

Master's Thesis Presentation for the 146th Dissertation Evaluation

A Study on Efficient Facial Expression Recognition for Fast Inference using Dilated Mixed Convolutions

빠른 추론을 위한 팽창 혼합 합성곱 기반
효율적 얼굴 표정 인식 연구

Haesung Kim

hskimz1217@cau.ac.kr

3rd Oct, 2025



중앙대학교
CHUNG-ANG UNIVERSITY



Contents

1. **Introduction**
2. **Related Work**
3. **Proposed Method**
4. **Experimental Settings**
5. **Overall Experimental Results**
6. **In-Depth Analysis**
7. **Conclusion and Contribution**
8. **Future Plan**
9. **Reference**

Introduction

Facial Expression Recognition

- Facial Expression Recognition (FER) is the task of recognizing basic human emotions from facial images, and has been applied in areas such as human–computer interaction, smart homes, patient pain monitoring, and driver safety management [1, 2].
- Beyond Lab-Controlled datasets where researchers direct specific facial expressions, the emergence of large-scale Real-World datasets captured naturally has led to active research in deep neural network-based supervised learning FER methodologies [3].



[Figure 1] Sample images of driver FER for safe driving

Lab-Controlled
CK+ (2010)



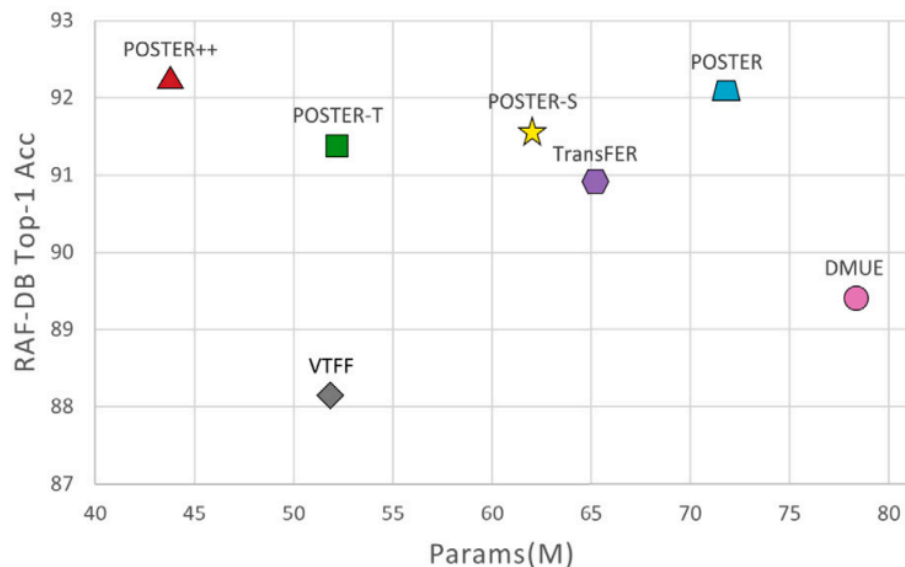
Real-World
RAFDB (2017)

[Figure 2] Comparison between lab-controlled and real-world datasets

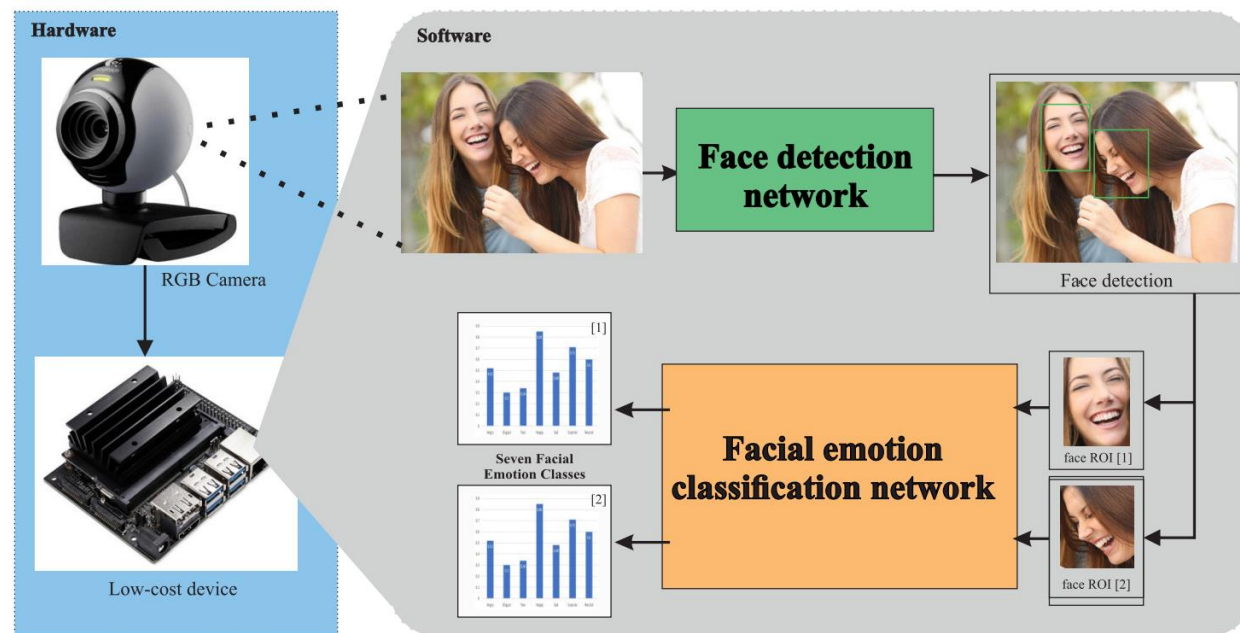
Introduction

Importance of Efficiency on Facial Expression Recognition

- While Real-World datasets reflecting subtle expression differences have driven research in complex deep neural network-based FER, these models involve excessive parameters and high computational costs, limiting practical applications [4].
- Since real-world FER applications such as driver drowsiness detection systems require immediate responses, it is important to ensure fast inference by effectively reducing computational load while maintaining recognition accuracies [5].



[Figure 3] Number of parameters in existing FER models

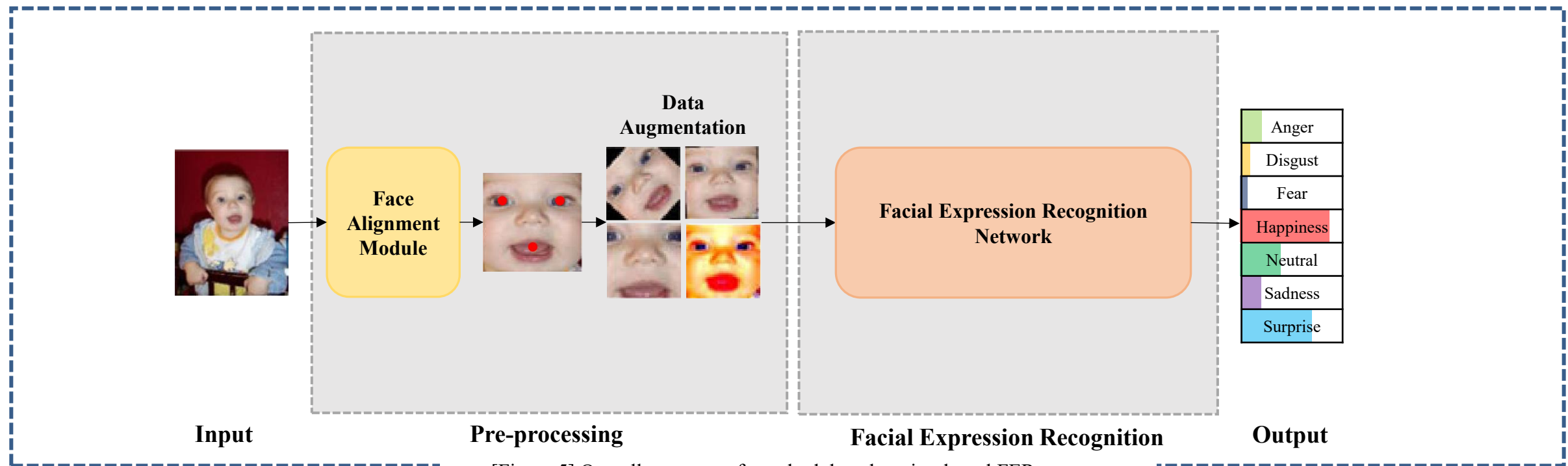


[Figure 4] Processing pipeline of real-time FER system utilizing low-cost devices

Introduction

Objective and Standard Procedure

- FER takes facial images as input and aims to classify images into one of the basic emotion categories by identifying and utilizing important facial regions that are crucial for distinguishing different expressions [6].
- FER takes detected and aligned facial images from Real-World images as input, extracts features through a feature extraction module, learns facial features and labels via deep neural networks, and classifies expressions through the trained model [7].

