The Inductive Bias Gap: A Local-Focus Analysis of CNN's Superiority over ViT in Facial Expression Recognition

1st Wangshu Gong
Stony Brook Institute at Anhui University
Hefei, Anhui, China
R32314008@stu.ahu.edu.cn

2nd Shizhan Chen
Stony Brook Institute at Anhui University
Hefei, Anhui, China
R22314045@stu.ahu.edu.cn

***Abstract*—In facial expression recognition (FER) tasks that rely on local details, Vision Transformer (ViT) often underperforms compared to Convolutional Neural Networks (CNN). This paper hypothesizes that the root cause lies in the lack of task-aligned inductive bias in standard ViT, leading to an architecture-task mismatch. Systematic experiments based on the FER-2013 dataset diagnose this issue from the perspectives of data efficiency and attention mechanisms. The findings reveal that: (1) ViT exhibits low efficiency under limited data; (2) the proposed "Local Focus Ratio" metric indicates that ViT's global attention struggles to consistently concentrate on key facial regions such as the eyes, eyebrows, and mouth, whereas CNN accurately localizes them; (3) the inherent locality and translation invariance priors of CNN align well with the local action units critical for expression recognition. This study provides a mechanistic explanation for ViT's limitations in fine-grained tasks and suggests directions for improving Transformer architectures by incorporating local inductive biases, such as windowed attention. The conclusions can be extended to other similar tasks such as medical imaging and defect detection.**

***Index Terms*—Vision Transformer; Convolutional Neural Networks; Facial Expression Recognition; Inductive Bias**

I. Introduction

In recent years, the Vision Transformer (ViT) has fundamentally reshaped the paradigm of large-scale image recognition by discarding traditional convolutions and applying pure attention mechanisms to sequences of image patches. After pre-training on massive datasets like ImageNet-21k, its powerful capacity for modeling long-range dependencies has enabled it to surpass many classic Convolutional Neural Networks (CNNs) in performance, marking a new direction in visual representation learning. However, the effectiveness of this successful paradigm begins to be questioned when it is directly transferred to visual tasks with limited data scale that critically depend on the accurate perception of subtle local features. Among these, Facial Expression Recognition (FER), as a typical fine-grained visual understanding problem, prominently reveals this contradiction: FER, typically conducted on only thousands to tens of thousands of annotated samples, relies heavily on the precise identification of local key regions such as the eyes, eyebrows, and mouth—a strength that does not seem innate to the native ViT architecture.

Notably, several recent empirical studies have consistently found that in many standard FER benchmarks, vanilla ViT models often underperform a well-tuned CNN baseline [1]. This counterintuitive phenomenon raises a pressing core question: why does ViT, which excels in large-scale generic vision tasks, suffer a significant performance drop in the important application of FER? Currently, most related research remains at the level of simple performance comparison, lacking a systematic diagnosis and in-depth attribution of the fundamental mechanisms behind this phenomenon. This gap in understanding not only hinders rational architecture selection for tasks like FER but also delays the process of customizing Transformer models for fine-grained visual tasks.

To fill this gap, this paper aims to conduct a microscopic attribution analysis of ViT's performance on the FER task. We propose that the root cause of its performance shortcoming lies in a fundamental architecture-task mismatch: ViT inherently lacks the inductive biases required for the FER task. To validate this core argument, we design three progressively structured diagnostic experiments to investigate the following sub-hypotheses:

- Data Efficiency Gap (H1): ViT's performance is more sensitive to the amount of training data; its disadvantage compared to CNN is amplified in the small-data scenarios common for FER.
- Attention Mechanism Defocusing (H2): ViT's global attention mechanism struggles to automatically and stably focus on the local facial regions essential for expression discrimination, leading to dispersed feature representations.
- Representation Transfer Dilemma (H3): Due to the aforementioned mismatch between its mechanism and task requirements, the feature representations ViT learns from large-scale generic data cannot be effectively transferred and adapted to the local semantics specific to FER.

The main contributions of this paper are:

- We present the first systematic empirical diagnosis of ViT's limitations on the FER task, revealing the intertwined chain of effects involving data efficiency, inductive

bias, and knowledge transferability.

- We propose a novel method combining visualization and quantitative analysis. Through metrics such as the "Local Focus Ratio," we intuitively and conclusively reveal the dispersed nature of ViT's attention mechanism.
- Our findings not only explain the phenomenon but also provide direct implications for architecture selection. By introducing models with local priors (e.g., Swin Transformer) as a comparative baseline, we point the way for future ViT improvements tailored for fine-grained tasks. The conclusions of this study can be broadly generalized to other vision tasks with limited data and reliance on local features, such as medical image analysis and subtle defect detection.

