Facial Expression Recognition in Anime and Manga Characters: A Comparative Study of Vision Transformers and Convolutional Neural Networks

Elia Santoro , Luigi Laura * , Marco Parrillo * , Valerio Rughetti

Posted Date: 20 April 2026

doi: 10.20944/preprints202604.0729.v2

Keywords: facial expression recognition; deep learning; vision transformer; convolutional neural network; resnet; anime; manga; transfer learning; computer vision

 doi:10.20944/preprints202604.0729.v2

Article

Facial Expression Recognition in Anime and Manga Characters: A Comparative Study of Vision Transformers and Convolutional Neural Networks

Elia Santoro [1], Luigi Laura [1,*], Marco Parrillo [1,*] and Valerio Rughetti [2]

[1] Faculty of Engineering, International Telematic University Uninettuno, 00186 Rome, Italy
[2] Department of Computer Science, LUMSA, Italy
* Correspondence: luigi.laura@uninettunouniversity.net (L.L); marcoparrillo@gmail.com (M.P.)

Abstract

Facial expression recognition (FER) is a well-established task in computer vision, yet its application to non-photorealistic domains, such as anime and manga, remains largely underexplored. The stylized, exaggerated, and often non-proportional facial features of illustrated characters present unique challenges for deep learning models trained predominantly on realistic imagery. In this work, we construct a balanced dataset of 3,000 manga and anime face images spanning six emotion categories (Angry, Embarrassed, Happy, Psycho-Crazy, Sad, Scared) and conduct a systematic comparison of two major deep learning paradigms: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Specifically, we evaluate ResNet-18, ResNet-50, ViT-B/16, and ViT-S/16 under four fine-tuning strategies: linear probing, partial fine-tuning, full fine-tuning, and progressive unfreezing; enabling a controlled comparison of both architectural families and transfer learning depth. Our results show that fine-tuning strategy significantly impacts performance: the best configuration (ViT-B/16 with progressive unfreezing) achieves 80.89% test accuracy, compared to 61.33% for the weakest linear probe baseline (ViT-S/16), a gap of 19.56 percentage points. Vision Transformers benefit disproportionately from fine-tuning, and the relative ranking of architectures changes across fine-tuning regimes. Confusion matrix analysis reveals persistent cross-class confusion between visually similar emotions (e.g., Happy vs. Embarrassed), while highly distinctive categories such as Psycho-Crazy are consistently well recognized across all architectures.

Keywords: facial expression recognition; deep learning; vision transformer; convolutional neural network; resnet; anime; manga; transfer learning; computer vision

1. Introduction

Emotion recognition from facial expressions is a central problem in affective computing and computer vision [3]. Its applications span human-computer interaction, healthcare, entertainment, and security. While significant progress has been made using large-scale photographic datasets such as FER-2013 [14], RAF-DB [13], and AffectNet [12], these advances are predominantly confined to realistic imagery depicting human faces.

The domain of anime and manga presents a distinct and largely unexplored challenge. Characters in these media are rendered with stylized, non-proportional features: exaggerated eyes, simplified mouths, and intentional deformations of the face designed to convey emotional states [15]. These artistic conventions diverge substantially from the anatomical regularities that deep learning models exploit when processing photographic faces. Consequently, feature extractors pre-trained on ImageNet or photographic FER datasets may not transfer effectively to this non-photorealistic domain. A further complication is that anime and manga imagery lacks the photorealistic textures, lighting gradients, and color naturalism that dominate ImageNet representations, raising an open question

 doi:10.20944/preprints202604.0729.v2

about whether Transformer-based architectures, which have recently surpassed CNNs on photographic FER benchmarks, offer a similar advantage in stylized illustrated domains.

Prior work on emotion recognition in illustrated media is limited and methodologically narrow. Hill [15] applied a custom CNN to a 4,800-image dataset of Tom and Jerry faces across three emotions, achieving approximately 80% accuracy. Köklü [16] used a PyTorch-based CNN for manga facial expression classification on a small dataset of five emotions. More recently, Parrillo et al. [7] conducted a systematic comparison of CNN architectures and transfer learning strategies for comic character recognition, demonstrating that pretrained ResNet-50 features transfer effectively to stylized imagery even under constrained data conditions. Critically, none of these studies has compared CNN and Vision Transformer architectures within the illustrated media domain, leaving open the question of whether the architectural advantages that ViTs demonstrate on photographic FER benchmarks generalise to the qualitatively different visual statistics of comic and manga imagery. The present study addresses this gap directly.

In this paper, we ask: *do Vision Transformers offer a systematic advantage over CNNs for facial expression recognition in anime and manga, and does the choice of fine-tuning strategy mediate this relationship?* We address this question with the following contributions:

1. **Dataset construction.** We assemble a balanced dataset of 3,000 anime and manga face images across six emotion categories, integrating multiple sources and performing manual annotation.
2. **Systematic architecture comparison.** We evaluate four models, ResNet-18, ResNet-50, ViT-B/16, and ViT-S/16, under identical preprocessing and evaluation conditions, enabling the first controlled comparison between CNN and Transformer paradigms specifically in the illustrated media domain.
3. **Multi-strategy transfer learning analysis.** We compare four fine-tuning strategies, linear probing, partial fine-tuning, full fine-tuning, and progressive unfreezing, across all architectures, demonstrating that fine-tuning depth significantly affects both absolute performance and the relative ranking of models.
4. **Detailed error analysis.** We provide per-class confusion matrix analysis, identifying systematic patterns of misclassification linked to visual similarity between emotion categories and to the artistic conventions of the illustrated domain.
5. **Efficiency analysis.** We report training speed, inference throughput, and computational cost alongside accuracy metrics, offering practical guidance for resource-constrained deployment on small non-photorealistic datasets.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset construction process. Section 4 details the architectures and training methodology. Section 5 presents and analyzes the experimental results. Section 6 discusses the results. Section 7 concludes with a discussion of limitations and future directions.

2. Related Work

2.1. Facial Expression Recognition

Computational models of facial expression recognition trace their origins to early neural network architectures [1,2], but practical progress accelerated with the introduction of hand-crafted feature descriptors and, later, deep convolutional networks. Facial expression recognition has since evolved from hand-crafted feature approaches, using descriptors such as LBP [6], HOG [5], and SIFT [4] paired with classifiers like SVM or Random Forest, to end-to-end deep learning methods. The introduction of CNNs, particularly AlexNet [8] and later ResNet [9], marked a paradigm shift by enabling automatic hierarchical feature extraction directly from raw pixel data.

ResNet [9] introduced residual connections (skip connections) that allow the training of very deep networks by mitigating the vanishing gradient problem. As shown in Figure 1, the identity shortcut connection bypasses weight layers, enabling the network to learn a residual function $\mathcal{F}(\mathbf{x})$ rather than

 doi:10.20944/preprints202604.0729.v2

a complete mapping. On standard benchmarks, ResNet-based models achieve accuracies exceeding 70% on FER-2013 and 80–85% on RAF-DB.

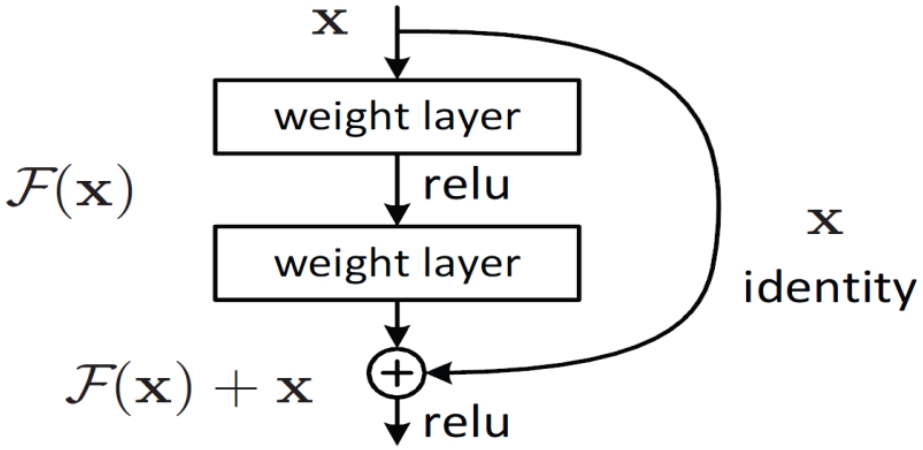


Figure 1. Residual learning block in ResNet. The identity shortcut connection bypasses two weight layers, allowing the network to learn a residual function $\mathcal{F}(\mathbf{x})$ rather than a complete mapping [9].