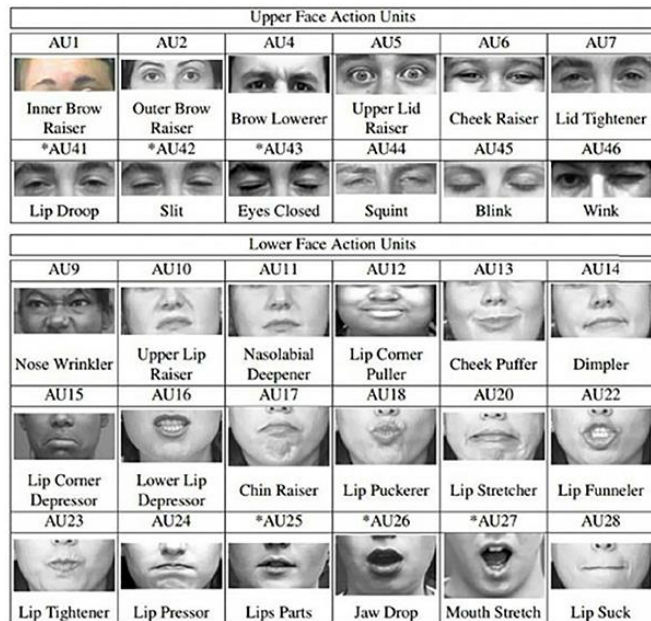


[Figure 5] Overall structure of standard deep learning-based FER systems

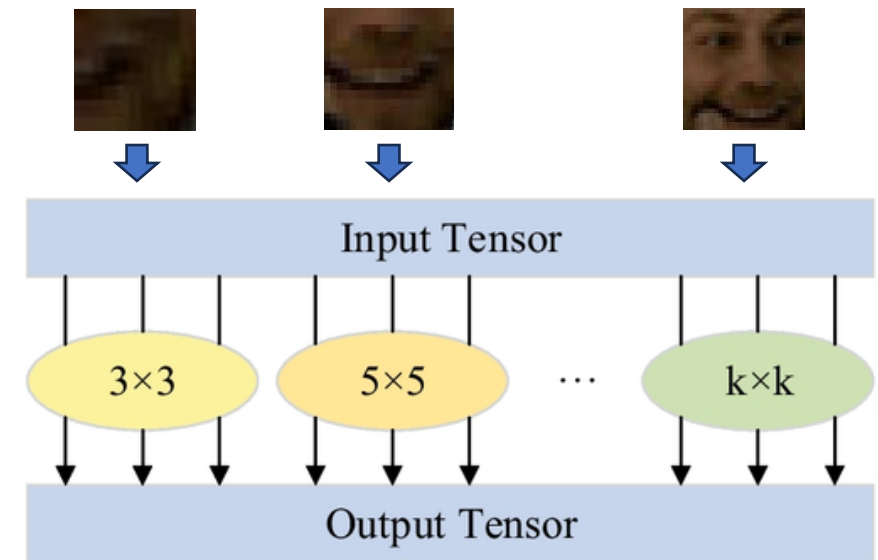
Introduction

Ideal Goal and Challenge

- Expression signals exist at multiple scales, with fine changes like eye wrinkles and mouth corner movements (small-scale) appearing with mouth opening and cheek elevation (medium-to-large scale), making it important to capture interactions across facial regions.
- Securing multiple receptive fields and fusing multi-scale regions inherently involves high computational complexity and increased parameters, hindering real-time performance in device-constrained environments.



[Figure 6] Sample images of Action Units: omponents of facial expressions

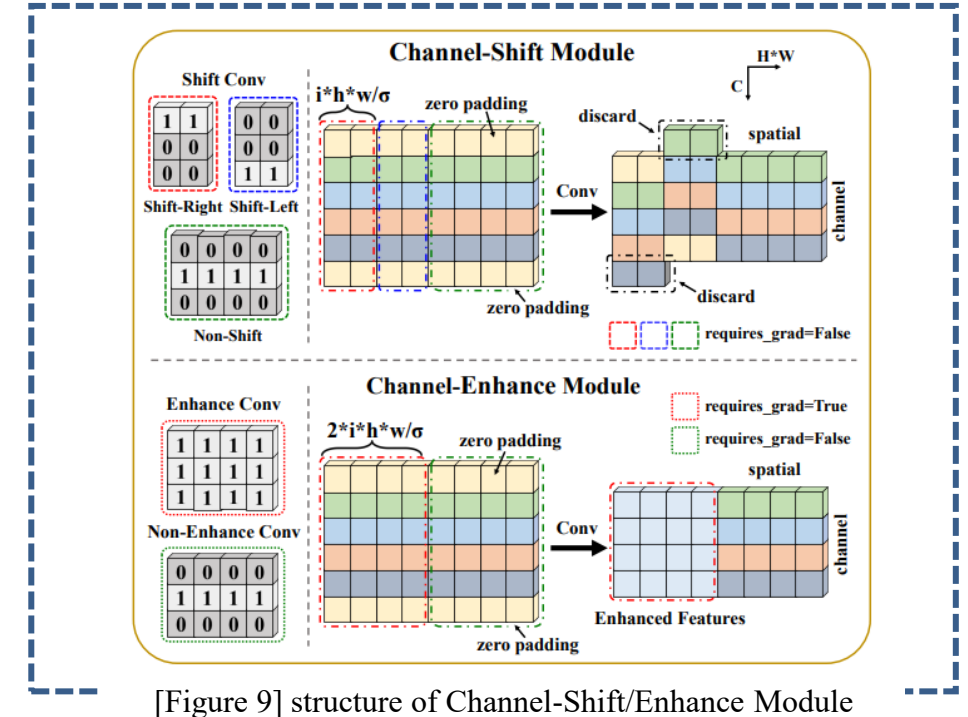
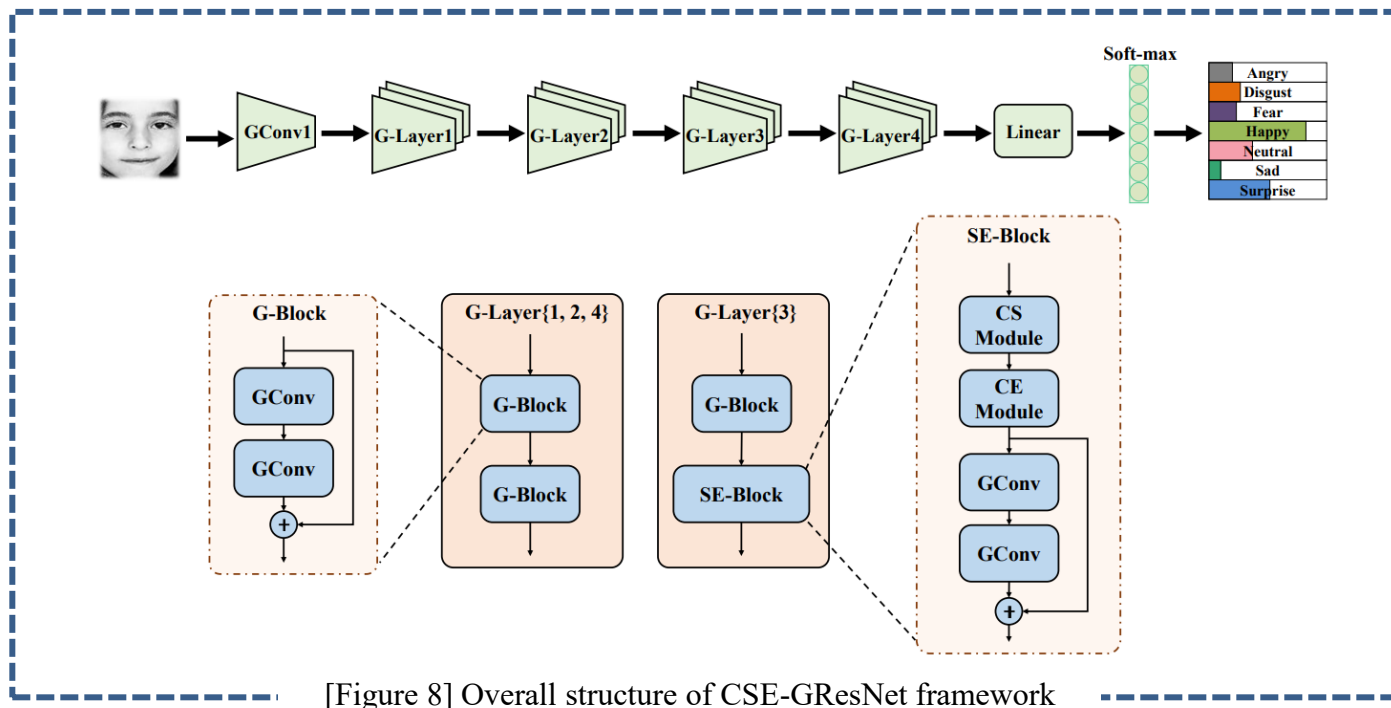


