Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33959

A Comparative Study of Efficient Neural Architectures for Facial Expression Recognition

Zitong Yang[1], Kexing Lian[2,*], Peichun Liao[3], Baoqi Gu[4]

[1]*School of Automation, North China Electric Power University, BaodingChina*
[2]*College of Science, Mathematics and Technology, Wenzhou-Kean University, WenZhou, China*
[3]*Glasgaw Colleague, University of Glasgaw, Glasgaw, UK*
[4]*AS-Computer science, Shenghua Zizhu Academy, Shanghai, China*
Corresponding author Email: liank@kean.edu

Abstract. We study facial expression recognition (FER) under the twin constraints of low-resolution inputs and on-device inference. We benchmark compact convolutional baselines—ResNet-50/152—and an attention-enhanced ResNet-50 with squeeze-and-excitation modules and class re-weighting. We also adopt a lightweight transformer, Iwin-T, that interleaves windowed self-attention with depthwise separable convolution to balance global context modeling and local inductive bias under limited computational resources. Using class-weighted cross-entropy on FER2013 with an 8:1:1 train–validation–test split, Iwin-T attains 68.09% Top-1 accuracy, surpassing ResNet-50 (63.22%), ResNet-152 (66.60%), and the attention-augmented ResNet-50 (67.19%). Beyond raw accuracy, we analyze training dynamics including loss/accuracy oscillations and training–validation divergence, and identify scheduling choices that improve stability under tight memory and compute budgets. Our findings suggest a practical guideline for edge-oriented FER: (i) a well-tuned ResNet-50 remains a robust default choice for stable deployment; (ii) channel attention and class re-weighting provide simple yet effective accuracy gains; (iii) when optimization can be carefully stabilized, Iwin-T delivers the best accuracy–efficiency trade-off for low-resolution FER tasks.

Keywords: Facial Expression Recognition, Resource-Constrained Devices, Lightweight Transformer, Iwin Transformer, FER2013.

1. Introduction

1.1. Topic

Problem. Facial expression recognition in the wild often processes small, compressed, or blurred facial images while satisfying strict latency and memory constraints on edge hardware. Larger models improve accuracy but violate efficiency constraints, while overly compressed networks lose sensitivity to subtle muscular changes.

Approach. We focus on low-resolution FER using compact residual architectures including ResNet-50 and ResNet-152 [1]. We enhance ResNet-50 with squeeze-and-excitation attention [2] and class-weighted cross-entropy to alleviate label imbalance. We also adopt Iwin-T [3], a lightweight transformer that uses interleaved window attention [4] and depthwise separable convolutions to retain

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33959

convolutional inductive bias. All experiments use FER2013 with an 8:1:1 train–val–test split [5]. Previous work on ResNet-based FER provides empirical context for our comparisons [6, 7]. We also tested YOLOv7 adapted for classification but abandoned it due to unstable batch behavior, as detailed in the methodology section.

Contributions.

- A controlled comparison of residual, attention-augmented CNN, and lightweight transformer backbones for low-resolution, edge-ready FER.
- Validation of Iwin-T as a strong efficiency–accuracy backbone for FER via interleaved window attention and depthwise convolutions.
- An analysis of training dynamics, overfitting patterns, and stability tradeoffs across models on FER2013.

1.2. Background

Challenges. Low-resolution inputs weaken fine-grained features, and edge deployment further restricts FLOPs and memory, exacerbating the accuracy–efficiency tradeoff and increasing overfitting risk on small datasets.

Protocol and baselines. We compare three families of models: (i) standard ResNet-50/152 [1], (ii) attention-augmented ResNet-50 with SE modules [2] and weighted loss, and (iii) Iwin-T, all evaluated on FER2013 [5]. This setup isolates the effects of model capacity, attention, and architectural structure on generalization and overfitting. Relevant prior work on ResNet FER systems is also referenced [7].

Applications. Low-latency on-device FER enables practical use cases in healthcare (pain and affect monitoring), education (student engagement detection), and security (behavioral analysis under degraded visual conditions).

2. Methodology

2.1. Base model architecture and initial setup

We began by testing YOLOv7 [8] repurposed for image classification by using full-image bounding boxes. However, this led to unstable training and incoherent sample behavior under large batches, so we switched to standard classification architectures.

