JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

A Case Study of EmoNeXt-Tiny Without Self-Attention for Facial Emotion Recognition

Ngoc Minh Tran[1], **Viet Tuan Le**[2*]
[1]*Nha Trang College of Technology, Vietnam*
[2]*Ho Chi Minh City Open University, Vietnam*

[*]Corresponding author. *Email: tuan.lv@ou.edu.vn*

ARTICLE INFO

Received: 04/03/2026
Revised: 13/03/2026
Accepted: 15/05/2026
Online First: 24/06/2026
Published:

KEYWORDS

Facial Emotion Recognition;
ConvNeXt-Tiny;
Spatial Transformer Network;
Squeeze-and-Excitation;
Self-Attention;
EmoNeXt-Tiny.

ABSTRACT

Facial emotion recognition (FER) on FER2013 remains challenging due to low-resolution grayscale images, class imbalance, and limited data diversity. This study presents a controlled Tiny-scale ablation of EmoNeXt-Tiny to isolate the contribution of the self-attention (SA) regularization term. Specifically, SA regularization is removed from the objective function, while all other components and conditions are preserved, including the ConvNeXt-Tiny backbone, the Spatial Transformer Network (STN), Squeeze-and-Excitation (SE) modules, the preprocessing pipeline, and the training/evaluation protocol. Experiments are conducted on FER2013 using the official train/validation/test split. Under an identical setup, the SA-free variant achieves 72.95% top-1 test accuracy, compared with 73.34% for the full EmoNeXt-Tiny and 71.99% for the ConvNeXt-Tiny baseline. Because the present study reports a single controlled run per configuration (i.e., without multi-seed repetition), the 0.39 percentage-point gap should be interpreted as a preliminary observation that may fall within random variance. Within this constraint, the findings indicate diminishing returns from SA regularization in this Tiny regime once geometric alignment and channel re-weighting are already incorporated. In addition, the SA-free model simplifies optimization by removing an auxiliary loss component and its associated tuning burden. Overall, an STN+SE-enhanced ConvNeXt-Tiny without SA offers a practical accuracy–complexity trade-off for resource-constrained FER applications.

Doi: https://doi.org/10.54644/jte.2026.2109

1. Introduction

Facial expressions constitute a key non-verbal communication channel, and facial emotion recognition (FER) has become an essential component in practical systems such as security, healthcare, driver-state monitoring, and human–machine interaction. Traditional FER pipelines based on hand-crafted descriptors and decoupled feature–classification stages tend to degrade under large data variability. In contrast, end-to-end deep learning approaches have shown stronger robustness to geometric distortions, illumination changes, and inter-subject diversity, with canonical backbones such as ResNet and Xception continuing to serve as reliable architectural baselines and reference points for modern FER systems [1], [2].

To enable controlled and comparable evaluation, FER2013 is widely adopted as a standard benchmark. It consists of low-resolution grayscale facial images with predefined training, validation, and test splits, and it exhibits notable class imbalance, which together make it a challenging yet convenient setting for fair comparison across methods [3]. Within this benchmark, recent work has increasingly emphasized the importance of backbone design for balancing recognition performance with efficiency. ConvNeXt exemplifies the trend of "CNNs for the 2020s" by modernizing convolutional architectures through design choices such as depthwise convolutions, inverted bottlenecks, GELU activations, and LayerNorm, and its Tiny configuration is particularly attractive when computational budgets are constrained while competitive representational capacity must be maintained [4], [5].

Beyond the backbone itself, robustness in FER is often improved by incorporating architectural mechanisms that explicitly address facial variability and noisy inputs. Spatial Transformer Networks provide differentiable geometric transformations to support alignment under pose and cropping variations, while Squeeze-and-Excitation modules recalibrate channel responses to emphasize informative cues and suppress noise, which is especially relevant under low-resolution grayscale conditions such as FER2013 [6], [7]. Complementary directions have also been explored, including localized channel attention mechanisms and FER-specific loss formulations for in-the-wild settings, as well as segmentation-informed pipelines that emphasize expression-relevant facial regions under limited-data regimes [8]–[11]. Building on ConvNeXt, EmoNeXt integrates STN at the input and SE after each stage, and further introduces a self-attention regularization term as an auxiliary component to cross-entropy with the aim of stabilizing attention statistics and enriching representations [12].