II. Related Work

In the field of image classification, pioneering research has demonstrated that pure Transformer-based architectures, namely Vision Transformers (ViT), can achieve leading performance without relying on any convolutional operations [2]. They do so by segmenting images into sequences of patches and leveraging attention weights learned from large-scale data to model global relationships. Unlike the strong inductive biases inherent to Convolutional Neural Networks (CNNs), such as locality and translation equivariance, ViT possesses relatively weaker inductive biases. These are primarily manifested in its capability to model global correlations between image patches and its sensitivity to the order of the patch sequence. Although subsequent research, such as Swin Transformer, has successfully re-injected locality bias into this architecture through hierarchical design and local windowed attention, the fundamental characteristic of standard ViT remains unchanged: it lacks explicit structural constraints to guide the model's focus on local details [3].

In contrast, Facial Expression Recognition (FER) has long been regarded as a critical fine-grained visual understanding task. Early research in this area relied on handcrafted features. With the advent of the deep learning era, CNNs and their variants rapidly became dominant. A significant body of work has since focused on designing more efficient network architectures, attention modules, or loss functions, with the core objective of precisely capturing subtle deformations in local facial regions [4]. The success of these methods is fundamentally rooted in the alignment between CNN's inductive biases—particularly its local receptive fields and hierarchical feature extraction process—and the FER task's heavy reliance on strongly discriminative features from local expression units (e.g., eyes, eyebrows, mouth).

With the rise of Vision Transformers, research efforts have naturally attempted to apply them to the FER domain. However, most existing works directly fine-tune ViT models pre-trained on generic datasets (e.g., ImageNet) and limit their reporting to final classification performance comparisons. These empirical studies have commonly observed a phenomenon: on standard FER benchmark datasets, ViT often struggles to consistently outperform carefully optimized CNN baseline models [5]. Regarding the root cause of this performance gap, existing explanations tend to be superficial, simplistically attributing it to "insufficient data scale" while lacking deeper mechanistic analysis [6]. Currently, there remains a gap in research that systematically diagnoses this issue from perspectives such as the alignment of inductive biases, the interpretability of feature formation mechanisms, and the efficiency of cross-domain knowledge transfer. Targeting this critical gap, this paper aims to move beyond superficial performance comparisons and provide an interpretable, mechanism-level attribution analysis for ViT's performance on the FER task.

III. Experimental Methodology

A. Overview of Experimental Design

This study employs two sets of progressive experiments to validate three core hypotheses. Experiment 1 focuses on the "data efficiency gap," while Experiment 2 centers on the "mismatch of inductive biases." The study culminates in a theoretical analysis to argue for "prior-task compatibility." All experiments are conducted on the FER-2013 dataset. A controlled variable approach ensures the validity of the results, which, combined with quantitative measurements and visualization analysis, forms a complete chain of evidence.

B. Experiment 1: Diagnostic Experiment on Data Efficiency Gap

1) Experimental Design Principle: To validate Hypothesis H1 (ViT has lower data efficiency than CNN in data-limited FER tasks), a controlled comparative experiment is designed. Key variables—data split, augmentation strategies, and optimizer settings—are held constant. The only variable altered is the scale of training data to investigate the performance progression patterns of the two model types. The experiment uses three training data proportions (30%, 60%, 100%). All models are trained from scratch without loading pre-trained weights to eliminate interference from transfer learning.

2) Dataset and Preprocessing: The experiment is based on the FER-2013 dataset, which contains 35,887 48×48 grayscale facial images covering seven expression categories (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral) [7]. The preprocessing pipeline consists of:

- Normalization to the [0, 1] interval;
- Data augmentation for the training set (random horizontal flipping (p=0.5), ±15° rotation, ±10% translation);
- Only normalization for the validation and test sets. The data is split into training, validation, and test sets with a ratio of 8:1:1.

3) Model Architecture Implementation:

- CNN Model:A four-layer convolutional architecture is employed to embody the inductive biases of locality and hierarchy [8]. The first two layers use 3×3 convolutions to extract low-level features, while the subsequent two layers combine features using 5×5 and 3×3 convolutions. Each stage is followed by batch normalization, max

pooling (2×2), and Dropout (p=0.5), culminating in classification via a fully connected layer. The convolutional operation is defined as:

$$F_{i,j} = \sigma\left(\sum_{m=0}^{K-1}\sum_{n=0}^{K-1} W_{m,n} \cdot I_{i+m,j+n} + b\right) \quad (1)$$

where K is the convolutional kernel size, W is the weight matrix, I is the input feature map, and σ denotes the ReLU activation function.