The Vision Transformer (ViT), introduced by Dosovitskiy et al. [11], adapted the Transformer architecture [10] from natural language processing to image classification. ViT divides an image into fixed-size patches (e.g., 16×16), linearly embeds them, and processes the resulting sequence through a standard Transformer encoder with global self-attention. This mechanism enables the model to capture long-range dependencies between distant regions of the image from the first layer, a capability that CNNs achieve only through stacking many convolutional layers.

2.2. Emotion Recognition in Illustrated Media

The literature on emotion recognition in non-photorealistic imagery is sparse. Hill [15] trained a custom CNN on 4,800 frames of Tom and Jerry across three emotions (happy, surprised, angry), reporting 80% accuracy, 56% recognition rate, and 59% validation rate. Köklü [16] developed a CNN-based classifier for manga facial expressions using the "Manga Facial Expression" dataset with a custom architecture, without leveraging pre-trained models or comparing with Transformer-based approaches.

Camerlingo et al. [21] explored a broader pipeline for neural network applications in comic strips, using Dilbert panels as a pedagogical case study. Their approach combined cascade classifiers for character face detection, OCR-based text extraction, and a Transformer network for dialogue generation. While not focused on emotion recognition per se, their work demonstrates the feasibility of applying deep learning to illustrated media for multiple tasks, including face detection, and highlights the unique challenges posed by the stylized, non-photorealistic visual conventions of comics. Their pipeline's reliance on face detection as a prerequisite step underscores the importance of robust feature extraction in illustrated domains, a challenge that directly motivates the present study.

Parrillo et al. [7] conducted a systematic evaluation of transfer learning strategies for comic character recognition under constrained data conditions, comparing a baseline CNN, a regularized CNN, a frozen ResNet-50 feature extractor, and a partially fine-tuned ResNet-50 on a custom four-class Dilbert dataset of 682 images. Using both a fixed 70/20/10 split and 5-fold cross-validation, they found that both ResNet-50 strategies achieve equivalent mean cross-validated accuracy of 95.0%, while shallow CNNs reach only 81–87%. A key methodological finding of that study is that evaluation-protocol selection critically affects perceived model ranking: fine-tuning nominally achieves 98.5% under a single fixed partition, but cross-validation deflates this figure to parity with frozen feature extraction, exposing it as a partition artifact. The authors therefore recommend frozen ResNet-50 feature extraction as the preferred strategy in low-data stylized domains, given its lower variance and 15× reduction in trainable parameters. The present work extends this line of inquiry in two important

 doi:10.20944/preprints202604.0729.v2

respects: first, by targeting a qualitatively different task (emotion classification rather than character identity recognition) on a larger and more diverse illustrated dataset; second, by incorporating Vision Transformers alongside CNNs, enabling a direct architectural comparison that the comic character recognition study did not address.

Taken together, these studies establish the feasibility of applying deep learning to illustrated media for both character recognition and emotion classification tasks, and demonstrate that pretrained convolutional features transfer effectively to stylized domains even under constrained data conditions [7]. However, a consistent limitation across this body of work is the exclusive reliance on CNN architectures: no prior study has evaluated Vision Transformers in the illustrated media domain, nor has any work systematically compared CNN and Transformer paradigms for facial expression recognition in anime or manga. Furthermore, the interaction between architectural family and fine-tuning strategy, which has been shown to critically affect performance in domain-shifted settings [7], remains unexplored for the emotion classification task. The present work addresses these gaps by providing the first controlled CNN–Transformer comparison for anime and manga FER, evaluated across multiple transfer learning strategies on a purpose-built balanced dataset of six emotion categories.

2.3. Subjectivity of Emotion Categories

Barrett et al. [17] argued that facial expressions do not constitute universal, unambiguous signals but depend on context and observer interpretation. Russell's Circumplex Model of Affect [18] further posits that emotions are not rigid categories but vary along two continuous dimensions: valence (positive–negative) and arousal (high–low). These theoretical perspectives underscore the inherent difficulty of discrete emotion classification, a challenge amplified in illustrated media where artistic style introduces additional variability.

3. Dataset Construction

3.1. Design Principles

A key requirement for reliable model evaluation is a balanced, sufficiently large, and consistently annotated dataset. We designed our dataset around three principles: (1) class balance to avoid bias toward dominant categories, (2) diversity of artistic styles to improve generalization, and (3) inclusion of both easily distinguishable and visually ambiguous emotion classes.

3.2. Data Sources

The final dataset comprises 3,000 images drawn from three sources:

(1) Manga Facial Expression Dataset [16]: 450 pre-annotated images across five emotions (happy, sad, angry, surprised, scared), serving as the initial seed.

(2) Face of Pixiv Top Daily Illustration 2018 [20]: Approximately 10,000 anime-style face images. Since this dataset lacked emotion labels, manual classification was performed, yielding ∼1,500 usable images. An initial severe imbalance (1,250 images in the "Happy" class alone) was resolved by down-sampling to 500 per class.

(3) Pinterest: Manual collection to fill remaining gaps, ensuring 500 images per class across all six categories. Images were collected manually from public search results for research purposes only, under fair use principles consistent with academic non-commercial use; no images were obtained by circumventing access restrictions or authentication mechanisms.

3.3. Emotion Categories

Six emotion classes were defined, each containing exactly 500 images: **Angry** (rage, frustration), **Embarrassed** (flushed, sheepish), **Happy** (smiling, joyful), **Psycho-Crazy** (maniacal, unhinged), **Sad** (sorrow, melancholy), and **Scared** (fear, shock).

The "Psycho-Crazy" class was introduced specifically to test model behavior on highly distinctive, exaggerated visual patterns that are common in manga but absent from standard FER benchmarks.

 doi:10.20944/preprints202604.0729.v2

3.4. Annotation and Data Split

All images were annotated and managed via Roboflow [19], which provided structured labeling, cloud storage, dataset versioning, and export in PyTorch-compatible format. The manual annotation process spanned several months and was performed by a single annotator. This constitutes a known limitation of the dataset: single-annotator labeling introduces the risk of subjective bias, particularly for visually ambiguous emotion pairs such as Happy and Embarrassed, where reasonable annotators may disagree. No inter-annotator agreement metric (e.g., Cohen's kappa) was computed during dataset construction, and the extent of label noise in the ground truth therefore cannot be quantified from the annotations alone. Readers should interpret classification performance on ambiguous categories with this caveat in mind; the implications for model evaluation are discussed further in Section 6.

The dataset was split into three stratified partitions maintaining class balance across all six categories: 70% training (2,100 images, 350 per class), 15% validation (450 images, 75 per class), and 15% test (450 images, 75 per class). The split was generated once via Roboflow's versioning system and frozen before any experiments began. The training set was used exclusively for gradient updates. The validation set was used for early stopping decisions and checkpoint selection. The held-out test set was reserved for final evaluation and was accessed only once, after all training and model selection decisions had been finalized, to provide an unbiased estimate of generalization performance.

4. Methodology

4.1. Architectures

We evaluate four models spanning two architectural families:

4.1.1. ResNet-18

A lightweight CNN with 18 layers and approximately 11.7M parameters. Its shallow depth makes it fast to train and resistant to overfitting on small datasets, while residual connections ensure stable gradient flow [9].

4.1.2. ResNet-50

A deeper CNN with 50 layers using bottleneck residual blocks, totaling approximately 25.6M parameters. It offers greater representational capacity but requires more computational resources and is more susceptible to overfitting on limited data.

4.1.3. ViT-B/16

The base Vision Transformer with 16×16 patch size, comprising 12 Transformer encoder layers, 12 attention heads, and approximately 86M parameters. The global self-attention mechanism enables the model to capture long-range spatial dependencies from the first layer [11].

4.1.4. ViT-S/16

A smaller Vision Transformer variant with reduced embedding dimension and fewer attention heads (∼22M parameters), providing a compromise between the computational cost of ViT-B/16 and the local inductive bias of CNNs.