Motivated by the practical constraints associated with Tiny-scale deployment and experimentation, this paper presents a focused Tiny-scale case study designed to isolate the contribution of the self-attention regularization term under a controlled setting. Specifically, the ConvNeXt-Tiny backbone and the STN and SE components are retained, and the training and evaluation protocol is kept unchanged, while only the self-attention term is removed from the objective. The study objective is to determine whether self-attention regularization remains necessary in a Tiny regime when alignment and channel re-weighting are already present. Accordingly, a reproducible comparison is reported among three Tiny configurations under the same FER2013 protocol, including the ConvNeXt-Tiny baseline [4], the full EmoNeXt-Tiny model [12], and the proposed No-SA variant.

The contribution of this work is empirical rather than architectural. Specifically, the paper presents a controlled engineering case study that isolates the role of the SA loss within a fixed Tiny-scale FER2013 pipeline.

2. Related Work

Classical facial emotion recognition methods have traditionally relied on hand-crafted geometric or appearance descriptors, including LBP and HOG, followed by shallow classifiers such as SVM, KNN, or MLP in a decoupled feature–classification pipeline. Although these approaches are computationally lightweight, they are often sensitive to geometric variation, illumination changes, and inter-subject diversity. With the emergence of deep learning, hierarchical representations have increasingly been learned in an end-to-end manner, and consistent gains have been reported for FER and related domains such as healthcare, security, and human–machine interaction. In this context, modern backbones, including ResNet and Xception, have commonly been adopted as strong baselines for both performance and training stability [1], [2].

Most contemporary FER studies are benchmarked on widely used datasets to enable controlled comparisons. FER2013 is frequently selected for this purpose because it provides predefined training, validation, and test splits, while also presenting practical challenges stemming from low-resolution grayscale inputs and class imbalance [3]. As a result, it has become a standard experimental setting for evaluating architectural modifications and training strategies under comparable protocols.

In addition to deep CNN backbones, efficient real-time FER pipelines have been explored using facial landmarks and lightweight camera-based processing, which can be suitable for deployment with limited compute resources [13]. More recently, dataset-centric efforts have expanded beyond visible imagery; for example, KTFEv2 provides a multimodal facial emotion database and analysis using visible and thermal infrared modalities, highlighting the role of modality and intensity variation in emotion recognition [14]. Such works complement FER2013-based benchmarking by emphasizing practical deployment constraints and cross-dataset variability.

Recent progress has also highlighted that backbone design remains a primary determinant of accuracy–efficiency trade-offs. ConvNeXt has been positioned as part of the broader trend of “CNNs for the 2020s,” incorporating depthwise convolutions, inverted bottlenecks, GELU activations, and LayerNorm to modernize convolutional networks and narrow the gap with Transformer-style performance in vision [4], [5]. The Tiny configuration, in particular, is attractive when computational

budgets are constrained, as competitive representational capacity can be retained with relatively low cost. These properties make ConvNeXt-Tiny a practical foundation for small-scale FER.

Beyond the backbone, robustness in FER has often been pursued through architectural components that address nuisance factors inherent to facial imagery. Spatial Transformer Networks introduce differentiable geometric transformations, thereby supporting alignment under pose changes or cropping deviations [6]. Squeeze-and-Excitation modules re-weight channel responses through a squeeze–excitation mechanism, amplifying informative signals while suppressing noise, which can be especially valuable for low-resolution grayscale inputs such as those in FER2013 [7]. In parallel, additional enhancement directions have been explored to emphasize expression-relevant regions and improve discrimination in unconstrained settings. Representative examples include segmentation-assisted designs that localize critical facial areas through U-Net-style branches [10], zoning or segmentation modules integrated with VGG- or ResNet-based pipelines [11], attention-oriented mechanisms such as local multi-head channel self-attention, and FER-specific objectives such as adaptive correlation-based losses [8], [9]. Efficient architectural families such as Xception, based on depthwise separable convolutions, have also continued to serve as lightweight foundations for recognition tasks [1].