- Vit Model: A standard architecture adapted for small input size is implemented: a 48×48 image is segmented into 6×6 patches (yielding 64 tokens). These are linearly projected into a 128-dimensional embedding space, added with learnable positional encodings. The model contains 4 Transformer encoder layers (with 4-head self-attention and feed-forward networks) and classifies via a [CLS] token. The self-attention mechanism is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the feature dimension.

4) Training and Evaluation Setup: The Adam optimizer is used with a learning rate of 0.0001 for the CNN and 0.0003 for the ViT [9]. The batch size is uniformly set to 64 for 30 training epochs, employing an early stopping strategy (halting if the validation accuracy does not improve for 5 consecutive epochs) [9]. The primary evaluation metric is the classification accuracy on the test set. Each experiment is repeated 3 times, and the average result is reported.

C. Experiment 2: Visualization and Quantitative Analysis

1) Experimental Design Rationale: To validate Hypothesis H2 (the underperformance of ViT is caused by a mismatch in inductive biases), the analysis unfolds across three dimensions: feature space distribution, attention mechanism, and local focus intensity. A combined approach of "qualitative visualization + quantitative metrics" is employed. The analysis uses the models trained in Experiment 1, with a random sample of 1000 instances from the test set.

2) Feature Space Visualization Method: The t-SNE dimensionality reduction technique is employed [10]. Features are extracted from the penultimate layer of the models (output of the CNN's fully connected layer, and the [CLS] token output for ViT). The t-SNE parameters are set with a perplexity of 30, a maximum of 1000 iterations, and a random seed of 42 to project the high-dimensional features onto a 2D space, facilitating the analysis of intra-class compactness and inter-class separation.

3) Attention Mechanism Visualization Technique:

- Heatmap Generation: The Grad-CAM technique is utilized to generate visual explanations [11]. The heatmap for class c is calculated as:

$$H_c(x, y) = \text{ReLU}\left(\sum_k \alpha_k^c \cdot A^k(x, y)\right) \quad (3)$$

where A^k is the k-th feature map, and α_k^c is the weight for class c corresponding to feature map k, obtained via gradient global average pooling.

- Contour Analysis and Bounding Box Annotation: Connected regions are extracted using 8-level thresholds (from 0.5 to 0.95). Noise regions smaller than 30 pixels are filtered out, and the Douglas-Peucker algorithm is applied to simplify contours []. Colored bounding boxes are drawn using an adaptive threshold (from 50% down to 30%), where the color intensity corresponds to the heatmap value.

4) Quantitative Analysis Metrics:

- Local Focus Ratio (LFR): This metric quantifies the model's degree of focus on key facial regions. It is defined as:

$$\text{LFR} = \frac{\frac{1}{N_{\text{key}}}\sum_{(x,y)\in\text{Key}} H(x, y)}{\frac{1}{N_{\text{non-key}}}\sum_{(x,y)\notin\text{Key}} H(x, y) + \epsilon} \quad (4)$$

where $\epsilon = 10^{-8}$, and the key region mask is defined as shown in Fig1.

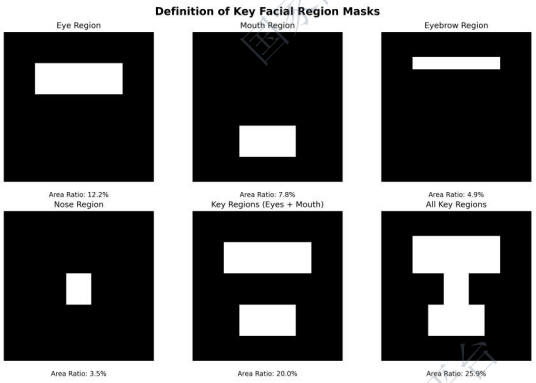

Fig. 1. Definition of Key Facial Region Masks

- Local Sensitivity Index (LSI): This metric quantifies the prior-task compatibility by measuring the alignment between the model's attention and predefined regions of interest. It is defined as:

$$\text{LSI}(M) = \frac{1}{n_c}\sum_{i=1}^{n_c} \text{IoU}\big(A_M(R_i),\, R_i\big) \quad (5)$$

where $A_M(R_i)$ denotes the attention region of model M for the i-th key region, and IoU represents the Intersection over Union [13].