4.2. Fine-Tuning Strategies

To investigate how the depth of fine-tuning affects performance across architectural families, we evaluate four transfer learning strategies under controlled conditions. All strategies share a common configuration: ImageNet-1K pre-trained weights, AdamW optimizer, cross-entropy loss on raw logits, batch size of 32, input image size of 224 × 224 pixels, and a fixed random seed of 42 for reproducibility. The complete hyperparameter configuration for each strategy is reported in Table 1.

 doi:10.20944/preprints202604.0729.v2

Table 1. Hyperparameter configuration for each fine-tuning strategy, as used in all experiments. `lr_head` and `lr_backbone` denote the learning rates applied to the classification head and the unfrozen backbone layers, respectively. `unfreeze_every` denotes the number of epochs between successive block-unfreezing steps, applicable only to the progressive strategy.

Strategy	lr_head	lr_backbone	wd	Epochs	Patience	Unfreeze every
Linear probe	10^{-3}	—	10^{-4}	20	5	—
Partial	10^{-3}	10^{-4}	10^{-4}	30	7	—
Full	10^{-3}	10^{-5}	10^{-4}	40	10	—
Progressive	10^{-3}	10^{-4}	10^{-4}	50	10	5 epochs

4.2.1. Strategy A: Linear Probe (Frozen Backbone)

All backbone parameters are frozen (`requires_grad = False`); only the final classification head (a single fully connected layer mapping to `NUM_CLASSES` = 6 output logits, replacing the original ImageNet 1,000-class layer) is trained. This strategy evaluates the quality of pre-trained features without any domain adaptation of the feature extractor.

Differential learning rates: $lr_{\text{head}} = 10^{-3}$; backbone learning rate is not applicable as all backbone parameters are frozen. Weight decay: 10^{-4}. Training is performed for a maximum of 20 epochs with early stopping enabled (patience = 5), monitoring validation loss. Model weights are restored to the best epoch upon termination.

4.2.2. Strategy B: Partial Fine-Tuning

The classification head and the final residual block (ResNet) or the final encoder block group (ViT) are unfrozen, while all earlier layers remain frozen. This strategy enables the model to adapt high-level feature representations to the illustrated domain while preserving low-level visual features learned from ImageNet.

Differential learning rates are applied: $lr_{\text{head}} = 10^{-3}$ for the classification head and $lr_{\text{backbone}} = 10^{-4}$ for the unfrozen backbone layers, a $10\times$ reduction relative to the head. Weight decay: 10^{-4}. Training is performed for a maximum of 30 epochs with early stopping enabled (patience = 7).

4.2.3. Strategy C: Full Fine-Tuning

All parameters are trainable from the start of training. Differential learning rates are used to mitigate catastrophic forgetting: $lr_{\text{head}} = 10^{-3}$ for the classification head and $lr_{\text{backbone}} = 10^{-5}$ for the backbone layers, a $100\times$ reduction relative to the head. Weight decay: 10^{-4}. Gradient clipping (max norm = 1.0) and cosine annealing with linear warmup (3 epochs) are applied to improve training stability. Training is performed for a maximum of 40 epochs with early stopping enabled (patience = 10).

4.2.4. Strategy D: Progressive Unfreezing

Progressive unfreezing gradually exposes backbone layers to gradient updates over the course of training, rather than unfreezing them all at once. Training begins with all backbone parameters frozen and only the classification head trainable. Every `unfreeze_every` = 5 epochs, one additional backbone block is unfrozen, proceeding from the deepest layers toward the earliest, following the principle that higher-level task-specific representations should be adapted before lower-level general-purpose features [11]. This schedule was held constant across all four architectures; no architecture-specific tuning of the unfreezing interval was performed.

Differential learning rates are applied throughout: $lr_{\text{head}} = 10^{-3}$ for the classification head and $lr_{\text{backbone}} = 10^{-4}$ for any currently unfrozen backbone layers. Weight decay: 10^{-4}. Training is performed for a maximum of 50 epochs (the longest budget of any strategy, to accommodate the gradual unfreezing schedule) with early stopping enabled (patience = 10). As with Strategy C, the longer patience and epoch budget reflect the slower convergence dynamics expected when backbone adaptation is introduced incrementally.

 doi:10.20944/preprints202604.0729.v2

4.3. Shared Experimental Settings

All four strategies share the following settings:

- **Pre-trained weights:** ImageNet-1K.
- **Loss function:** Cross-entropy on raw logits.
- **Optimizer:** AdamW with per-group learning rates as specified above and weight decay $= 10^{-4}$ uniformly.
- **Batch size:** 32.
- **Image size:** 224 $\times$ 224 pixels (`IMG_SIZE` = 224).
- **Input preprocessing (training):**
  - resize to 256, random resized crop to 224 $\times$ 224 (scale 0.8–1.0)
  - random horizontal flip ($p = 0.5$)
  - color jitter (brightness $= 0.2$, contrast $= 0.2$, saturation $= 0.2$, hue $= 0.05$)
  - random rotation ($\pm 10^{\circ}$)
  - normalization to ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$)
  - random erasing ($p = 0.1$)
- **Input preprocessing (evaluation):** resize to 256, center crop to 224 $\times$ 224, normalization to ImageNet statistics.
- **Hardware:** Google Colab with CUDA-accelerated GPU.
- **Framework:** PyTorch 2.x with Torchvision.
- **Random seed:** 42 (fixed across all runs for reproducibility; `SEED` = 42).
- **Class labels:** `CLASSES` = [Angry, Embarrassed, Happy, Psycho-Crazy, Sad, Scared], `NUM_CLASSES` = 6.

4.4. Evaluation Metrics

Models were evaluated using accuracy, cross-entropy loss, precision (macro), recall (macro), F1-score (macro and weighted), training speed (average time per epoch and images per second), and per-class confusion matrices. The core per-class metrics are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Evaluation followed a two-stage protocol. During training, all metrics were computed on the validation set to guide early stopping and checkpoint selection. After all training decisions were finalized, each model's selected checkpoint was evaluated a single time on the held-out test set. The test-set metrics constitute the primary results reported in this paper. Validation metrics are reported alongside for transparency, allowing readers to assess the gap between the model-selection criterion and true generalization performance.

5. Experimental Results

All the code used in this study is available at [22].

5.1. Overall Performance

Table 2 summarizes the test set performance of all four models across four fine-tuning strategies. The results reveal two distinct performance regimes.

Under **linear probing** (frozen backbone), all models cluster within a narrow 3.3 percentage point band (61.33%–64.67%), with ViT-B/16 achieving the highest accuracy (64.67%) and lowest loss (0.972). This narrow spread indicates that ImageNet pre-trained features transfer to the anime/manga

 doi:10.20944/preprints202604.0729.v2

domain with roughly comparable effectiveness across architectures when no domain adaptation of the backbone is permitted.

Under **fine-tuning**, performance improves substantially for all models, but the gains are unevenly distributed. The best configuration, ViT-B/16 with progressive unfreezing, reaches 80.89% test accuracy ($F1_{macro} = 0.807$), representing a 16.22 percentage point improvement over its linear probe baseline. At the other end, ViT-S/16 with partial fine-tuning achieves only 66.22%, a modest 4.89 percentage point gain. The spread between the best and worst configurations across the entire experiment matrix is 19.56 percentage points (80.89% vs. 61.33%), demonstrating that the choice of fine-tuning strategy has a larger impact on performance than the choice of architecture.

Both CNN models achieve their best results under partial fine-tuning (ResNet-18: 75.11%; ResNet-50: 74.22%), while the best ViT results require deeper adaptation: full fine-tuning for ViT-S/16 (76.22%) and progressive unfreezing for ViT-B/16 (80.89%). Section 6.4 provides a detailed analysis of these architecture–strategy interactions.

ResNet-50, the weakest model under linear probing (61.78%), recovers to 74.22% with partial fine-tuning, suggesting that its poor frozen-backbone performance reflects limited feature transferability rather than a fundamental capacity mismatch. ViT-B/16 consistently achieves the lowest test loss across most strategies, indicating well-calibrated probability estimates that may be valuable for downstream applications where confidence scores matter.

Table 2. Final comparison across models and training strategies.