We selected ResNet-50 as our primary baseline, a well-established residual network known for stable feature learning [1]. It provided a reliable reference compared to YOLOv7 and MobileNetV3 [9].

2.2. Model refinements: attention and loss function

We introduced two improvements to the baseline ResNet-50:

1. **Squeeze-and-Excitation (SE) Blocks.** SE modules enable adaptive channel recalibration by squeezing global spatial information via average pooling, followed by two fully connected layers that learn channel-wise weights. This helps the network focus on semantically important regions such as eyes and mouth, which is critical for low-resolution FER.
2. **Weighted Cross-Entropy Loss.** FER datasets typically suffer from class imbalance. We weighted classes inversely proportional to their frequency to reduce bias toward majority classes and improve generalization on underrepresented expressions.

Combined, these changes consistently improved performance over the vanilla ResNet-50, as shown in the results.

DOI: 10.54254/2755-2721/2026.GU33959

2.3. Iwin transformer

Drawing from related work on window-based transformers [10,11], we adopt the Iwin Transformer [3] as our lightweight transformer backbone. It is especially suitable when training data is limited compared to standard vision transformers [12].

Its core design fuses Interleaved Window Attention (IWA) and Depthwise Separable Convolution (DWConv) [13] in a unified module.

IWA uses a reshape–transpose–reshape (RTR) operation to reorganize tokens so each window contains spatially dispersed tokens, enabling global communication within one block. This is more efficient than the shifted window scheme used in Swin Transformer [4].

The parallel DWConv branch captures local patterns and provides convolutional inductive bias while naturally encoding positional information, removing the need for explicit position embeddings. This improves resolution flexibility when fine-tuning from high-resolution pretraining to low-resolution FER inputs.

Theoretical guarantees in [3] ensure global information flow under mild conditions. Overall, Iwin-T efficiently combines global context, linear complexity, and local bias, making it highly suitable for low-resolution, resource-constrained FER.

3. Results

We trained all models on FER2013 with an 8:1:1 train–val–test split. The class distribution of the dataset is shown in Figure 1, which illustrates the inherent imbalance that motivates our use of weighted loss.

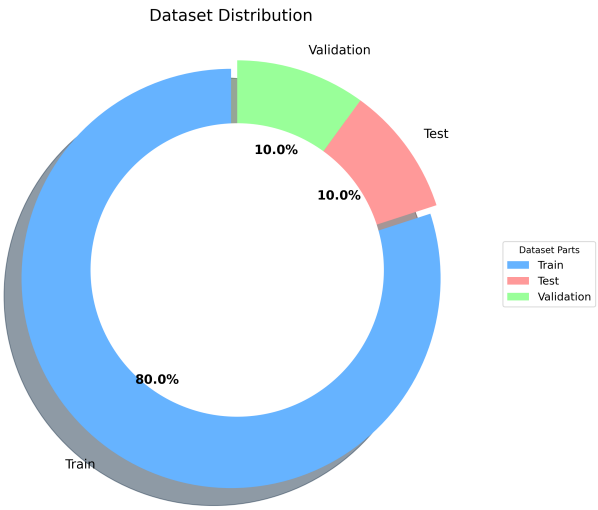


Figure1. Dataset class distribution

3.1. ResNet-50

As shown in Figure 3, ResNet-50 performs stably on FER2013, achieving a peak Top-1 accuracy of 63.22% at epoch 384 and Top-5 accuracy of 98.58%. While effective for general feature learning, it struggles with highly similar expressions. We also observed low-quality and mislabeled samples in the dataset, as shown in Figure 2, which limit achievable accuracy.