IV. Experimental Conclusions

A. Overview of Core Findings

Through two sets of comparative experiments and theoretical analysis, this study validates three core hypotheses: in data-limited FER tasks, ViT exhibits significantly lower data efficiency than CNN (H1), the root cause of which

lies in the mismatch between its inductive bias and task requirements (H2). In contrast, CNN's locality prior forms a precise match with the local feature dependency of the FER task (H3). The experimental results demonstrate that the degree of compatibility between architectural priors and task characteristics is a key factor determining model performance, rather than mere differences in representational capacity.

B. Data Efficiency Gap

Experiment 1 quantified the performance difference between the two model types across varying data scales, as shown in Fig2. CNN outperformed ViT at all data proportions, leading

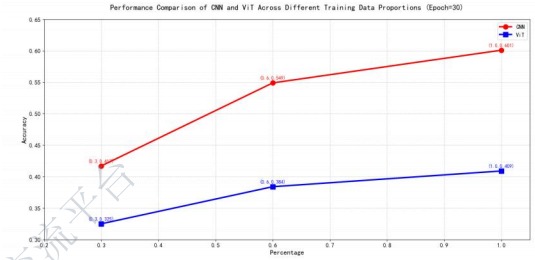

Fig. 2. Performance Comparison of CNN and Vit Across Different Training Data Proportions (Epoch=30)

by 9.2% at 30% data, 16.5%points at 60%, and 19.2% at 100% data. This refutes the superficial attribution of the performance gap solely to "insufficient data". More critically, the performance improvement rates differed: when data increased from 30% to 100%, CNN's accuracy improved by 18.4%, whereas ViT's improved by only 8.4%, with CNN's improvement curve being steeper.

C. Mismatch of Inductive Biases

1) Differences in Feature Space Distribution: The t-SNE visualization, shown in Fig3, reveals that CNN features form clear cluster structures (intra-class compactness and inter-class separation). This indicates that CNN's local convolu-

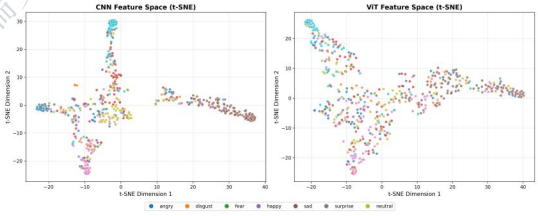

Fig. 3. CNN and Vit Feature Space (t-SNE)

tions compel the model to learn discriminative local features, whereas ViT's global attention struggles to distill effective local patterns.

2) Attention Mechanism and Quantitative Metrics: A comparison of attention heatmaps shows that CNN's strongest attention regions are 100% concentrated on key facial areas (eyes, mouth), highly consistent with facial Action Units (AUs). In contrast, only 26% of ViT's attention resides in these key regions, often focusing on irrelevant areas like the background. Quantitative comparisons are shown in TABLE 1. The results indicate that CNN's LFR and LSI are significantly higher than ViT's. This quantitatively confirms H2: CNN's locality bias enables it to spontaneously focus on key regions, while ViT's lack of such constraints leads to dispersed attention.

TABLE I
QUANTITATIVE METRICS FOR ATTENTION REGIONS OF CNN AND VIT (N=200)

Region	Mean LFR ± SD	
	CNN	ViT
EYES	1.556 ± 0.495	0.874 ± 0.222
MOUTH	1.931 ± 0.750	0.865 ± 0.254
EYEBROW	1.372 ± 0.530	0.856 ± 0.278
NOSE	1.497 ± 0.425	0.769 ± 0.209
KEY Region	1.846 ± 0.358	0.861 ± 0.213
ALL Region	1.896 ± 0.398	0.828 ± 0.216

Prop. ¿1.0	Effect Size	
CNN / ViT	(Cohen's d)	Conclusion
87.0% / 27.5%	1.777	Sig. higher
98.5% / 26.0%	1.903	Sig. higher
75.0% / 27.5%	1.218	Sig. higher
92.0% / 11.0%	2.171	Sig. higher
100.0% / 26.0%	3.345	Sig. higher
100.0% / 18.0%	3.335	Sig. higher

D. Prior-Task Compatibility

1) The Adaptive Advantage of the Locality Prior: CNN's locality prior, enforced through limited receptive fields and weight sharing, achieves precise compatibility with the local feature dependency of the FER task. Mathematically, the convolution operation (Formula 1) compels feature extraction based on local neighborhoods, perfectly aligning with the spatial locality (a single AU affects i20% of facial area) and combinatorial sparsity (2-4 AUs determine an expression) of expression features. In contrast, ViT's global attention (Formula 2) suffers from a local detail dilution effect and is prone to overfitting spurious global correlations in small-sample scenarios [14].