[Figure 7] FER utilizing multiple-scale kernel sizes

Related Work

CSE-GResNet (2025) [17]

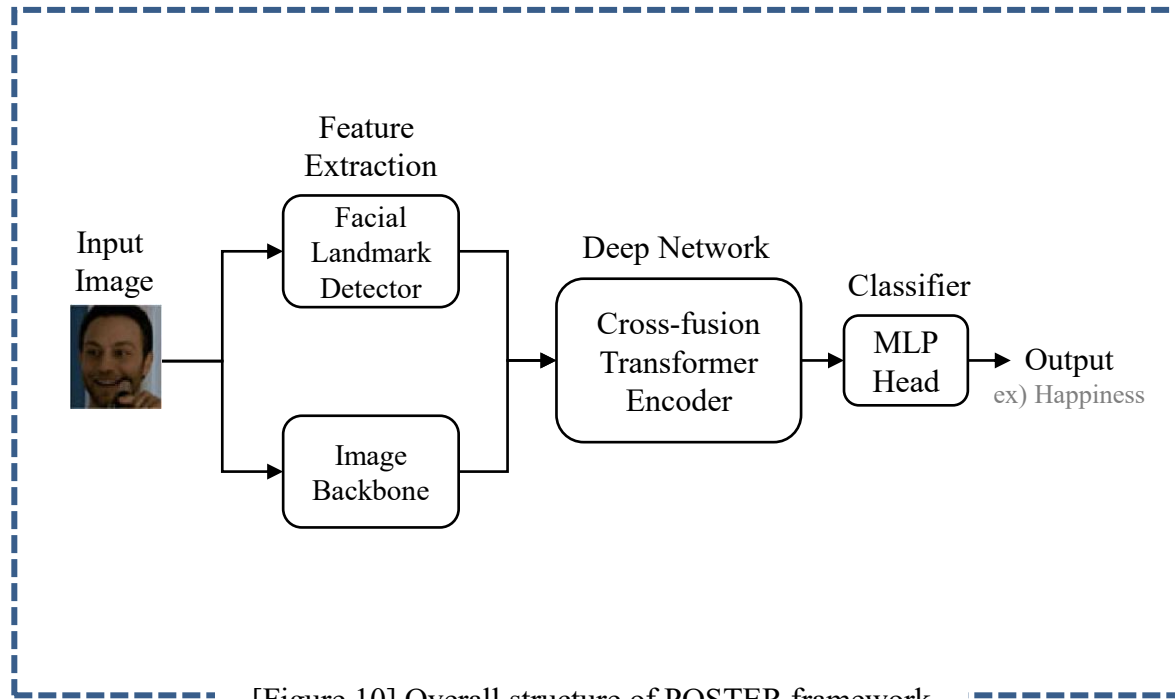
- GResNet is created by applying Gabor convolution to ResNet, then inserting channel-shift and channel-enhancement modules that exchange and strengthen information between adjacent channels, effectively extracting and integrating subtle expression features.
- The proposed CS/CE modules only focus on 'local' information exchange in the channel dimension, failing to secure spatially rich receptive fields and relying on model depth to capture facial region interactions.



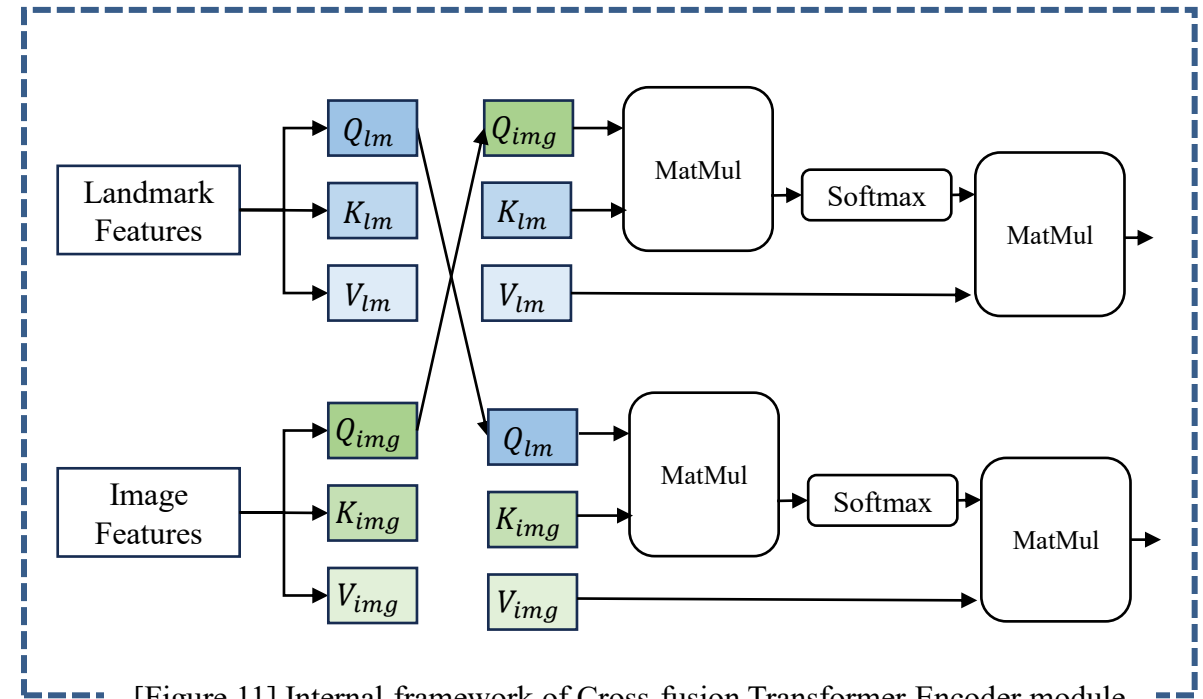
Related Work

POSTER++ (2025) [8]

- POSTER++ enables complementary learning between facial landmark features and global image features through smooth fusion (cross-exchanging Key values), while reducing the complexity of traditional cross-attention through window-based cross attention.
- Although attention mechanisms effectively capture rich interactions between facial regions, the inherent complexity of transformer model architectures is not suitable for real-time performance.