Building directly on ConvNeXt, EmoNeXt integrates an STN at the input stage and inserts SE modules after each stage, while further introducing a self-attention regularization term that supplements cross-entropy and is intended to stabilize representations by regulating attention statistics [12]. Under standard FER conditions in the Tiny regime, the full EmoNeXt variant has been reported to improve upon the ConvNeXt baseline [12]. However, the isolated contribution of the self-attention regularization term remains less clear when alignment and channel re-weighting are already present in the same Tiny-scale configuration.

To address this gap, we retain the ConvNeXt-Tiny backbone together with STN and SE, while removing only the SA regularization term, following the FER2013 official split and evaluation protocol [3], [12]. The architectural components are based on ConvNeXt [4], [5], Spatial Transformer Networks [6], and Squeeze-and-Excitation blocks [7]. The analysis is intended to quantify whether the additional loss component is necessary once geometric alignment and channel re-balancing have been incorporated, and to provide a compact reference point for resource-constrained deployments. This positioning is established relative to common baselines such as ResNet and Xception, and to recent enhancement directions that emphasize region awareness, attention mechanisms, and FER-specific loss design [1], [2], [8]–[11].

3. Models and Methods

Our objective in this setting is to classify seven emotions – Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral – from 48×48 grayscale face images in FER2013 (Figure 1). We restrict all experiments to the Tiny configuration to reflect a "narrow" variant and common computational constraints [3].

Figure 1. *Sample images from the FER2013 benchmark dataset [3].*

3.1. Loss function

Given an input image x, ground-truth label y∈{1,…,7}, and model posterior $p_\theta(y \mid x)$, we employ the standard cross-entropy loss without any Self-Attention (SA) term:

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta\,(y_i|x_i) \tag{1}$$

No additional regularizer related to SA is used.

3.2. Overall architecture

The model retains two architectural pillars known to be effective for FER – Spatial Transformer Network (STN) at the input and Squeeze-and-Excitation (SE) after each backbone stage – while removing only the SA term from the loss (relative to the full EmoNeXt) [4], [6], [7], [12]. The computational flow is:

$$\mathbf{x} \xrightarrow{\text{STN}} \tilde{\mathbf{x}} \xrightarrow{\text{Backbone (ConvNeXt-Tiny + SE)}} \mathbf{f} \xrightarrow{\text{Head}} \hat{\mathbf{z}} \xrightarrow{\text{Softmax}} \hat{\mathbf{p}}$$

Here $\tilde{\mathbf{x}}$ denotes the STN-aligned input, $\mathbf{f}$ the extracted features, $\hat{\mathbf{z}}$ the output logits and $\hat{\mathbf{p}}$ the predicted class distribution. The overall architecture of the proposed No-SA variant is illustrated in Figure 2.