Model	Strategy	%Tr	Acc	$F1_m$	$F1_w$	Loss	Ep	Time
resnet18	full	100.0	74.67	0.7405	0.7405	0.7008	26	25.9s
resnet18	linear_probe	0.0	62.22	0.6119	0.6119	1.0158	20	22.9s
resnet18	partial	75.1	75.11	0.7480	0.7480	0.7358	**10**	**22.7s**
resnet18	progressive	98.6	74.89	0.7480	0.7480	0.8431	20	22.8s
resnet50	full	100.0	73.78	0.7325	0.7325	0.7554	19	44.9s
resnet50	linear_probe	0.1	61.78	0.6014	0.6014	1.0118	19	26.4s
resnet50	partial	63.7	74.22	0.7367	0.7367	0.8209	12	29.9s
resnet50	progressive	99.0	73.56	0.7321	0.7321	0.9097	18	32.0s
vit_b_16	full	100.0	74.44	0.7421	0.7421	0.6547	12	105.2s
vit_b_16	linear_probe	0.0	64.67	0.6431	0.6431	0.9722	20	47.4s
vit_b_16	partial	24.8	70.67	0.7035	0.7035	0.7838	**10**	61.0s
vit_b_16	progressive	99.8	**80.89**	**0.8066**	**0.8066**	0.8701	29	79.9s
vit_s_16	full	100.0	76.22	0.7607	0.7607	**0.6422**	18	43.8s
vit_s_16	linear_probe	0.0	61.33	0.6073	0.6073	0.9742	20	26.4s
vit_s_16	partial	24.6	66.22	0.6579	0.6579	0.8920	11	31.0s
vit_s_16	progressive	73.7	66.67	0.6604	0.6604	0.9186	19	32.6s

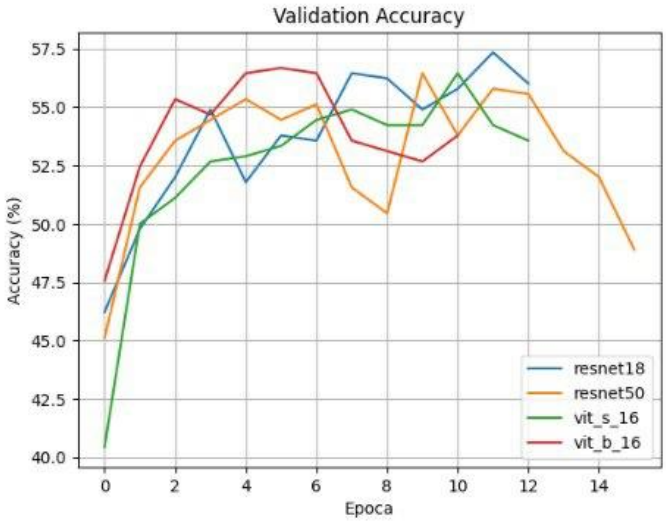
Note: Balanced dataset $\Rightarrow F1_w \approx F1_m$. *Best values in bold.*

5.2. Training Dynamics

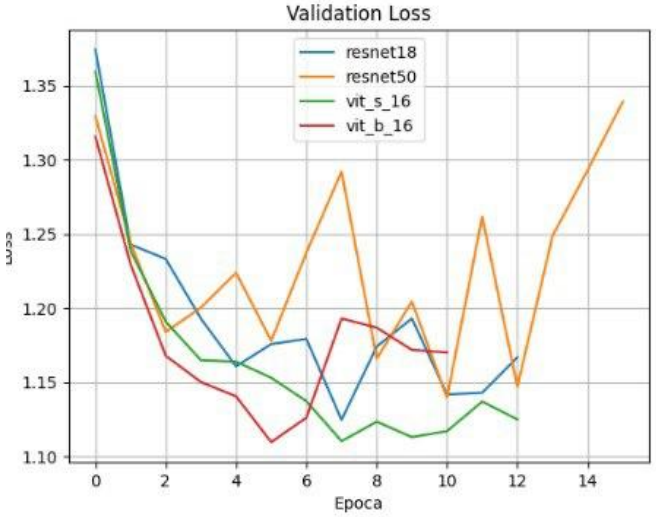
Figure 2 shows the validation accuracy and loss curves across all epochs **under the linear probe setting**. Fine-tuning training dynamics are not shown but follow qualitatively similar convergence patterns; per-epoch training times for all strategies are reported in Table 2.

All models exhibit rapid improvement during the first 5 epochs, followed by convergence between epochs 8 and 12.

 doi:10.20944/preprints202604.0729.v2



(a) Validation accuracy across epochs.



(b) Validation loss across epochs.

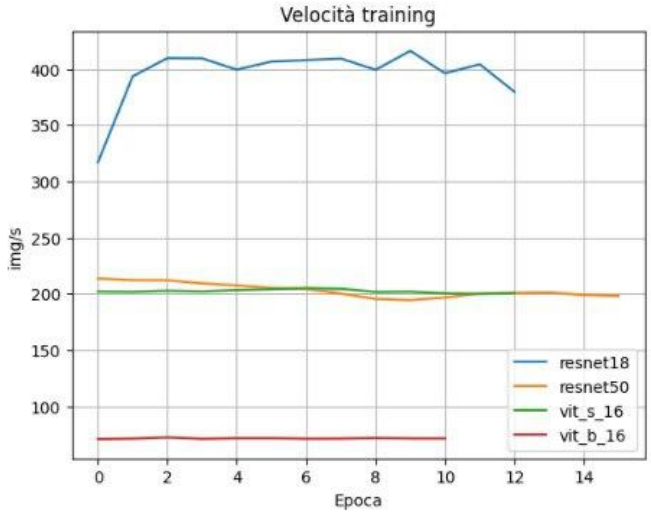
Figure 2. Training dynamics for all four models. (a) Accuracy converges between epochs 8–12 for all architectures. (b) ViT-B/16 achieves the lowest validation loss, indicating better-calibrated predictions.

5.3. Computational Efficiency

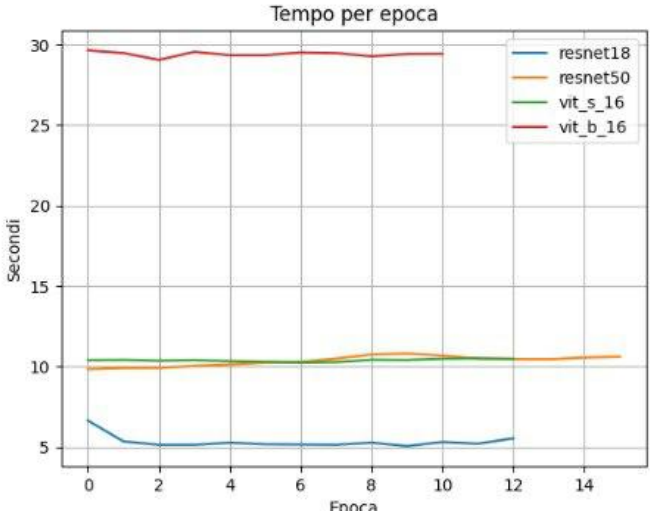
Figure 3 illustrates the training throughput and per-epoch time **under the linear probe setting**. Table 2 reports per-epoch times for all fine-tuning strategies, which increase with the number of trainable parameters (e.g., ViT-B/16 rises from 47.4 s under linear probing to 105.2 s under full fine-tuning).

ResNet-18 is by far the fastest model, approximately 2× faster than ResNet-50 and ViT-S/16 under linear probing, and over 2× faster than ViT-B/16 (Table 2).ViT-B/16 requires 47.4 s per epoch under linear probing, rising to 105.2 s under full fine-tuning, due to the quadratic complexity of self-attention with respect to the number of patches.

 doi:10.20944/preprints202604.0729.v2



(a) Training throughput (images per second).



(b) Time per training epoch (seconds).

Figure 3. Computational efficiency comparison under the initial experimental configuration. Per-epoch times for all strategies under the standardized pipeline are reported in Table 2.Throughput measurements reflect the initial experimental configuration; per-epoch times for all strategies under the standardized pipeline are reported in Table 2.