[Figure 10] Overall structure of POSTER framework

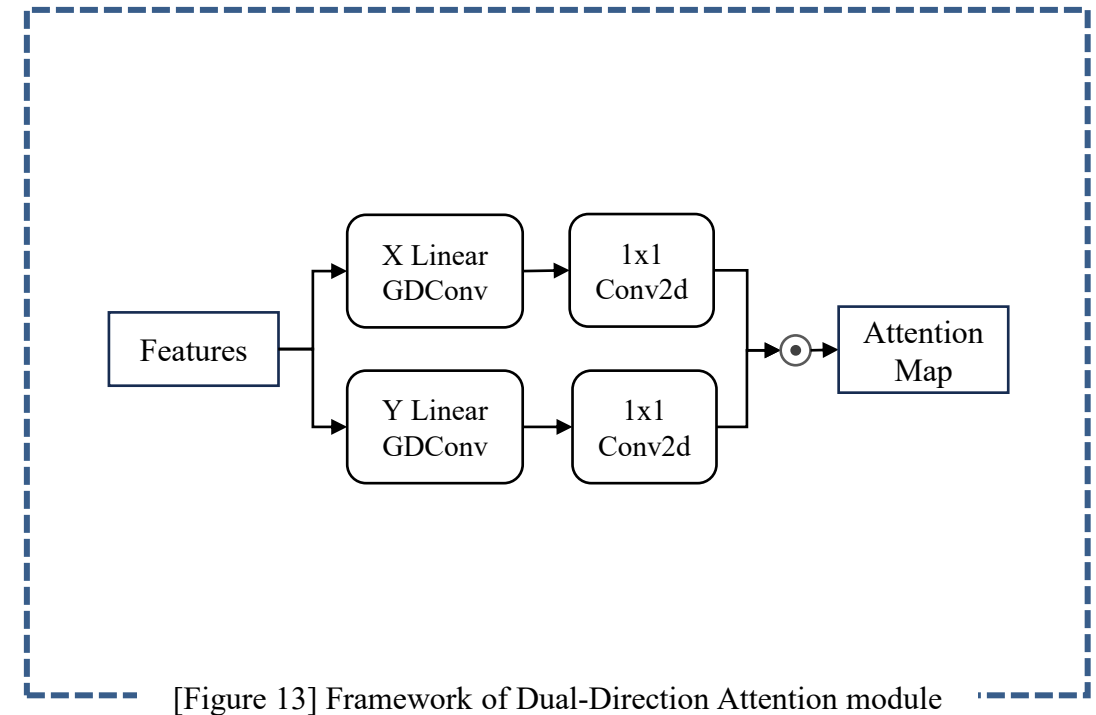
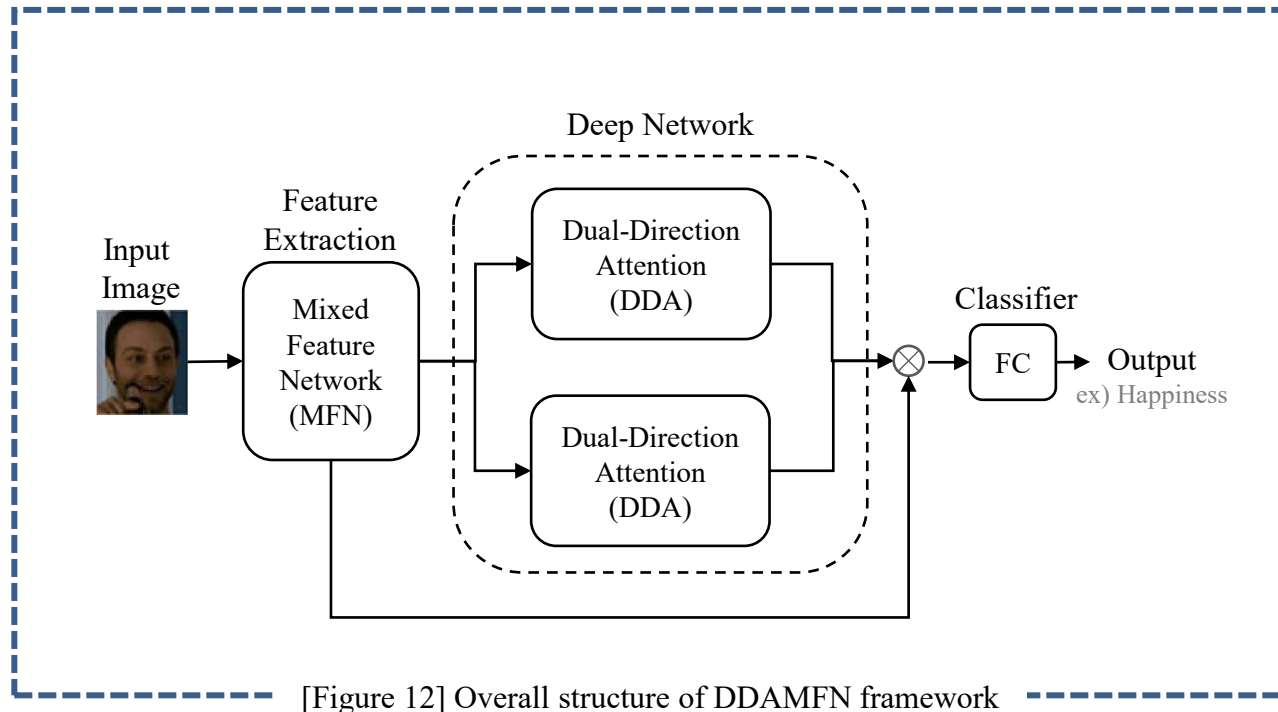


[Figure 11] Internal framework of Cross-fusion Transformer Encoder module

Related Work

DDAMFN (2023) [10]

- DDAMFN extracts rich features by introducing kernels of various sizes (MixConv) to the Mixed Feature Network (MFN) based on MobileFaceNet, and complements global information through the Dual-Direction Attention (DDAN) attention module.
- While applying multiple kernel sizes, it has the limitation of high computational cost by applying large-sized (7x7) convolutions in parallel without efficient alternative techniques such as dilated convolutions.

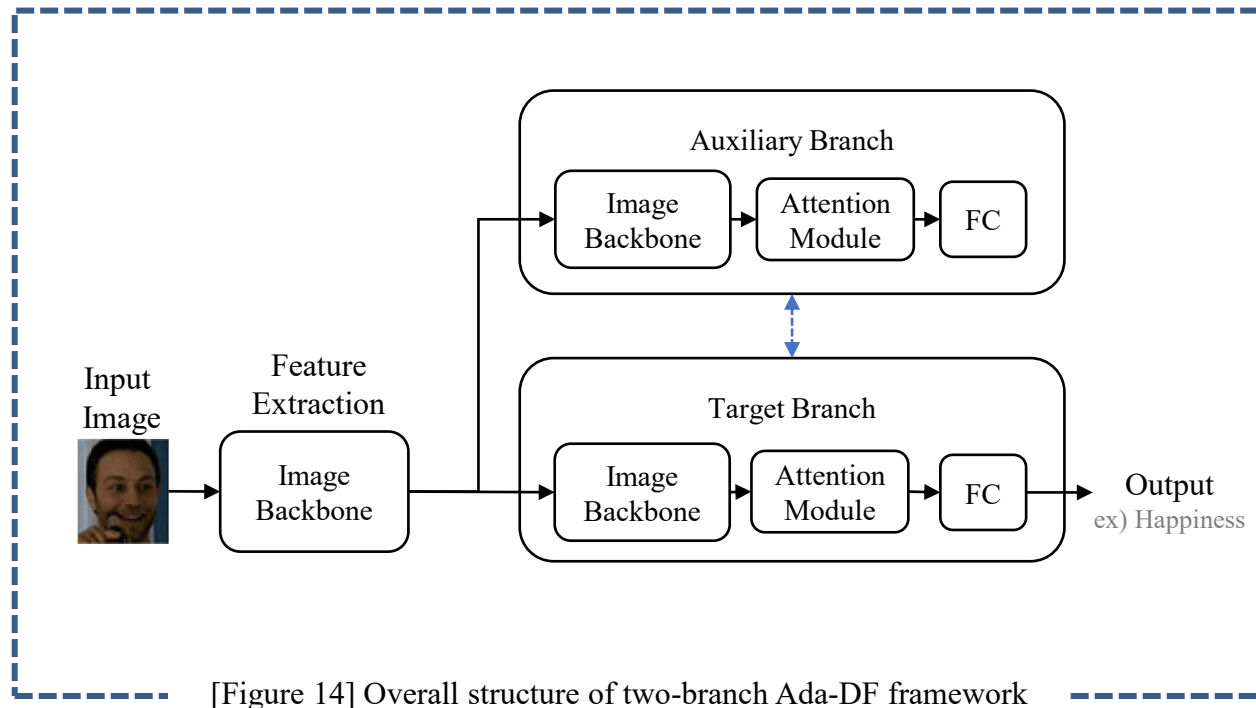


[10] Zhang, Saining, et al. "A dual-direction attention mixed feature network for facial expression recognition." *Electronics* 12.17 (2023): 3595.

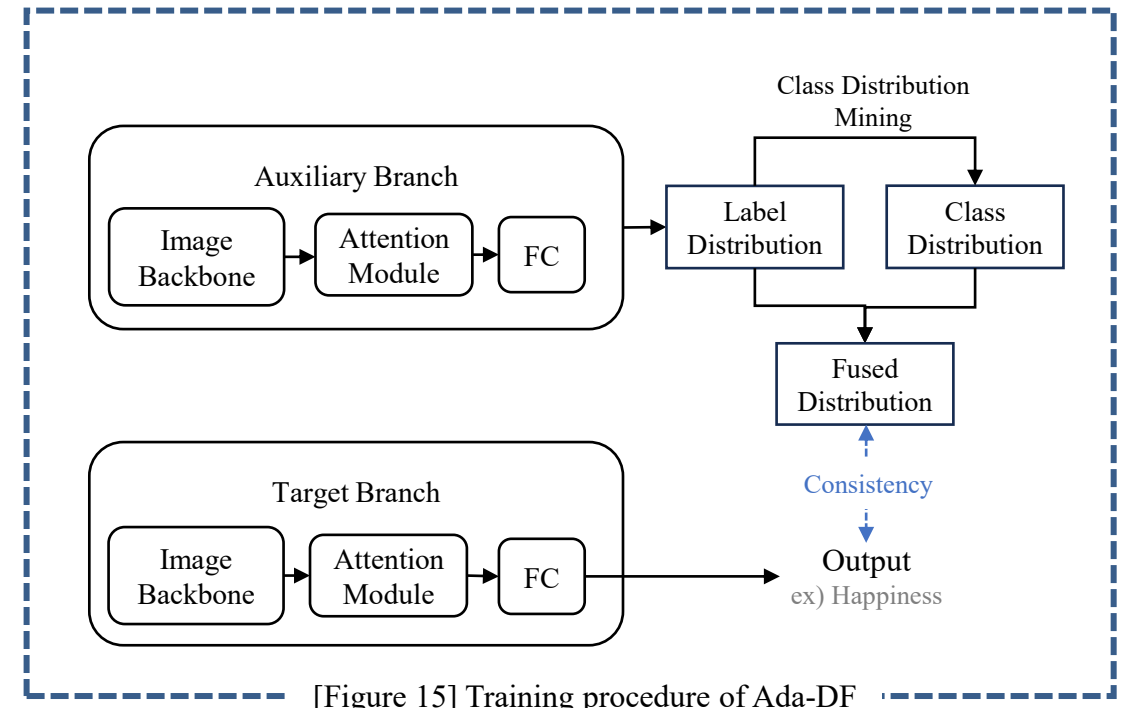
Related Work

Ada-DF (2023) [11]