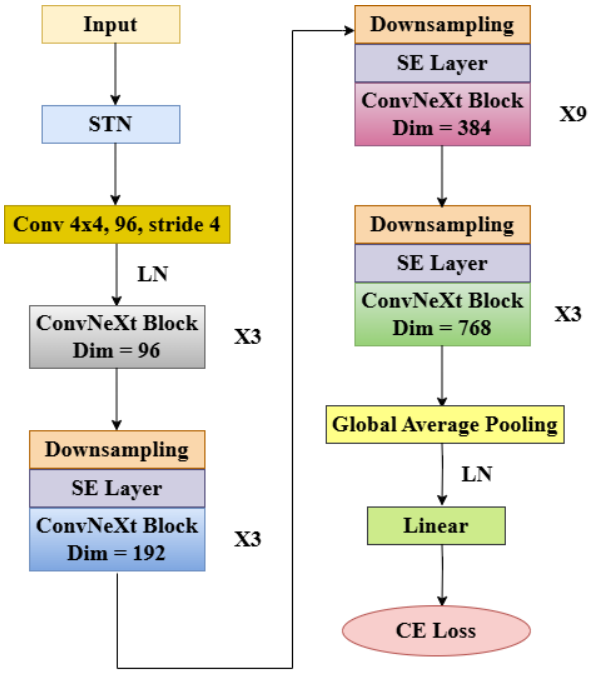

***Figure 2**. Architecture design for EmoNeXt-Tiny (No-SA).*

3.3. Spatial Transformer Network (STN)

STN is placed before the backbone to learn differentiable geometric transformations (translation, rotation, scaling) for facial alignment under pose/cropping variations [6]. A shallow localization network predicts a 2×3 affine matrix θ; a grid generator produces sampling coordinates from θ; and a sampler performs bilinear interpolation to obtain $\tilde{\mathrm{x}} = \mathrm{STN}(\mathrm{x}; \theta)$ [6]. STN is optimized end-to-end solely with $\mathcal{L}_{\mathrm{CE}}$ in (1) (no auxiliary constraints), allowing it to learn transformations beneficial for emotion classification.

3.4. Backbone: ConvNeXt-Tiny with SE

The backbone comprises four stages using the modernized ConvNeXt design: depthwise convolution, pointwise/MLP (inverted bottleneck), and GELU + LayerNorm, providing an efficient Tiny-scale CNN for the "2020s" [4], [5]. After every stage, an SE block re-balances channels: Squeeze via global average pooling $\mathbf{s} \in \mathbb{R}^C$; Excitation via two fully connected layers (channel reduction then expansion, reduction ratio r typically 16) to produce weights $\alpha \in (0,1)^C$; and Scale by channel-wise multiplication $\boldsymbol{\alpha} \odot$ (feature tensor) [7].

Classification head. Following the backbone + SE, features undergo global average pooling, optional LayerNorm, and a linear classifier to seven output classes.

3.5. Preprocessing and data input

Because FER2013 images are grayscale, inputs are replicated to three channels to match ConvNeXt pretrained weights [3], [4]. Images are then resized and cropped to the backbone's standard resolution (e.g., 224×224), and normalized with ImageNet statistics to align the input distribution with the pretrained model. The train/validation/test protocol and preprocessing/evaluation steps follow common practice and the original setup; we introduce no new augmentation so as to ensure a fair comparison focused solely on removing SA.

3.6. Training and inference

Training uses only the cross-entropy loss in (1), with no SA term or attention-related regularization [12]. Inference/evaluation follows the Tiny protocol (e.g., center-crop for validation/test), keeping all factors other than SA fixed to isolate the effect of SA on overall performance [3], [4], [12].

4. Results and Discussion

4.1. Results

Seven-class facial emotion recognition was evaluated on FER2013, a benchmark consisting of 48×48 grayscale facial images. The official data split was used, including 28,709 training images, 3,589 validation images, and 3,589 test images, as summarized in Table 1. All experiments were restricted to the Tiny configuration in order to reflect the narrow-scope setting of this study.

Table 1. *Summary of the FER2013 Dataset.*

Class	Training	Validation	Testing	Class Total
angry	3,995	467	491	4,953
disgust	436	56	55	547
fear	4,097	496	528	5,121
happy	7,215	895	879	8,989
sad	4,830	653	594	6,077
surprise	3,171	415	416	4,002
neutral	4,965	607	626	6,198
Total	28,709	3,589	3,589	35,887

Three configurations were assessed under the same Tiny backbone setting and an identical training/evaluation protocol:

- ConvNeXt-Tiny (baseline).
- EmoNeXt-Tiny (full components as in the original work).
- EmoNeXt-Tiny (No-SA): identical to EmoNeXt-Tiny except that the Self-Attention term is removed.

No new techniques were introduced in this evaluation. The experimental setup follows common Tiny-scale practice:

- Grayscale images are replicated to three channels to use pretrained weights.
- Input resolution 224×224; preprocessing and evaluation follow the standard Tiny pipeline.
- ImageNet-22k pretrained, batch size 64, 300 epochs, AMP enabled.
- The best checkpoint is selected by validation accuracy, then test accuracy is reported.
- Hardware: Google Colab GPU (L4/T4/A100), num_workers = 1.