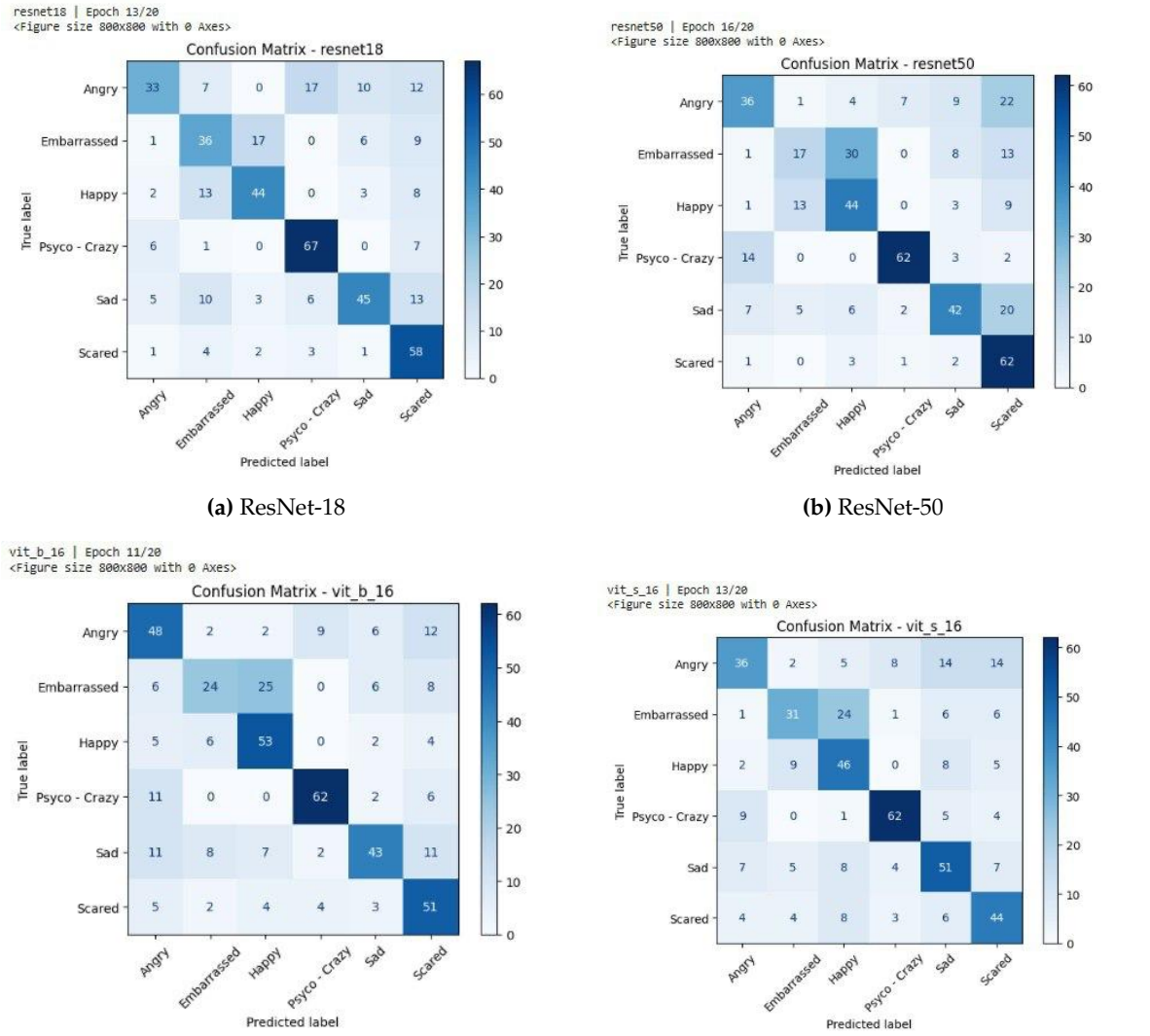
5.4. Confusion Matrix Analysis

Figure 4 presents the confusion matrices for all four models under the **linear probe** setting, computed on the held-out test set (450 images). Table 3 provides the corresponding per-class F1-scores on the same partition. Several consistent patterns emerge across all architectures.

Table 3. Per-class F1-scores: linear probe vs. best fine-tuned configuration for each model. LP = Linear Probe; Best = best fine-tuning strategy for that model (see Table 2).

	ResNet-18		ResNet-50		ViT-B/16		ViT-S/16	
Class	LP	Best	LP	Best	LP	Best	LP	Best
Angry	0.52	0.71	0.52	0.72	0.58	0.73	0.52	0.74
Embarrassed	0.51	0.66	0.32	0.70	0.43	0.75	0.52	0.66
Happy	0.65	0.80	0.56	0.76	0.66	0.80	0.57	0.73
Psycho-Crazy	0.77	0.84	0.81	0.83	0.78	0.91	0.78	0.86
Sad	0.61	0.79	0.56	0.71	0.60	0.83	0.59	0.74
Scared	0.66	0.8	0.63	0.8	0.63	0.80	0.59	0.77
Macro avg	0.62	0.76	0.57	0.75	0.61	0.80	0.60	0.75

 doi:10.20944/preprints202604.0729.v2



(a) ResNet-18 **(b)** ResNet-50

(c) ViT-B/16 **(d)** ViT-S/16

Figure 4. Confusion matrices for all four models on the held-out test set (450 images). Rows represent true labels and columns represent predicted labels. Darker cells along the diagonal indicate higher correct classification counts. All models show strong diagonal dominance for Psycho-Crazy and persistent off-diagonal confusion between Embarrassed and Happy.

5.4.1. Best-Recognized Class: Psycho-Crazy

The Psycho-Crazy category achieves the highest F1-score across all models (0.77–0.81). This class features highly distinctive visual patterns, exaggerated eyes, distorted facial proportions, and extreme expressions, that provide strong discriminative cues for both CNNs and Transformers. ResNet-50 achieves the highest F1 (0.81) for this class, with a precision of 0.86 and recall of 0.77, indicating that its deeper feature hierarchy captures the distinctive cues of this category effectively despite its weaker overall performance. All models correctly classify 62–67 out of 81 Psycho-Crazy test images, confirming that globally exaggerated visual patterns are robust to architectural choice.

5.4.2. Most Confused Pair: Happy vs. Embarrassed

The confusion between Happy and Embarrassed remains the most persistent source of error across all architectures, with a pronounced asymmetry: Embarrassed images are frequently misclassified as Happy, while the reverse occurs less often. In ResNet-50 (Figure 4b), 30 out of 69 Embarrassed images are misclassified as Happy, the single largest off-diagonal cell in any confusion matrix. ViT-B/16 (Figure 4c) shows a similar pattern, with 25 Embarrassed images predicted as Happy. In the opposite direction, ViT-B/16 misclassifies only 6 Happy images as Embarrassed, while ResNet-18 misclassifies

 doi:10.20944/preprints202604.0729.v2

13. This asymmetry suggests that features characteristic of the Embarrassed class (slight smiles, flushed cheeks, closed eyes) are a visual subset of the broader Happy category, making it easier for models to absorb Embarrassed into Happy than vice versa.

5.4.3. Confusion among Negative High-Arousal Emotions

Angry is frequently confused with Scared and Psycho-Crazy across all models, reflecting the shared visual features of high-arousal negative emotions: furrowed brows, wide eyes, and tense mouths. This pattern is particularly pronounced in ResNet-50 (Figure 4b), where 22 Angry images are misclassified as Scared. ResNet-18 (Figure 4a) shows a different confusion profile for Angry, with 17 misclassifications toward Psycho-Crazy rather than Scared, suggesting that CNNs of different depths pick up on different subsets of high-arousal features. Sad images also show notable dispersion: in ResNet-50, 20 Sad images are misclassified as Scared, while in ResNet-18, 13 are misclassified as Scared and 10 as Embarrassed.

5.5. Architecture-Specific Observations

The following observations are based on the **linear probe** results reported in Table 2. Section 5.6 extends this analysis to the fine-tuned configurations.

