4쪽. [58] S. Miao, H. Xu, Z. Han, Y. Zhu, "얕은 합성곱 신경망을 사용하여 얼굴 표정 인식", IEEE Access, vol. 7, pp. 78 000 – 78 011, 2019. [59] X. Fan, Z. Deng, K. Wang, X. Peng, 및 Y. Qiao, "불확실성에서 얼굴 표정 인식을 위한 차별 표현 학습", Proc. IEEE Int. Conf. Image Process., 2020, pp. 903 – 907. [60] J. Zeng, S. Shan, 및 X. Chen, "일관되지 않은 주석이 달린 데이터 세트를 사용한 얼굴 표정 인식", Proc. Eur. Conf. Comput. Vis., 2018, pp. 222 – 237. [61] D. P. Kingma 및 J. Ba, "Adam: 확률적 최적화 방법", 2014, arXiv:1412.6980. [62] A. Paszke et al., "PyTorch: 필수 스타일, 고성능 딥 러닝 라이브러리," 2019, arXiv: 1912.01703. [63] S. Happy, A. Dantcheva, F. Bremond, "전체 및 부분 기반 모델에서 추출한 딥 기능을 사용한 표정 인식," Image Vis. Comput., vol. 105, 2021, Art. no. 104038. [64] Y. Li et al., "야생에서 얼굴 표정 인식을 위한 정보적 및 차별적 기능 학습," IEEE Trans. Circuits Syst. Video Technol., 출판 예정, doi 10.1109/TCSVT.2021.3103760. 영어: [65] F. Zhang, M. Xu 및 C. Xu, "노이즈가 있는 데이터를 사용한 야생에서의 약한 감독 얼굴 표정 인식", IEEE Trans.Multimedia, 출판 예정, doi: 10.1109/TMM.2021.3072786. [66] W. Xie, H. Wu, Y. Tian, M. Bai 및 L. Shen, "얼굴 표정 인식을 위한 다단계 이상치 억제 및 클래스 쌍 마진을 사용한 삼중 손실", IEEE Trans.Circuits Syst.Video Technol., 출판 예정, doi: 10.1109/TCSVT.2021.3063052. [67] W. Hayale, P. S. Negi 및 M. Mahoor, "야생에서의 얼굴 표정 인식을 위한 딥 시암 신경망", IEEE Trans. Affective Comput., 출판 예정, doi: 10.1109/TAFFC.2021.3077248. [68] Y. Xia, H. Yu, X. Wang, M. Jian, 및 F.-Y. Wang, "관계 인식 얼굴 표정 인식," IEEE Trans. Cogn. Devel. Syst., 출판 예정, doi: 10.1109/TCDS.2021.3100131.



푸얀 마는 2018년 중국 우한의 중남민족대학 전자정보공학과에서 학사 학위를 받았습니다. 그는 현재 중국 창사의 후난대학 전기정보공학과에서 박사 학위를 취득하기 위해 노력하고 있습니다. 그의 연구 관심사는 이미지 처리, 감정 컴퓨팅, 멀티모달 정보 융합입니다.



Bin Sun(IEEE 회원)은 2010년과 2016년에 각각 중국 창사의 후난 대학에서 제어 과학 및 공학 학사 학위와 박사 학위를 받았습니다. 2017년부터 2019년까지 그는 후난 대학 전기 및 정보 공학 대학에서 전기 공학 박사후 연구원으로 재직했으며 현재 준교수로 재직하고 있습니다. 그의 연구 관심사는 컴퓨터 비전과 인간-로봇 자연적 상호 작용입니다.

Shutao Li(IEEE 펠로우)는 1995년, 1997년, 2001년에 각각 중국 창사의 후난 대학에서 학사, 석사, 박사 학위를 받았습니다. 2001년에 그는 후난 대학 전기정보공학과에 합류했습니다. 2001년 5월부터 2001년 10월까지 그는 홍콩과학기술대학 컴퓨터과학과에서 연구원으로 일했습니다. 2002년 11월부터 2003년 11월까지 그는 런던 대학의 Royal Holloway College에서 박사후 연구원으로 일했습니다. 2005년 4월부터 2005년 6월까지

2005년, 그는 홍콩과학기술대학교 컴퓨터과학과의 방문교수였습니다. 그는 현재 후난대학교 전기정보공학과의 정교수입니다. 그는 200편 이상의 추천 논문을 저술하거나 공동 저술했습니다. 그는 2004년과 2006년에 중국의 2급 국가 과학기술 진보상을 두 번 수상했습니다. 그의 현재 연구 관심사는 이미지 처리, 패턴 인식, 인공 지능입니다. 그는 IEEE 지구과학 및 원격탐사 저널과 IEEE 계측 및 측정 저널의 부편집자입니다. 그는 정보 융합 및 감지 및 이미징 저널의 편집 위원회 위원입니다.