Top-1 accuracy (%) on the FER2013 test split is reported in Table 2. In this controlled single-run setting, the No-SA configuration remains competitive with several reported baselines. However, the baseline values in Table 2 are reproduced from EmoNeXt [12] under the same official split and should therefore be interpreted accordingly. The No-SA result is obtained in this study.

***Table 2**. Accuracy comparison on FER2013. Baseline accuracies are reproduced from EmoNeXt [12] under the FER2013 official split; primary references of the baseline architectures are also cited (e.g., GoogLeNet [15], Deep-Emotion [16], Inception [17]). The No-SA result is obtained in this study.*

Model	Test Top-1 Acc.
GoogLeNet [15] (reproduced from [12])	65.20%
Deep-Emotion [16] (reproduced from [12])	70.02%
Inception [17] (reproduced from [12])	71.60%
ConvNeXt-Tiny [12]	**71.99%**
Ad-Corre [12]	72.03%
ConvNeXt-Small [12]	72.34%
SE-Net50 [12]	72.50%
EmoNeXt-Tiny (No-SA) (this study)	**72.95%**
EmoNeXt-Tiny (full) [12]	**73.34%**

To provide a more informative view beyond overall accuracy, the class-wise performance of the No-SA model on the FER2013 test split is presented in Figure 3. The figure highlights which emotion categories are recognized more reliably and which remain challenging.

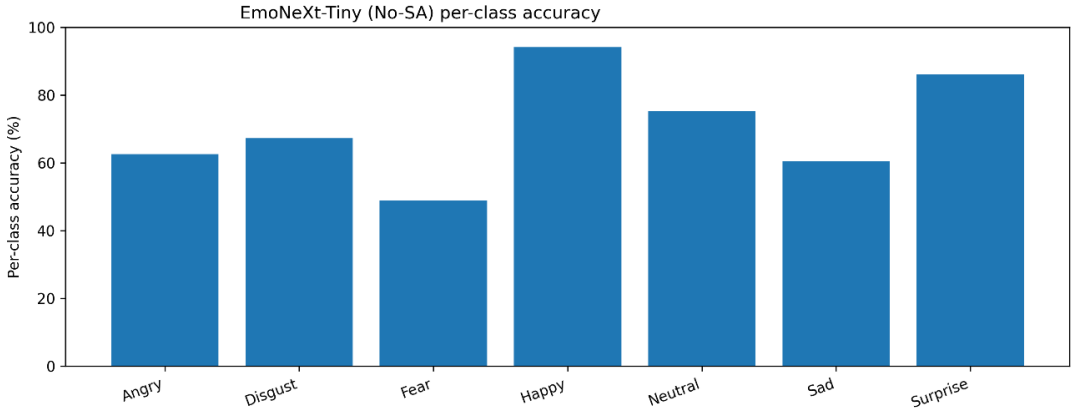

***Figure 3**. Per-class accuracy (recall) of EmoNeXt-Tiny (No-SA) on the FER2013 test split.*

For a more detailed class-level interpretation, Table 3 reports the per-class accuracy, or recall, of the No-SA model on the FER2013 test set.

***Table 3**. Per-class accuracy (recall) of EmoNeXt-Tiny (No-SA) on FER2013 test split.*

Class	Emotion	Test	Correct	Per-class Acc.	Incorrect
0	Angry	491	307	62.53%	184
1	Disgust	55	37	67.27%	18
2	Fear	528	258	48.86%	270
3	Happy	879	828	94.20%	51
4	Neutral	626	471	75.24%	155
5	Sad	594	359	60.44%	235
6	Surprise	416	358	86.06%	58