5.5.1. ResNet-18

Under linear probing, ResNet-18 achieves a test accuracy of 62.22% and F1-macro of 0.612. It is the most stable CNN, completing all 20 training epochs without triggering early stopping. It obtains the highest per-class F1 among CNNs for both Sad (0.61) and Scared (0.66), and is competitive on all other categories. Its lightweight architecture (11.7M parameters, 22.9 s per epoch) makes it the most efficient model for this dataset size. Under fine-tuning, ResNet-18 peaks at 75.11% with partial fine-tuning (Table 2), confirming its strong performance across training regimes.

5.5.2. ResNet-50

Under linear probing, ResNet-50 achieves the lowest test accuracy among all models (61.78%, F1-macro = 0.601). Despite greater depth, it shows higher validation loss and lower accuracy than ResNet-18, consistent with its excess capacity being poorly utilized when only the classification head is trained on 2,100 images. The Embarrassed class is particularly degraded (F1 = 0.32), with 30 out of 69 Embarrassed images misclassified as Happy (Figure 4b). However, ResNet-50 achieves the highest Psycho-Crazy F1 (0.81) among all linear probe models, suggesting that its deeper feature hierarchy captures highly distinctive patterns even when it fails on subtler distinctions. Notably, ResNet-50 recovers substantially under partial fine-tuning, reaching 74.22% (Table 2), a 12.44 percentage point improvement that indicates its poor linear probe performance reflects limited feature transferability rather than a fundamental capacity mismatch.

5.5.3. ViT-B/16

Under linear probing, ViT-B/16 achieves the highest test accuracy (64.67%) and the lowest test loss (0.972) among all models, indicating the most well-calibrated probability estimates. It obtains the highest per-class F1 for both Angry (0.58) and Happy (0.66), suggesting that global self-attention helps disambiguate categories where spatial relationships between facial features matter. Its per-class metrics are more uniform than those of ResNet-50, though the Embarrassed class remains challenging (F1 = 0.43). Its computational cost is approximately 2× that of ResNet-18 under linear probing (47.4 s vs. 22.9 s per epoch). Under fine-tuning, ViT-B/16 shows the largest improvement of any model, reaching 80.89% with progressive unfreezing, the overall best result in our experiments.

5.5.4. ViT-S/16

Under linear probing, ViT-S/16 achieves a test accuracy of 61.33% (F1-macro = 0.607) and a competitive Psycho-Crazy F1 (0.78), demonstrating that even a compact Transformer can effectively

 doi:10.20944/preprints202604.0729.v2

capture globally distinctive patterns. It obtains the highest Embarrassed F1 (0.52) among all linear probe models, though this remains modest. Its overall performance falls between ResNet-50 and the top two models, at a training cost comparable to ResNet-50 (26.4 s per epoch). Under full fine-tuning, ViT-S/16 reaches 76.22%, the second-highest result overall, though its progressive unfreezing performance (66.67%) is anomalously low, as discussed in Section 5.6.

5.6. Impact of Fine-Tuning Strategy

Figures 5 and 6 depict respectively, the Validation Loss and Validation Accuracy for each model during the Fine-Tuning strategy.

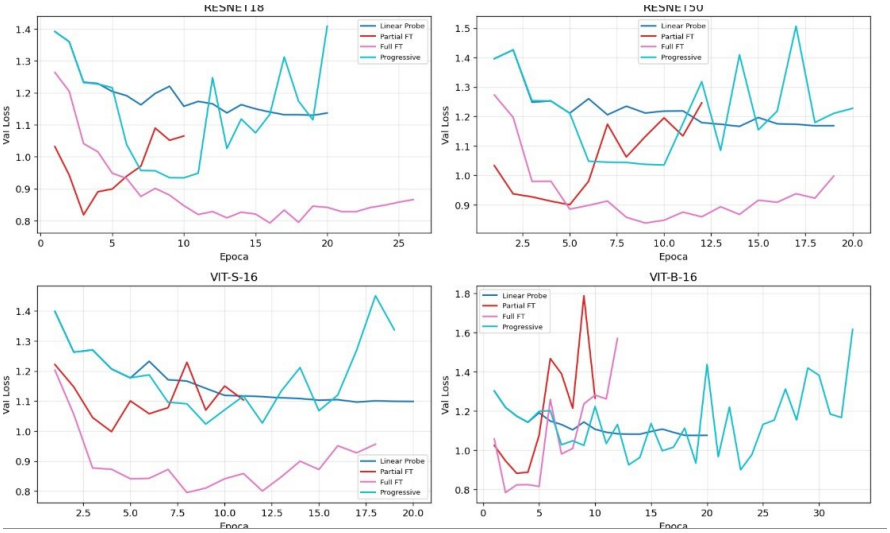


Figure 5. Validation Loss plots for all the models.

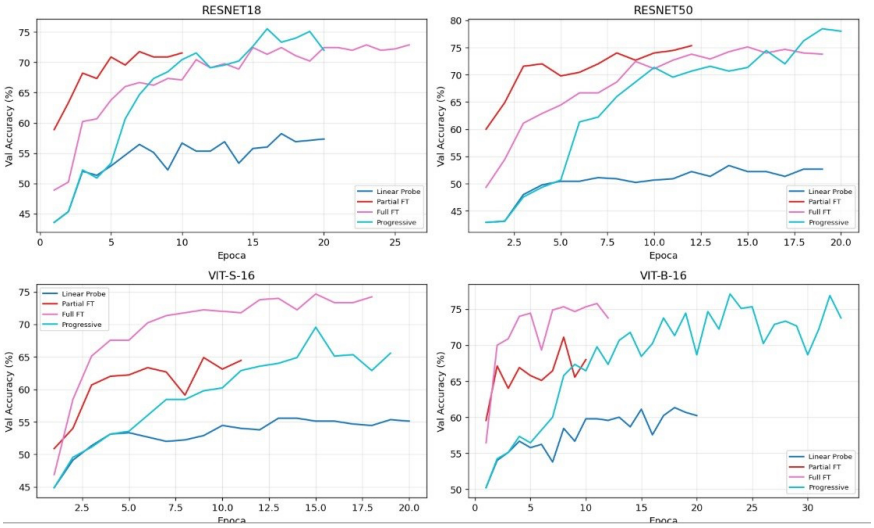


Figure 6. Validation Accuracy plots for all the models.

Figure 7 resumes the impact of the Fine-Tuning strategies on the Validation Accuracy.

Table 2 reports performance across four fine-tuning strategies for each architecture. Three key patterns are observable in the data.

Universal improvement. Every model improves substantially when backbone parameters are allowed to adapt. The smallest gain from linear probing to the best strategy for a given model is +12.44 percentage points (ResNet-50), and the largest is +16.22 percentage points (ViT-B/16). No model achieves its best performance under linear probing.

 doi:10.20944/preprints202604.0729.v2

Asymmetric gains across architectures. The two ViT models show a combined average improvement of 15.56 percentage points over their respective linear probe baselines (under their best strategy), compared to 12.67 percentage points for the two ResNets. ViT-B/16 progressive unfreezing (80.89%) outperforms all other configurations by at least 4.67 percentage points.

Strategy preference by architecture family. Both ResNets peak under partial fine-tuning, with full fine-tuning and progressive unfreezing offering no additional benefit (and marginally lower accuracy in some cases). In contrast, ViT-B/16 peaks under progressive unfreezing and ViT-S/16 under full fine-tuning. A notable anomaly is ViT-S/16 under progressive unfreezing, which achieves only 66.67% — lower than its partial fine-tuning result (66.22%) and far below its full fine-tuning peak (76.22%). This suggests sensitivity to the unfreezing schedule hyperparameters, which were held constant across all models.

Loss vs. accuracy trade-off. ViT-S/16 full fine-tuning and ViT-B/16 full fine-tuning achieve the lowest test losses (0.642 and 0.655, respectively), even though ViT-B/16 progressive unfreezing achieves a higher accuracy (80.89%) at a substantially higher loss (0.870). This divergence between loss and accuracy under progressive unfreezing warrants further investigation and may indicate that the model's probability estimates become less calibrated as more layers are gradually unfrozen.

Robustness caveat. The ViT-B/16 progressive result (80.89%) represents a single run with seed 42. Given that this configuration outperforms the next best result (ViT-S/16 full, 76.22%) by 4.67 percentage points, we caution that some portion of this gap may reflect variance due to random initialization or data ordering rather than a systematic advantage of the progressive unfreezing schedule for this architecture. Future work should verify this result across multiple seeds and report confidence intervals.