DOI: 10.54254/2755-2721/2026.GU33959

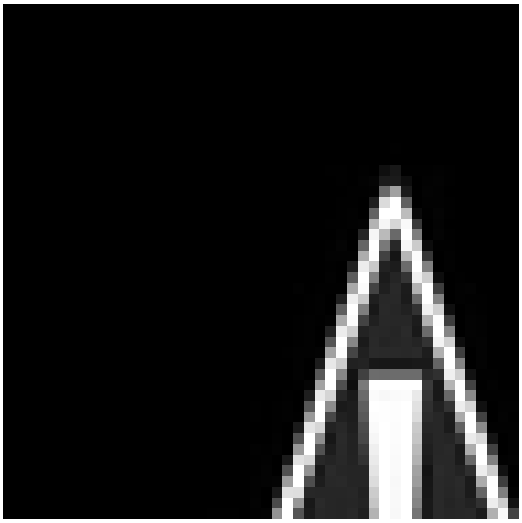


Figure2. Example of a low-quality sample in FER2013

Training converged smoothly with training loss consistently below validation loss, indicating stable generalization. Final accuracy was 62.98% Top-1 and 98.52% Top-5, very close to peak values, showing minimal overfitting.

The learning rate decayed from 1×10^{-5} to 4.63×10^{-8} with an adjustment at epoch 200. A strong negative correlation between loss and accuracy (Pearson $r = -0.992$, $p < 0.001$) confirms healthy optimization. The complete training curves are shown in Figure 3.

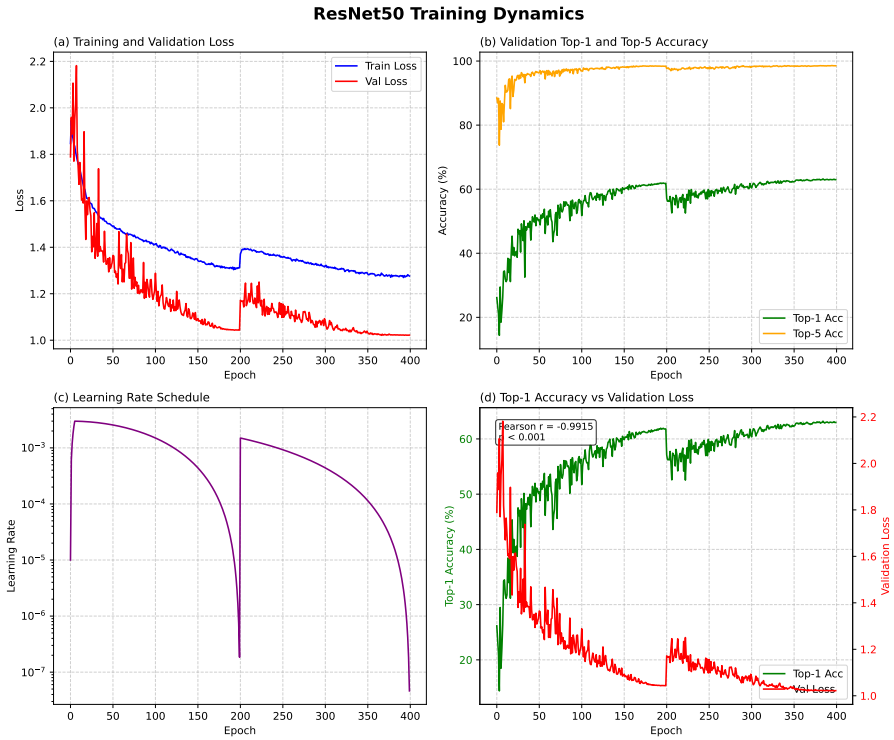


Figure3. ResNet-50 training dynamics

DOI: 10.54254/2755-2721/2026.GU33959

Table1. Best performance metrics of ResNet50 and the epochs at which they were achieved

Metric	Metric	Best Value	Epoch of best
Top1 Accuracy	62.98%	63.22%	384
Top5 Accuracy	98.52%	98.58%	339
Validation Loss	1.0220	1.0209	366
Training Loss	1.2763	1.2695	392

3.2. ResNet-50 + Attention + Weighted Cross-Entropy

We refer to the enhanced model as Model II and the baseline as Model I. Model II achieved peak Top-1 accuracy 67.19% (epoch 365) and Top-5 98.89%, clearly outperforming the baseline. The complete training dynamics are presented in Figure 4.

However, after reaching minimum validation loss (0.9807) at epoch 110, validation loss gradually increased to 1.1123 while training loss continued falling, indicating overfitting. Final accuracy was 66.73% Top-1 and 96.07% Top-5.