2) Theoretical Trade-off and Implications: Architecture design fundamentally involves a trade-off between inductive bias and flexibility. In the FER task, CNN's strong locality bias introduces beneficial bias, reducing variance. While ViT's weaker bias enhances flexibility, it leads to sharply increased variance when data is insufficient. The Local Sensitivity Index (LSI) quantifies this compatibility: the CNN/ViT focus ratio in key facial regions is 2.15. CNN's attention focus in key facial areas is significantly higher than ViT's. This theory points to directions for improving ViT:

- Introducing local attention constraints (e.g., windowed attention) [15].
- Hybridizing CNN's local feature extraction with ViT's global modeling.

- Regularizing attention distribution through loss functions.

The core principle is to introduce a moderate locality bias while preserving global modeling advantages, thereby matching the requirements of locally-sensitive tasks.

V. Conclusion

This study, through systematic empirical diagnosis and mechanistic analysis, has clearly identified the core reason for the underperformance of Vision Transformers (ViT) compared to Convolutional Neural Networks (CNN) in Facial Expression Recognition (FER) tasks: a fundamental mismatch between the architectural inductive biases and the requirements of the FER task, rather than merely limitations in data scale or representational capacity. This finding provides crucial theoretical guidance for architecture selection and model improvement in fine-grained visual tasks.

Experimental data thoroughly validated three core hypotheses: In the typical small-data scenario of FER, ViT's data efficiency is significantly lower than CNN's. Even when training data increased to the full set, its accuracy lagged by 19.2 percentage points, with slower loss reduction and longer convergence cycles, highlighting its "data-hungry" nature. Attention mechanism analysis revealed that ViT's global attention lacks explicit local constraints, with only 26

The core contribution of this study lies in moving beyond superficial performance comparisons. For the first time, we constructed a complete attribution chain for ViT's shortcomings in fine-grained tasks from the perspectives of inductive bias compatibility, feature formation mechanisms, and knowledge transferability. We also proposed quantitative metrics such as the "Local Focus Ratio," providing actionable tools for mechanistic analysis. The conclusions of this research can be widely generalized to other vision tasks with limited data and reliance on local features, such as medical image analysis and subtle defect detection.

Future improvement directions should focus on "injecting locality bias": by incorporating windowed attention, hybridizing CNN-based local feature extraction modules, or designing attention regularization loss functions. The goal is to enhance ViT's perception of local key regions while preserving its advantages in global modeling, thereby achieving precise alignment between architectural inductive biases and the requirements of fine-grained tasks. This will promote the effective application of Transformers in a wider range of specific visual tasks.

References

[1] Kim H, Ko B C. Rethinking attention mechanisms in vision transformers with graph structures[J]. Sensors, 2024, 24(4): 1111.

[2] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[3] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[4] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[5] Song H. Facial Expression Recognition with ViT Considering All Tokens towards More Informative Self-attention Outputs[J]. network, 2023, 41.

[6] Li J, Zhang Z. Facial expression recognition using vanilla vit backbones with mae pretraining[J]. arXiv preprint arXiv:2207.11081, 2022.

[7] Goodfellow I J, Erhan D, Carrier P L, et al. Challenges in representation learning: A report on three machine learning contests[C]//International conference on neural information processing. Berlin, Heidelberg: Springer berlin heidelberg, 2013: 117-124.

[8] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[9] Kinga D, Adam J B. A method for stochastic optimization[C]//International conference on learning representations (ICLR). 2015, 5(6).

[10] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(Nov): 2579-2605.

[11] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[12] Hershberger J E, Snoeyink J. Speeding up the Douglas-Peucker line-simplification algorithm[J]. 1992.

[13] Montavon G, Binder A, Lapuschkin S, et al. Explainable AI: interpreting, explaining and visualizing deep learning[J]. Springer LNCS, 2019, 11700(1).

[14] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.

[15] Fang Y, Wang X, Wu R, et al. What makes for hierarchical vision transformer?[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12714-12720.