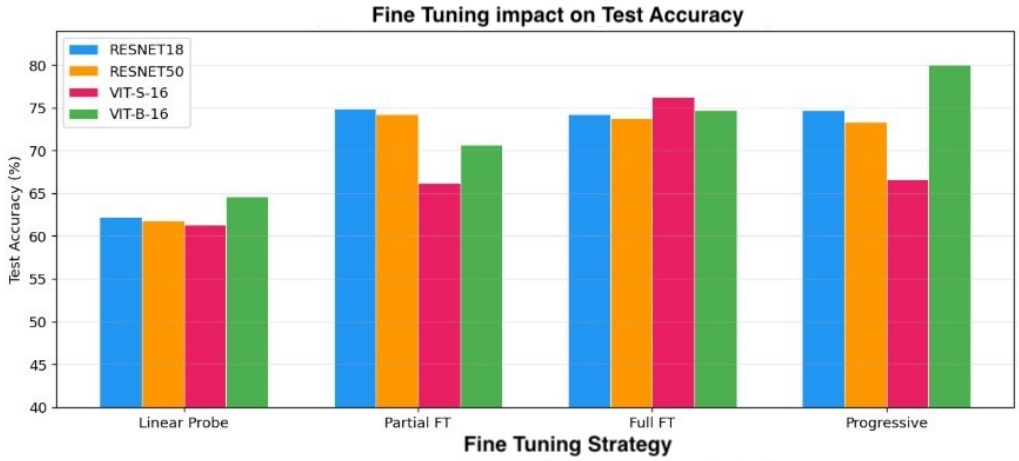


Figure 7. Fine-Tuning impact on the Validation Accuracy.

5.6.1. Confusion Patterns Under Fine-Tuning

Figure 8 presents the confusion matrices for ViT-B/16 under full fine-tuning (74.44% test accuracy) and progressive unfreezing (80.89% test accuracy). Comparing these with the linear probe confusion matrix (Figure 4) reveals how increasing fine-tuning depth reshapes error patterns.

Psycho-Crazy recognition improves monotonically with fine-tuning depth. ViT-B/16 correctly classifies 62 out of 81 Psycho-Crazy images under linear probing (recall = 0.765), rising to 67 under full fine-tuning (0.827) and 74 under progressive unfreezing (0.914). This confirms that the already-strong discriminative cues of this class are further refined when backbone features are allowed to adapt.

The Happy/Embarrassed confusion is substantially reduced but exhibits an interesting reversal. Under linear probing, 25 out of 69 Embarrassed images are misclassified as Happy (the largest single off-diagonal cell), while only 6 Happy images are misclassified as Embarrassed. Progressive unfreezing reduces the Embarrassed→Happy errors from 25 to 9, improving Embarrassed recall from 0.348 to 0.826. However, the reverse direction worsens: Happy→Embarrassed misclassifications increase from

 doi:10.20944/preprints202604.0729.v2

6 (linear probe) to 14 (progressive), slightly reducing Happy recall from 0.757 to 0.786 overall but shifting the confusion profile. This suggests that progressive unfreezing helps the model learn features specific to the Embarrassed class (e.g., flushed cheeks, averted gaze), but in doing so it occasionally over-applies these cues to Happy images that share partial visual overlap. The asymmetry of the original confusion is thus partially corrected, moving toward a more balanced—though not fully resolved—bidirectional error pattern.

Sad and Scared recognition improve markedly. Sad recall increases from 0.524 (linear probe) to 0.780 (progressive), and Scared recall from 0.739 to 0.913. These are among the largest per-class gains in the experiment, suggesting that fine-tuning is particularly effective for classes whose distinguishing features (e.g., downturned mouths for Sad, wide eyes with tense expressions for Scared) require domain-adapted intermediate representations that frozen ImageNet features do not provide.

Angry shows an uneven pattern across strategies. Angry recall improves from 0.608 (linear probe) to 0.734 under full fine-tuning, but drops back to 0.646 under progressive unfreezing. This decline is driven by a resurgence of Angry→Scared misclassifications: only 2 under full fine-tuning, but 13 under progressive unfreezing—comparable to the 12 observed under linear probing. The shared high-arousal visual features of these two categories (wide eyes, tense mouths, furrowed brows) appear to be re-entangled as progressive unfreezing adapts deeper layers, a pattern not observed for full fine-tuning where all layers are updated simultaneously from the start of training.

These results indicate that fine-tuning does not uniformly improve all classes. While overall accuracy rises from 64.67% to 80.89%, the per-class gains are uneven: Embarrassed, Sad, Scared, and Psycho-Crazy benefit substantially, Happy shows modest improvement, and Angry actually regresses under the best overall configuration. This class-level heterogeneity reinforces the importance of reporting per-class metrics alongside aggregate accuracy, and suggests that the residual errors at 80% accuracy reflect a combination of genuine visual ambiguity in the illustrated domain and annotation subjectivity from single-annotator labeling, rather than insufficient model capacity.

These patterns are analyzed in detail in Section 6.4.

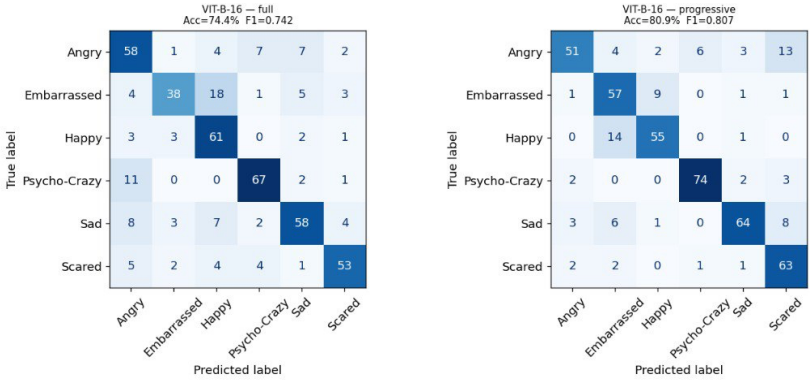


Figure 8. Confusion matrix for ViT-B/16 with full and progressive unfreezing on the held-out test set (450 images). Compared to the linear probe confusion matrices (Figure 4).

6. Discussion

6.1. CNN vs. Transformer on Small Non-Photorealistic Datasets

The interaction between architecture and fine-tuning depth produces a nuanced picture of CNN–Transformer trade-offs on small illustrated datasets.

Under linear probing, the advantage of Vision Transformers over CNNs is marginal: all four models fall within a 3.3 percentage point band (61.33%–64.67%). The strong inductive bias of CNNs, locality and translation equivariance, proves beneficial when no domain adaptation of the backbone is permitted, as it reduces the space of learnable functions and partially compensates for the domain gap between ImageNet and anime/manga imagery.

 doi:10.20944/preprints202604.0729.v2

This picture changes substantially under fine-tuning. When backbone parameters are allowed to adapt, ViT-B/16 with progressive unfreezing reaches 80.89%, outperforming the best CNN configuration (ResNet-18 partial, 75.11%) by 5.78 percentage points. ViT-S/16 with full fine-tuning (76.22%) also surpasses both ResNets. The average gain from linear probing to the best strategy is 15.56 pp for ViTs versus 12.67 pp for CNNs, indicating that Transformers benefit disproportionately from domain adaptation. This is consistent with the view that, lacking spatial inductive biases, Transformers are more dependent on, and more responsive to, fine-tuning of intermediate representations.

A complementary finding concerns probability calibration. ViT-B/16 consistently achieves the lowest test loss under full fine-tuning (0.655), even though its highest accuracy is obtained under progressive unfreezing at a higher loss (0.870). This divergence suggests that progressive unfreezing trades calibration for discriminative accuracy, a trade-off that may matter in downstream applications where confidence scores are used for human–AI collaborative annotation or rejection thresholds.

6.2. The Role of Dataset Size and Domain

Under linear probing, test accuracies of 61–65% fall substantially below the 70–85% typically achieved on photographic FER benchmarks such as FER-2013 and RAF-DB, reflecting the inherent difficulty of the illustrated domain. Fine-tuning narrows this gap considerably: the best configuration (ViT-B/16 progressive, 80.89%) enters the lower range of photographic benchmark performance, demonstrating that modern architectures can achieve competitive accuracy on anime and manga faces when appropriate transfer learning strategies are applied.