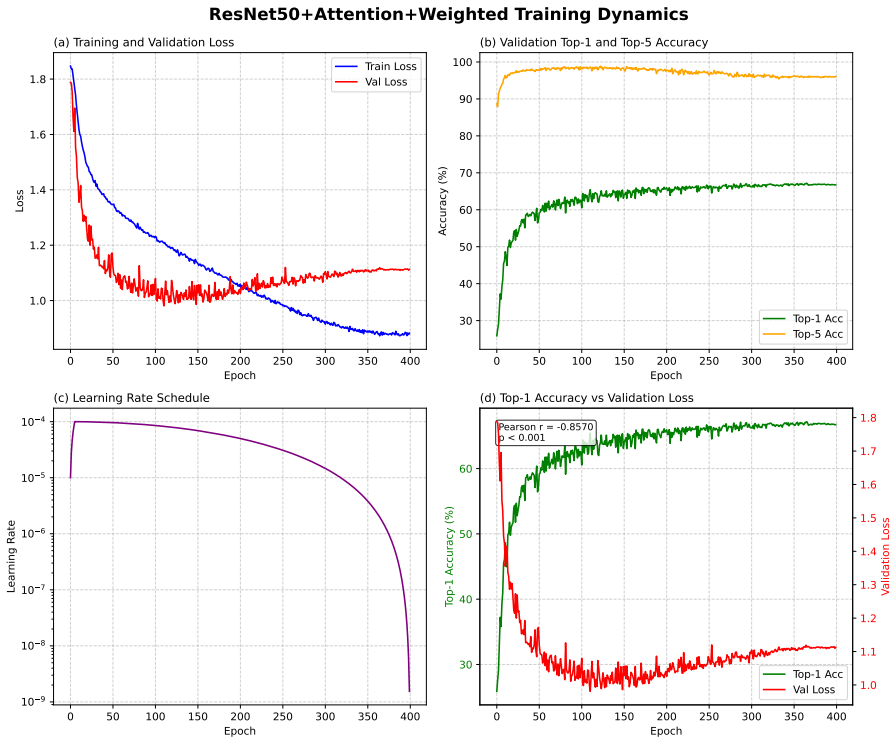


Figure4. ResNet-50+Attention+Weighted Cross-Entropy training dynamics

During epochs 110–250, accuracy remained stable despite rising loss, showing that loss and accuracy are not perfectly aligned. The correlation weakened to $r = -0.776$, reflecting less stable dynamics. The model benefits from attention and reweighting but requires stronger regularization to control overfitting.

DOI: 10.54254/2755-2721/2026.GU33959

Table2. Best performance metrics of ResNet50 + Attention + Weighted Cross-Entropy and the epochs at which they were achieved

Metric	Metric	Best Value	Epoch of best
Top1 Accuracy	66.73%	67.19%	365
Top5 Accuracy	96.07%	98.89%	122
Validation Loss	1.1123	0.9807	110
Training Loss	0.8823	0.8726	364

3.3. ResNet-152

ResNet-152 achieved peak Top-1 accuracy 66.60% (epoch 287) and Top-5 98.66%, comparable to the enhanced ResNet-50 but significantly better than vanilla ResNet-50. Its training behavior is shown in Figure 5.