- Ada-DF uses a dual-branch structure through Dual-Branch Adaptive Distribution Fusion, employing an auxiliary branch during training to reduce label ambiguity, but utilizing only a single-branch structure during inference to secure fast inference.
- Similarly, due to the fixed kernel structural limitations of ResNet backbones, it is difficult to capture associations between various receptive fields, and there exists the limitation that ResNet depth must inevitably increase to capture rich interactions of regions.



[Figure 14] Overall structure of two-branch Ada-DF framework



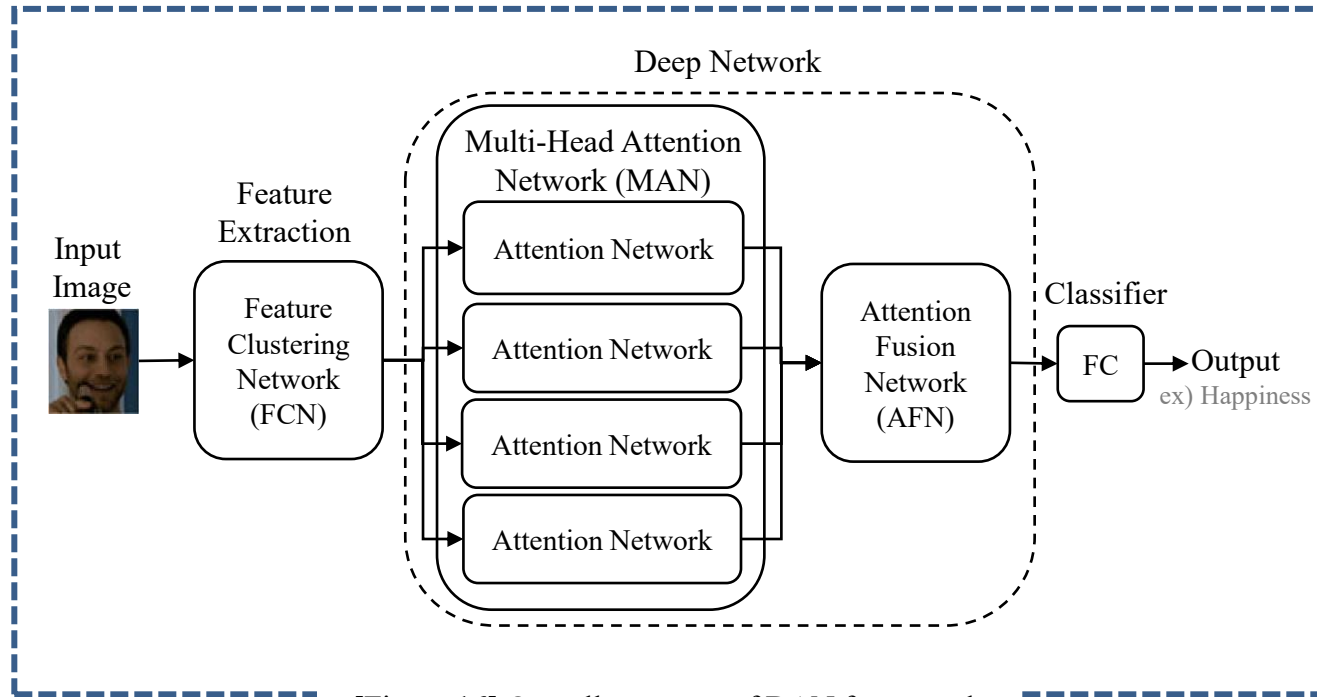
[Figure 15] Training procedure of Ada-DF

[11] Liu, Shu, et al. "A Dual-Branch Adaptive Distribution Fusion Framework for Real-World Facial Expression Recognition." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

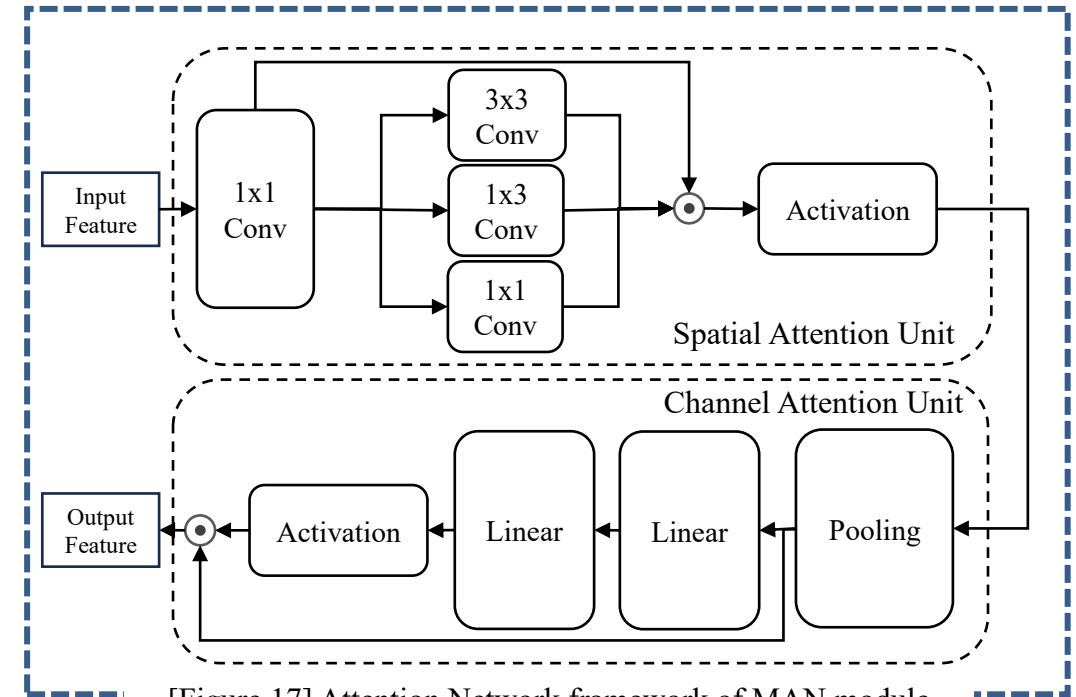
Related Work

DAN (2023) [9]

- DAN enables multiple attention modules to independently focus on key facial regions, allowing for more detailed capture of each region's features, and effectively reflects interactions between key facial regions by learning interactions among these features.
- While multiple attention heads and AFN structures are effective at capturing multi-regional information of facial areas, the parallel processing significantly increases computational load, limiting suitability for real-time applications.



[Figure 16] Overall structure of DAN framework



[Figure 17] Attention Network framework of MAN module

Related Work

Common Drawback and Hypothesis

- Existing methods lack the ability to extract rich multi-scale features—which is crucial for distinguishing subtle expressions—and rely on fixed kernel sizes and model depth to capture facial region interactions, making inference speed degradation inevitable.
- By efficiently capturing multi-scale features and global context, facial region interactions can be captured without relying on network depth, thereby preventing unnecessary inference speed degradation caused by fixed kernels and excessive model depth.

[Table 1] Summary of recent facial expression recognition methods, outlining their key contributions and a common drawback.