4.2. Discussion

Relative to ConvNeXt-Tiny, the No-SA variant improves test accuracy by +0.96 percentage points (72.95% vs. 71.99%), indicating that the combination of STN and SE already provides a meaningful benefit even without the SA loss. Compared with the full EmoNeXt-Tiny model, removing SA leads to only a 0.39-percentage-point reduction (73.34% vs. 72.95%). Under the constraints of FER2013, including 48×48 grayscale inputs, limited data diversity, and class imbalance, this observed gap is small, suggesting that SA may contribute only limited marginal benefit once geometric alignment (STN) and channel re-weighting (SE) are already in place and the rest of the pipeline is held constant. At the same time, because the present comparison is based on a fixed-run setting rather than repeated multi-seed trials, this small difference should be interpreted with caution and should not be treated as a statistically definitive estimate of the true effect of SA. Accordingly, we refrain from making conclusive claims and present the gap as a run-specific empirical observation.

The No-SA variant also remains competitive with several reported baselines. It surpasses classical CNN models such as GoogleNet (65.20%) and Inception (71.60%), exceeds Ad-Corre (72.03%) and SE-Net50 (72.50%), and slightly outperforms ConvNeXt-Small (72.34%) despite operating at Tiny capacity. This pattern suggests that careful integration of STN and SE, even without SA, can recover most of the performance associated with stronger Tiny-scale models and some non-Tiny alternatives. From an engineering perspective, removing SA simplifies the training objective by reducing it to pure cross-entropy, eliminates the need to tune an additional SA-related loss weight, and avoids the implementation overhead associated with that auxiliary component. These practical advantages make the No-SA variant attractive for resource-constrained experimentation and deployment scenarios, including Colab-class GPU environments, where simpler optimization often translates into easier reproduction and faster iteration.

In practical terms, No-SA offers near-full EmoNeXt-Tiny performance without the additional tuning burden introduced by the SA term. This can shorten experimental turnaround and reduce optimization complexity by lowering the number of interacting hyperparameters. Even so, differences below one percentage point should not be over-interpreted as universally decisive. FER2013 is a relatively small grayscale benchmark, and SA may play a more important role in higher-resolution, more diverse datasets or under larger-capacity model settings. Accordingly, the present findings are best understood as evidence of diminishing returns for SA within the specific Tiny-scale FER2013 regime examined here, rather than as a general claim that SA is unnecessary in FER.

The per-class results in Table 3 and Figure 3 provide additional insight into this behavior. No-SA performs strongly on Happy (94.20%) and Surprise (86.06%), which are often associated with more visually salient and distinctive facial cues, whereas performance is lower on Fear (48.86%) and remains moderate on Angry (62.53%) and Sad (60.44%). This pattern suggests that, in the Tiny-scale low-resolution grayscale setting, the remaining errors are concentrated in emotion categories with subtler, more ambiguous, or more easily confusable appearance variations. Disgust (67.27%) should be interpreted with caution because its test support is very small (55 images), meaning that only a few misclassifications can substantially change the reported percentage. These class-wise observations complement the overall accuracy comparison by clarifying which categories are more reliably recognized and which contribute most to the residual performance gap.

At the same time, several limitations should be acknowledged. This study is confined to the FER2013 benchmark, which consists of low-resolution grayscale images and exhibits class imbalance; therefore, the reported gains and class-wise trends may not transfer directly to higher-resolution or more diverse FER datasets. In addition, the conclusions are specific to the Tiny regime and to the fixed training and evaluation protocol used to isolate the effect of removing the SA regularization term. Different model capacities, augmentation strategies, loss formulations, or optimization settings may alter the relative contribution of SA. Moreover, because the current study reports one run per configuration under limited compute conditions, the observed 0.39-point gap between the No-SA and full variants remains a run-specific observation that should be validated across multiple random seeds in future work. Broader validation on additional FER datasets, along with cross-dataset generalization analysis, therefore remains an important direction for future work.