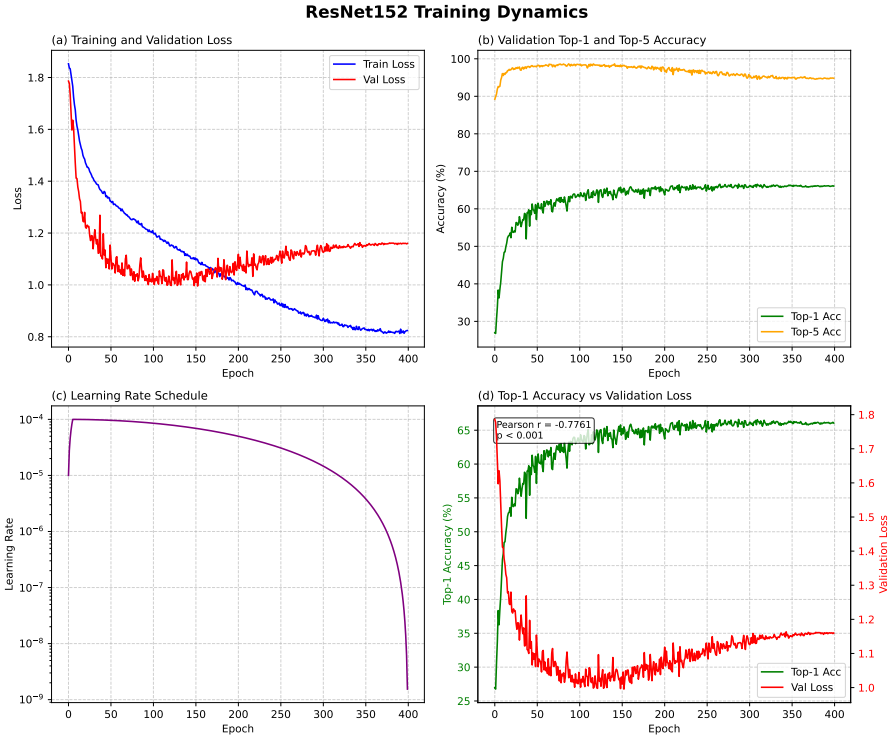


Figure5. ResNet-152 training dynamics

However, it suffered from severe overfitting: validation loss bottomed out early and rose substantially, widening the gap from training loss. Final Top-5 accuracy dropped to 94.80%, a large decline from peak.

ResNet-152's greater depth increases capacity but also raises overfitting risk on FER2013 without strong regularization. In contrast, ResNet-50 provides the best balance of stability and efficiency.

DOI: 10.54254/2755-2721/2026.GU33959

Table3. Best performance metrics of ResNet152 and the epochs at which they were achieved

Metric	Metric	Best Value	Epoch of best
Top1 Accuracy	66.06%	66.60%	287
Top5 Accuracy	94.80%	98.66%	141
Validation Loss	1.1599	0.9955	152
Training Loss	0.8240	0.8123	390

3.4. Iwin transformer

We evaluated four Iwin variants: Base, Small22t01, Base22t01, and Tiny. All showed similar training trends as summarized in Figure 6. Base and Tiny achieved the highest peak Top-1 accuracy at 68.09%, while other variants remained below 60%.

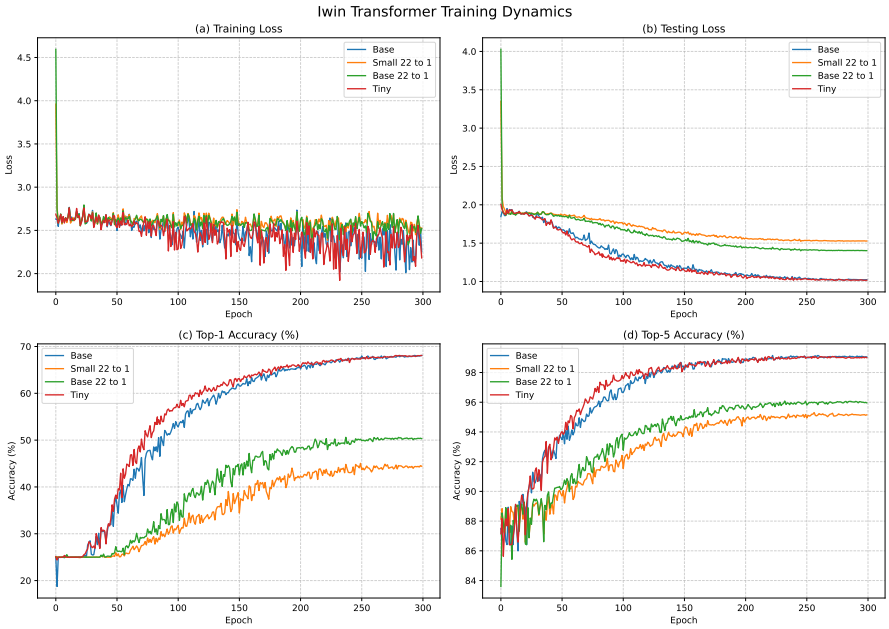


Figure6. Iwin Transformer training dynamics

All variants showed a large train–val accuracy gap (around 70% vs 50%), suggesting overfitting, though loss curves did not show classic rebound behavior. Significant fluctuations implied unstable optimization, likely due to sensitive learning rate scheduling.