Method	Venue	Summary	Common Drawback
CSE-GResNet	IEEE Transactions on Affective Computing 2025	Efficiently fuses facial landmarks and global image features with window-based cross attention	Existing methods lack the ability to extract rich multi-scale features and rely on fixed kernel sizes and model depth to capture facial region interactions, making inference speed degradation inevitable.
POSTER++	Pattern Recognition 2025	Efficiently fuses facial landmarks and global image features with window-based cross attention	
DDAMFN [10]	Electronics 2023	Introduces MixConv and DDAN modules to MobileFaceNet-based MFN for rich feature extraction and global information complementation	
Ada-DF [11]	ICASSP 2023	Reduces label ambiguity with dual-branch during training and secures fast speed with single-branch during inference	
DAN [9]	Biomimetics 2023	Multiple attention modules independently focus on key facial regions and learn inter-regional interactions for detailed feature capture	

Proposed Method

Notation and Definition

- The facial image dataset is defined as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i is a facial image and y_i is an expression label from basic emotion categories such as anger, disgust, fear, happiness, sadness, surprise, neutral. Here, N denotes dataset size and, i denotes the index.
- FER aims to learn a classification function $f: x_i \mapsto y_i$ that maps input facial images x_i to expression labels y_i using dataset $\mathcal{D}$. It learns a deep neural network f based on feature vectors $\phi(x_i)$ extracted from faces to perform inference on unknown facial images.

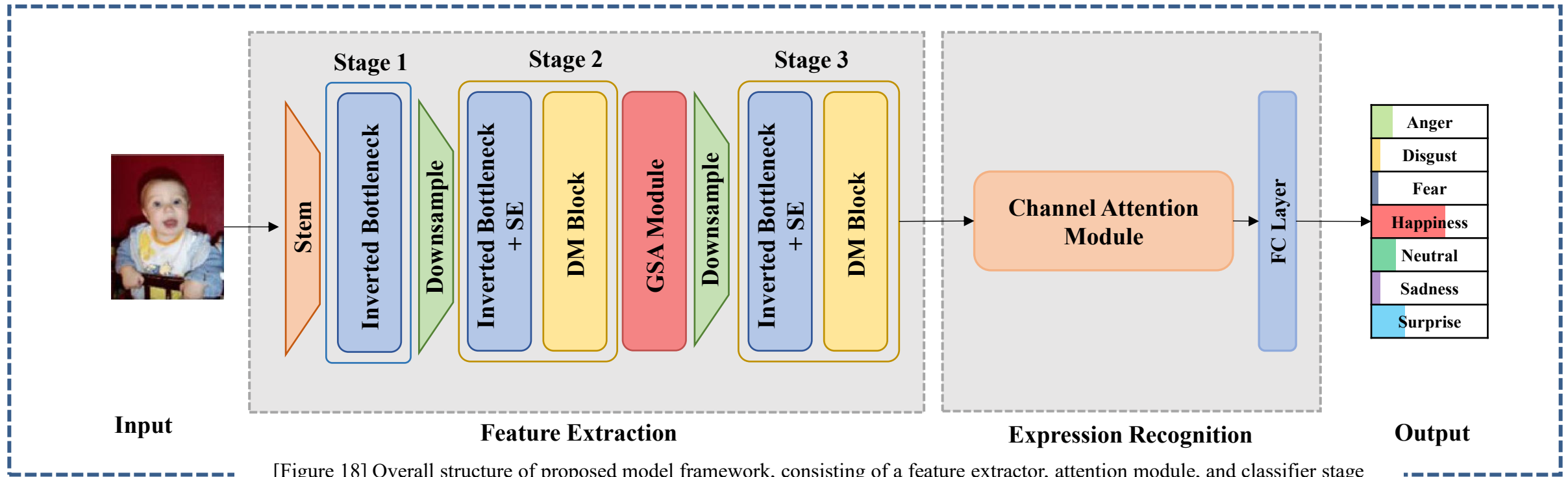
[Table 2] Notations and definitions used in the formulation of the facial expression recognition frameworks

Symbol	Name	Description
$\mathcal{D}$	Dataset	Overall Dataset
x_i	Face Image	Normalized facial image
y_i	Expression Label	Expression label corresponding to one of the basic emotion categories
N	Dataset Size	Dataset size
i	Index	Index for distinguishing individual samples within dataset $\mathcal{D}$
f	Classification Function	Classification function that maps input facial images to expression labels
$\phi(x_i)$	Feature Vector	Feature vector extracted from facial image x_i

Proposed Method

Graphical Abstract

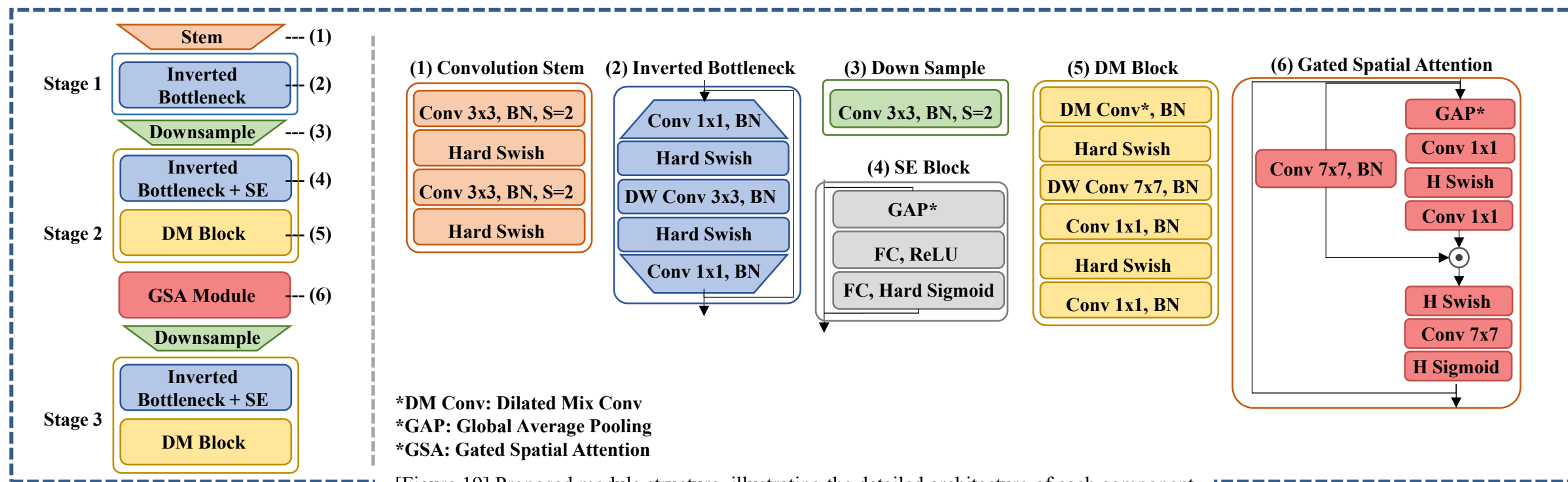
- The proposed model extracts low-level features through Stem and InvertedResidual blocks, then mixes multi-scale context through Dilated Mix Conv (DM) blocks. Subsequently, Gated Spatial Attention (GSA) is applied to reflect global contextual information.
- In the final stage, channels are projected to embedding dimensions and features with compressed spatial information are extracted. This is followed by Channel Attention and classification head to ultimately predict FER classes.



Proposed Method

Details

- Features are extracted through various kernels via Dilated Mixed convolutions, then channels are mixed through wide kernel depthwise convolutions and 1x1 convolutions.
- GSA emphasizes expression-related regions using global average pooling-based context vectors, and due to its relatively high computational cost, it is applied only once after stage 2, while SE blocks are used after stages 2 and 3 to capture channel importance.



Experimental Settings

Dataset

- Lab datasets are formed by researchers instructing subjects to make specific expressions, while Wild dataset images are crawled from the internet, showing significant variations in age, gender, ethnicity, head pose, lighting, occlusion, and post-processing, with annotations by trained annotators.
- 7-Classes expression datasets have seven expression categories: the six basic human emotions defined by Charles Darwin (happiness, anger, sadness, fear, surprise, disgust) plus neutral, while 8-Classes expression datasets additionally include the contempt expression category.

[Table 3] Summary of facial expression datasets, contrasting Lab vs. Wild settings and 7-class vs. 8-class annotations.