5. Conclusions

This study presents a controlled comparison in the Tiny-scale FER2013 setting to isolate the role of the Self-Attention loss while keeping the overall architecture and training procedure unchanged. Under the present fixed-run evaluation setting, the results provide a preliminary indication that, once STN and SE are incorporated, adding SA regularization may yield diminishing returns in this Tiny-scale FER2013 configuration. The No-SA variant remains close to the full model in accuracy, stays competitive with common baselines, and is simpler to optimize and maintain because it removes an auxiliary loss term and its associated tuning coefficient.

From a practical perspective, these findings are most relevant to scenarios that value simplicity, reproducibility, and low computational overhead, including teaching environments, lightweight prototypes, and resource-constrained deployment. In such settings, retaining STN and SE while omitting Self-Attention may serve as a reasonable default choice, as it preserves most of the observed performance while reducing training friction, implementation complexity, and hyperparameter sensitivity.

At the same time, these conclusions should be interpreted within the scope of the present study. The analysis is limited to FER2013, a low-resolution grayscale benchmark with class imbalance, and the current results are reported under a fixed-run setting rather than repeated multi-seed evaluation. Future work should therefore validate the present observations on additional FER datasets with higher resolution and more diverse capture conditions, examine cross-dataset generalization and robustness to pose, occlusion, and illumination, and explore capacity scaling from Tiny to larger backbones under matched training recipes to determine more clearly the regimes in which Self-Attention provides measurable benefit. Openly released code and evaluation protocols can further support reproducibility and practical adoption in FER scenarios that require a compact and dependable baseline.

Conflict of Interest

The authors declare no conflict of interest

REFERENCES

[1] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1251–1258.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[3] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2013, pp. 117–124. (FER2013).
[4] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11976–11986.
[5] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 16133–16143.
[6] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015.
[7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
[8] R. Pecoraro, V. Basile, and V. Bono, "Local Multi-Head Channel Self-Attention for Facial Expression Recognition," *Information*, vol. 13, no. 9, p. 419, 2022.
[9] A. P. Fard and M. H. Mahoor, "Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild," *IEEE Access*, vol. 10, pp. 26756–26768, 2022.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.
[11] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A Novel Facial Emotion Recognition Model Using Segmentation VGG-19 Architecture," *Int. J. Inf. Technol.*, 2023, pp. 1–11.
[12] Y. El Boudouri and A. Bohi, "EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition," *arXiv preprint* arXiv:2501.08199, Jan. 14, 2025.
[13] B. Nguyen et al., "An Efficient Real-Time Emotion Detection Using Camera and Facial Landmarks," in *Proc. 7th Int. Conf. Inf. Sci. Technol. (ICIST)*, 2017, pp. 251–255.
[14] H. Nguyen, N. Tran, H. D. Nguyen, L. Nguyen, and K. Kotani, "KTFEv2: Multimodal Facial Emotion Database and Its Analysis," *IEEE Access*, vol. 11, pp. 17811–17822, 2023, doi: 10.1109/ACCESS.2023.3246047.
[15] C. Szegedy, W. Liu, and Y. Jia, "Going Deeper with Convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[16] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, vol. 21, no. 9, Art. no. 3046, 2021.
[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

Ngoc Minh Tran works in the Information Technology Department, Faculty of Electrical - Electronics at Nha Trang College of Technology since 2009 as a lecturer. He earned a Master's degree in Information Technology from Hanoi University of Science and Technology in 2012, specialized in teaching software development, web, and mobile applications, has been a member, consultant, editor, and author for technology communities such as CodeProject and DZone since 2016. In free time, he writes articles sharing professional experiences and insights on various fields on the personal blog ngocminhtran.com. Currently, he focuses on researching the application of Deep Learning to enhance vocational education activities under the supervision of Viet-Tuan Le, co-author of this paper.

Email: tnminh.cdktcn@khanhhoa.edu.vn. ORCID: https://orcid.org/0009-0006-6220-5134

Viet Tuan Le received the Ph.D. degree in computer science and engineering from the Sejong University, Korea, in 2024. He is currently an assistant professor within the Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam. His research interests include diverse network architectures for video anomaly detection and generative models.

Email: tuan.lv@ou.edu.vn. ORCID: https://orcid.org/0000-0002-2289-8128