We hypothesize that large gradient steps cause the model to enter sharp local minima specific to training batches [14], leading to unstable generalization. Nonetheless, with proper tuning, gradient-based methods can still avoid spurious minima and converge to strong solutions [15], highlighting the need for architecture-aware optimization.

4. Discussion

Our experiments reveal meaningful tradeoffs among capacity, stability, attention, and transformer-style design for low-resolution edge FER.

DOI: 10.54254/2755-2721/2026.GU33959

Capacity, Overfitting, and Regularization. ResNet-50 converged stably with minimal overfitting. Deeper (ResNet-152) and attention-augmented models achieved higher peak accuracy but overfitted more severely. Controlling capacity and strengthening regularization is essential to unlock their full potential on FER2013.

Architectural Bias for FER. SE attention improved focus on facial regions critical for expression recognition. Iwin-T further improved accuracy by combining global window attention and local convolutional bias, outperforming both baseline and enhanced CNNs.

Training Dynamics and Optimization. Iwin-T exhibited more volatile training dynamics than ResNets, as seen in Figure 6, showing greater sensitivity to learning rate and batch settings. Careful scheduling and warm-up are needed to stabilize convergence.

Deployment Implications. ResNet-50 is ideal for stable, low-effort edge deployment. Enhanced ResNet-50 offers better accuracy with moderate regularization cost. Iwin-T delivers top accuracy but requires careful optimization. Future work will profile latency and memory for real-world deployment.

Overall, a spectrum of models exists to balance accuracy, stability, and deployment convenience. ResNet-50 is a safe default, while attention-augmented CNNs and Iwin-T offer promising paths for higher performance when properly regularized and optimized.

5. Conclusion

We present a comparative study of efficient architectures for low-resolution, on-device facial expression recognition. We benchmark ResNet-50, ResNet-152, attention-augmented ResNet-50, and the lightweight Iwin-T transformer on FER2013.

ResNet-50 provides reliable, stable performance (63.22% Top-1). Attention and class re-weighting improve accuracy to 67.19% but increase overfitting. Iwin-T achieves the best result at 68.09% Top-1, though it requires more careful optimization.

For practitioners, ResNet-50 remains an excellent stable baseline. For higher accuracy, enhanced CNNs are accessible with stronger regularization, while Iwin-T offers state-of-the-art potential with dedicated tuning. Future work will explore quantization, further regularization, and hardware deployment to realize efficient real-world FER systems.

Acknowledgment

LIAO PEICHUN, YANG ZITONG, LIAN KEXING and GU BAOQI contributed equally to this work.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[3] Simin Huo and Ning Li. Iwin transformer: Hierarchical vision transformer using interleaved windows, 2025.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[5] Ian J. Goodfellow et al. Challenges in representation learning: A report on three machine learning contests, 2013.

[6] Bin Li and Dimas Lima. Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering*, 2:57–64, 2021.

[7] Surya Petluru and Pradeep Singh. Transfer learning-based facial expression recognition with modified resnet50, 2022.

DOI: 10.54254/2755-2721/2026.GU33959

[8] Ryan Satria Wijaya et al. Comparative study of yolov5, yolov7 and yolov8 for robust outdoor detection. *Journal of Applied Electrical Engineering*, 8(1):37–43, 2024.

[9] Andrew Howard et al. Searching for mobilenetv3, 2019.

[10] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. Iwin: Human-object interaction detection via transformer with irregular windows, 2022.

[11] Daixin Li, Hai Nan, and Kai Zhao. Revolutionizing face recognition: Leveraging frformer and swin transformer. In *Proceedings of the International Conference on Computer Information and Big Data Applications (ICCIBA)*, pages 581–589, 2024.

[12] Dazhi Yao and Yunxue Shao. A data efficient transformer based on swin transformer. *The Visual Computer*, 40(4):2589–2598, 2023.

[13] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[14] Nitish Shirish Keskar et al. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.

[15] Simon S. Du et al. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima, 2018.