*no labels

Dataset	Image Size	Type	#Channels	#Classes	#Images (train/validation/test)	#anger	#disgust	#fear	#happiness	#sadness	#surprise	#neutral	#contempt
CK+ [16]	48 × 48	Lab	1	8	981 (589/196/196)	135	177	75	207	84	249	0	54
SFEW2.0 [19]	143 × 181	Wild	3	7	1634 (859/411/*364)	241	73	122	240	225	145	224	0
RAF-DB [20]	100 × 100	Wild	3	7	15,339 (9205/3067/3067)	705	717	281	4772	1982	1290	2524	0
FERPlus [21]	48 × 48	Wild	1	8	35711 (21463/7154/7154)	3124	253	825	9367	4415	4493	13013	221
FER2013 [22]	48 × 48	Wild	1	7	35887 (21533/7177/7177)	4953	547	5121	8989	6077	4002	6198	0
ExpW [23]	112 × 112	Wild	3	7	90560 (54336/18112/18112)	3671	3995	1088	30537	10559	7060	34883	0

| Experimental Settings

Evaluation Measures

- Accuracy

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Cross Validation

- Number of Iterations: 10
- Repeated Shuffle Split
- Data Split Ratio (Train:Validation:Test) = 6:2:2

Hyperparameter Settings

- Batch size: 32
- Epochs : 100
- Optimizer : AdamW (Weighted Decay=1e-4)
- Learning Rate: 1e-3
- LR Schedule: Cosine Annealing ($T_{max}=100$)

Experimental Settings

Hyperparameter Settings

- Batch size: 32
- Epochs : 100
- Optimizer : AdamW (Weighted Decay=1e-4)
- Learning Rate: 1e-3
- LR Schedule: Cosine Annealing (T_{max} =100)

[Table 4] detailed hyperparameter settings

High-level Categories	Low-level Categories	Option
Training Parameters & Strategy	Epochs	100
	Early Stopping Patience	20
	Batch Size	32
	Model Selection	Epoch with Minimum Validation Loss
Optimizer Settings	Optimizer	AdamW
	Learning Rate	1e-3
	Momentum	0.9
	Weight Decay	1e-4
Learning Rate Scheduler	Scheduler	Cosine annealing scheduler
	Maximum Number of Iterations	100
	Minimum Learning Rate	1e-6
Data Transform	Resize	(112,112)
Data Augmentation	Random Horizontal Flip	50%
	Random Rotation	5-20, 20%
	Random Crop	8-32, 20%
	Random Erasing	(0.02, 0.1)

Experimental Results

Performance Comparison

- The proposed model achieves the highest accuracy on various dataset, outperforming all comparison methods including the recent state-of-the-arts models such as CSE-GResNet.

Evaluation Measure: Accuracy

[Table 5] Comparison results of five FER methods on six datasets in terms of the accuracy

Dataset	Proposed	CSE-GResNet [17]	POSTER++ [12]	Ada-DF [11]	DAN [9]
CK+	0.983 ± 0.015	0.979 ± 0.010	0.944 ± 0.018	0.865 ± 0.017	0.978 ± 0.011
SFEW2.0	0.464 ± 0.003	0.342 ± 0.028	0.247 ± 0.035	0.214 ± 0.028	0.441 ± 0.012
RAFDB	0.802 ± 0.005	0.790 ± 0.008	0.624 ± 0.015	0.733 ± 0.002	0.801 ± 0.009
FERPlus	0.823 ± 0.003	0.811 ± 0.003	0.577 ± 0.024	0.804 ± 0.003	0.815 ± 0.004
FER2013	0.655 ± 0.003	0.639 ± 0.004	0.541 ± 0.011	0.652 ± 0.011	0.653 ± 0.006
ExpW	0.642 ± 0.003	0.632 ± 0.004	0.392 ± 0.007	0.641 ± 0.001	0.639 ± 0.003

Experimental Results

Performance Comparison

- Compared to CSE-GResNet, the proposed approach reduces inference time by 53.5%, while requiring 46.6% fewer FLOPs and 34.4% fewer parameters, demonstrating superior efficiency for real-time applications.

Evaluation Measure: Inferenc Time, FLOPs, Number of Parameters

[Table 6] Comparison results of five FER methods in terms of Inference Time, FLOPs, Number of Parameters

Metric	Proposed	CSE-GResNet [17]	POSTER++ [12]	Ada-DF [11]	DAN [9]
GPU Inference Time (ms)	5.516 ± 0.242	11.866 ± 0.376	37.588 ± 0.385	13.594 ± 0.187	15.808 ± 0.128
FLOPs (G)	0.921	1.726	8.403	2.872	2.249
#Params	1.842	2.807	43.70	22.204	19.724

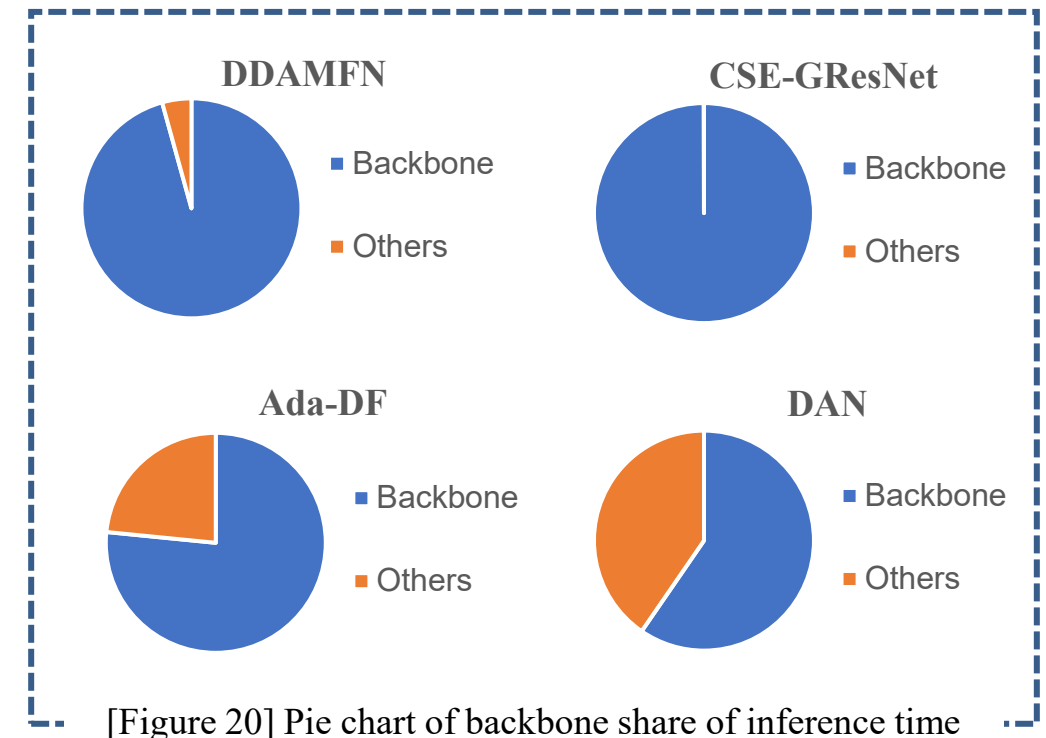
Experimental Results

In-depth Analysis

- Facial expression recognition systems conventionally follow a three-stage pipeline comprising feature extraction, attention modules, and classification layers.
- As the backbone network accounts for the dominant portion of total inference time in FER, computational reduction at the feature extraction stage is inherently crucial for real-time performance.

[Table 7] Ratio of backbone to total inference time per image, by counterpart models

Model	Backbone	Total Inference Time (ms)	Feature Extraction Time (ms)	Percent of Backbone
DDAMFN	MixedFeatureNet	21.176	20.265	95.70%
CSE-GResNet	GResNet	11.866	11.863	99.98%
Ada-DF	ResNet-18	1.759	1.340	76.15%
DAN	ResNet-18	4.046	2.411	59.60%



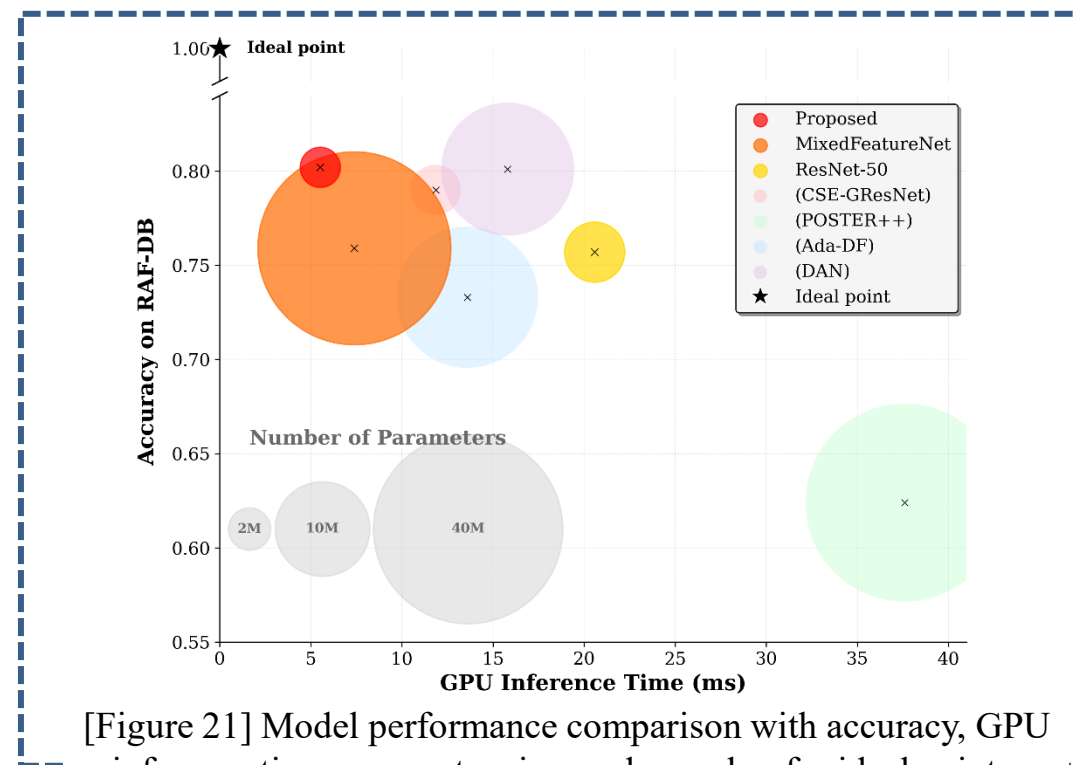
Experimental Results

In-depth Analysis

- To demonstrate the effectiveness of our backbone architecture, we conducted comparative experiments by integrating different backbone networks with a simple channel attention module and measuring their recognition performance.
- Compared to widely-used efficient backbones including ResNet-50 and MixedFeatureNet, our architecture exhibits advantages across multiple dimensions: reduced computational complexity, faster inference time while achieving higher accuracy

[Table 8] Performance comparison of backbone architectures: inference time, FLOPs, parameters, and RAF-DB accuracy

Metric	Proposed Backbone	MixedFeatureNet	ResNet-50
GPU Inference Time (ms)	5.516 ± 0.242	20.583 ± 0.357	7.393 ± 0.196
CPU Inference Time (ms)	26.759	56.035	39.103
FLOPs (G)	0.919	0.569	6.314
#Params	1.801	4.042	41.567
Accuracy on RAFDB	0.802 ± 0.005	0.757 ± 0.006	0.759 ± 0.011



[Figure 21] Model performance comparison with accuracy, GPU inference time, parameter size, and a marker for ideal points.

| Conclusion and Contribution

Conclusion

- This study introduces a FER network that efficiently captures crucial facial information while minimizing inference delay by leveraging Dilated Mix Convolutions and Gated Spatial Attention.
- The proposed model achieved the best performance among six comparison models, slightly surpassing state-of-the-art accuracy while reducing computational cost by 50% and inference time by about 30%.
- The study further demonstrated that our backbone can serve as an effective and generalizable feature extractor, making it a strong candidate for integration into diverse FER frameworks.

Achievement

- Projects
 - Lightweight Micro-Facial Expression Recognition for near-Neutral Faces Based on Self-Knowledge Distillation, *National Research Foundation of Korea*, 2025~2026
 - Development of Integrated Development Framework that supports Automatic Neural Network Generation and Deployment optimized for Runtime Environment, *IITP*, 2025
 - Teaching Assistant, *Hyundai Bootcampus*, 2025
 - The Development of Automatic Search Technology for Hyperparameter Optimization for Vision AI Model, *Hyundai*, 2025
- In Domestic Conferences
 - Enhancing Facial Expression Recognition via Fine-Tuning with Supervised Contrastive Learning, *IEIE*, 2025

Future Plan

Requirement Skill and Plan

석사 학위 취득 후, 컴퓨터비전 관련 기술을 연구하는 기업에 병역특례 전문연구요원으로 취업하는 것을 목표로하며 관련 회사에서 필요로 하는 요구 사항, 우대 사항, 프로그래밍 기술을 확보하고자 함.

[표 9] 회사 별 직무 요구 역량

직무 (회사)	요구 역량
딥러닝 연구원 - 전문연구요원 (딥브레인 AI)	휴먼AI 관련 연구 (얼굴 및 몸 영상 생성 및 합성, 음성 합성, 음성 변환, 딥페이크 탐지, 자연어처리 등) 개발 경험 또는 논문 경험 요구
컴퓨터비전 엔지니어 - 전문연구요원 (플라잇)	C/C++, Python 기술 요구, AI 비전 프로젝트 배포 경험 요구, 3D 딥러닝 모델 개발 경험 우대

[표 10] 해당 요구 역량 확보 및 보충 계획

요구 역량	확보 및 보충 계획
휴먼AI 관련 연구 개발 경험 또는 논문 경험	현재 석사 논문 연구 주제로 선정한 Facial Expression Recognition (FER) 분야의 연구 및 논문 경험을 통해 서 확보하고자 함.
Python, C/C++ 기술	Python을 사용하여 PCCP 자격증을 취득하고, C++을 사용하여 개인 프로젝트를 배포함으로써 해당 요구 역량을 보충하고자 함.
AI 비전 프로젝트 배포 경험	현재 참여 중인 연구실의 현대자동차 프로젝트로 해당 경험을 확보하고, 추가적으로 FER 관련 개인 프로젝 트를 웹/앱으로 배포함으로써 확보하고자 함.

Reference

- [1] Tao, Huanjie, and Qian Yue Duan. "Hierarchical attention network with progressive feature fusion for facial expression recognition." *Neural Networks* 170 (2024): 337-348.
- [2] Dhall, Abhinav, et al. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark." 2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011.
- [3] She, Jiahui, et al. "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
- [4] Ale, Laha, et al. "Lightweight deep learning model for facial expression recognition." 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2019.
- [5] Zhao, Zengqun, Qingshan Liu, and Feng Zhou. "Robust lightweight facial expression recognition network with label distribution training." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 4. 2021.
- [6] Xue, Fanglei, Qiangchang Wang, and Guodong Guo. "Transfer: Learning relation-aware facial expression representations with transformers." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
- [7] Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." *IEEE transactions on affective computing* 13.3 (2020): 1195-1215.
- [8] Zheng, Ce, Matias Mendieta, and Chen Chen. "Poster: A pyramid cross-fusion transformer network for facial expression recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
- [9] Wen, Zhengyao, et al. "Distract your attention: Multi-head cross attention network for facial expression recognition." *Biomimetics* 8.2 (2023): 199.
- [10] Zhang, Saining, et al. "A dual-direction attention mixed feature network for facial expression recognition." *Electronics* 12.17 (2023): 3595.
- [11] Liu, Shu, et al. "A Dual-Branch Adaptive Distribution Fusion Framework for Real-World Facial Expression Recognition." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.
- [12] Mao, Jiawei, et al. "Poster++: A simpler and stronger facial expression recognition network." *Pattern Recognition* (2024): 110951.
- [13] Xu, Jie, et al. "Multiscale facial expression recognition based on dynamic global and static local attention." *IEEE Transactions on Affective Computing* (2024).
- [14] Zhang, Xin, et al. "A gradual self distillation network with adaptive channel attention for facial expression recognition." *Applied Soft Computing* 161 (2024): 111762.
- [15] Zhang, Yuhang, et al. "Leave no stone unturned: mine extra knowledge for imbalanced facial expression recognition." *Advances in Neural Information Processing Systems* 36 (2024).
- [16] Lucey, Patrick, et al. "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression." 2010.
- [17] Jiang, Shaoping, et al. "CSE-GResNet: A Simple and Highly Efficient Network for Facial Expression Recognition." *IEEE Transactions on Affective Computing* (2025).
- [18] Zhang, Xin, et al. "A gradual self distillation network with adaptive channel attention for facial expression recognition." *Applied Soft Computing* 161 (2024): 111762.
- [19] Dhall, Abhinav, et al. "Static Facial Expression in the Wild (SFEW) Database for Facial Expression Recognition in the Wild." 2013.
- [20] Li, Shan, et al. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild." *Computer Vision and Pattern Recognition*, 2017.
- [21] Barsoum, Emad, et al. "Training Deep Neural Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution." *Proceedings of the ACM International Conference on Multimedia*, 2016.
- [22] Goodfellow, Ian, et al. "Challenges in Representation Learning: Facial Expression Recognition Challenge." *International Conference on Machine Learning (ICML) Workshop*, 2013.
- [23] Li, Shan, et al. "Expression Analysis Based on Face Regions in Real-World Conditions." *International Journal of Automation and Computing*, 2018.

Thank you