A direct comparison with the two prior studies most closely related to this work further contextualises our results. Hill [15] reported approximately 80% accuracy on a three-class emotion recognition task (happy, surprised, angry) applied to Tom and Jerry cartoon frames using a custom CNN trained from scratch on 4,800 images. Our best configuration (ViT-B/16 with progressive unfreezing) achieves 80.89% on a substantially harder problem: six emotion categories, a more visually diverse illustrated domain spanning both anime and manga styles, and a larger class set that includes deliberately ambiguous categories such as Embarrassed and Psycho-Crazy that have no equivalent in Hill's benchmark. Matching Hill's accuracy on a more complex task, and with architectures that generalise rather than specialise, represents a meaningful advance. Köklü [16] did not report a comparable aggregate accuracy figure for their manga facial expression classifier, precluding a direct numerical comparison; however, their reliance on a custom architecture trained without pre-trained weights on a five-class dataset places it methodologically below even our linear probe baselines (61–65%), which use frozen ImageNet features without any domain adaptation. More broadly, Parrillo et al. [7] demonstrated 95.0% cross-validated accuracy on a four-class comic character *identity* recognition task using a frozen ResNet-50 backbone, a figure that is not directly comparable to ours because character identity and facial expression classification involve qualitatively different discriminative features. Nonetheless, the fact that both studies converge on the conclusion that pretrained ResNet-50 features transfer effectively to stylized illustrated domains, and that fine-tuning depth is the primary performance determinant, provides independent corroboration of the findings reported here.

A residual gap nonetheless persists, attributable to several domain-specific factors. Manga and anime expressions do not follow anatomical constraints; they are stylistically diverse across artists and series, and they often employ visual metaphors, steam emanating from the head for anger, heart-shaped eyes for love, spiral eyes for dizziness, that are difficult for pixel-level models to interpret without domain-specific priors. Furthermore, the subjectivity of emotion categories (Section 2.3) is amplified in illustrated media, where the same expression may be drawn in radically different styles depending on the artist's conventions. The persistent Happy/Embarrassed confusion observed even at 80% overall accuracy (Section 5.6.1) likely reflects this combination of visual ambiguity and annotation subjectivity rather than a limitation of model capacity.

Expanding the dataset beyond 3,000 images, particularly with greater diversity of artistic styles and multiple annotators, remains the most promising avenue for closing the remaining gap with photographic benchmarks.

 doi:10.20944/preprints202604.0729.v2

6.3. Implications for Model Selection

The results suggest a two-tier recommendation depending on deployment constraints.

When training time and computational resources are limited, ResNet-18 with partial fine-tuning offers the strongest speed–accuracy trade-off: 75.11% test accuracy ($F1_{macro}$ = 0.748) at 22.7 s per epoch, with only 75.1% of parameters trainable. Its shallow architecture is resistant to overfitting on small datasets, and partial fine-tuning recovers nearly all of the accuracy gains available from deeper strategies (full fine-tuning yields 74.67%, i.e., no additional benefit). ResNet-50 with partial fine-tuning achieves comparable accuracy (74.22%) but at 1.3× the training cost, making it a less efficient alternative.

When accuracy is the primary objective and computational cost is acceptable, ViT-B/16 with progressive unfreezing achieves the highest accuracy in our experiments (80.89%, $F1_{macro}$ = 0.807), outperforming the best CNN by 5.78 percentage points. This configuration requires approximately 3.5× longer training time per epoch than ResNet-18 partial (79.9 s vs. 22.7 s), but the absolute training time remains modest given the small dataset size (under 40 minutes total on a single GPU). For applications where well-calibrated probability estimates matter—such as human–AI collaborative annotation or confidence-based rejection—ViT-B/16 under full fine-tuning may be preferable despite its lower accuracy (74.44%), as it achieves the lowest test loss in the experiment (0.655).

ViT-S/16 occupies a middle ground: its full fine-tuning result (76.22%) exceeds both ResNets at a training cost (43.8 s per epoch) substantially below that of ViT-B/16. It may therefore be the preferred choice when a moderate accuracy improvement over CNNs is desired without incurring the full computational overhead of the base Transformer.

6.4. Fine-Tuning Depth and Architectural Sensitivity

The comparison across fine-tuning strategies (Table 2) reveals three key findings regarding the interaction between architecture and transfer learning depth.

Finding 1: Fine-tuning strategy matters more than architecture. Under linear probing, all four models cluster within a narrow 3.3 percentage point band (61.33%–64.67%). Under the best fine-tuning strategy for each model, this band widens to 6.67 percentage points (74.22%–80.89%), and every model improves by at least 12 percentage points over its linear probe baseline. The gap between the worst and best configurations across the entire experiment matrix, ResNet-50 linear probe (61.78%) versus ViT-B/16 progressive (80.89%), is 19.11 percentage points, far larger than any inter-architecture difference within a single strategy. This demonstrates that the choice of fine-tuning regime has a greater impact on performance than the choice of architecture for this task and dataset size.

Finding 2: Vision Transformers benefit more from fine-tuning than CNNs. The two ViT models show an average improvement of 15.56 percentage points from linear probing to their respective best strategies (ViT-B/16: +16.22 pp via progressive unfreezing; ViT-S/16: +14.89 pp via full fine-tuning), compared to an average of 12.67 percentage points for the two ResNets (ResNet-18: +12.89 pp via partial; ResNet-50: +12.44 pp via partial). This is consistent with the hypothesis that Transformers' lack of spatial inductive bias makes them more dependent on domain-specific adaptation of intermediate representations. Under linear probing, the CNN's built-in locality and translation equivariance provide a stronger prior for the illustrated domain, partially compensating for the domain gap between ImageNet and anime/manga imagery. When backbone parameters are allowed to adapt, the Transformer's greater representational flexibility enables it to learn domain-specific spatial relationships that frozen features could not capture.

Finding 3: CNNs and ViTs favor different fine-tuning regimes. Both ResNet models achieve their best performance under partial fine-tuning (ResNet-18: 75.11%; ResNet-50: 74.22%), with full fine-tuning offering no additional benefit and in some cases slightly degrading performance. This suggests that for CNNs on this dataset size, unfreezing only the last residual block provides sufficient domain adaptation while the frozen early layers act as an effective regularizer.

 doi:10.20944/preprints202604.0729.v2

In contrast, ViT-B/16 achieves its best result under progressive unfreezing (80.89%), substantially outperforming both its partial (70.67%) and full (74.44%) fine-tuning configurations. The progressive schedule, which gradually exposes deeper layers to gradient updates, appears to provide a favorable regularization–adaptation trade-off for the large Transformer on this small dataset. ViT-S/16 instead peaks under full fine-tuning (76.22%), while its progressive result (66.67%) is anomalously low. This discrepancy may reflect sensitivity to the unfreezing schedule hyperparameters (e.g., the number of epochs between unfreezing steps), which were kept constant across all models rather than tuned per-architecture. Future work should investigate architecture-specific unfreezing schedules.

ResNet-50 recovery. A notable result is the recovery of ResNet-50, which was the weakest model under linear probing (61.78%) but becomes competitive under partial fine-tuning (74.22%, F1 = 0.74). This indicates that its poor frozen-backbone performance was primarily caused by the limited transferability of its deeper features to the illustrated domain, rather than by a fundamental capacity mismatch. Partial fine-tuning allows the final bottleneck block to adapt its high-level representations, recovering the model's capacity advantage while the frozen early layers prevent overfitting.

Architectural ranking shift. Under linear probing, ResNet-18 (62.22%) outperforms both ViT-S/16 (61.33%) and ResNet-50 (61.78%), and is within 2.5 percentage points of the best model (ViT-B/16, 64.67%). Under the best fine-tuning strategy for each model, the ranking shifts: ViT-B/16 (80.89%) leads by a substantial margin, followed by ViT-S/16 (76.22%), ResNet-18 (75.11%), and ResNet-50 (74.22%). This reversal confirms that architectural comparisons based solely on frozen-backbone evaluation can be misleading, and underscores the importance of evaluating multiple transfer learning strategies when comparing deep learning architectures.

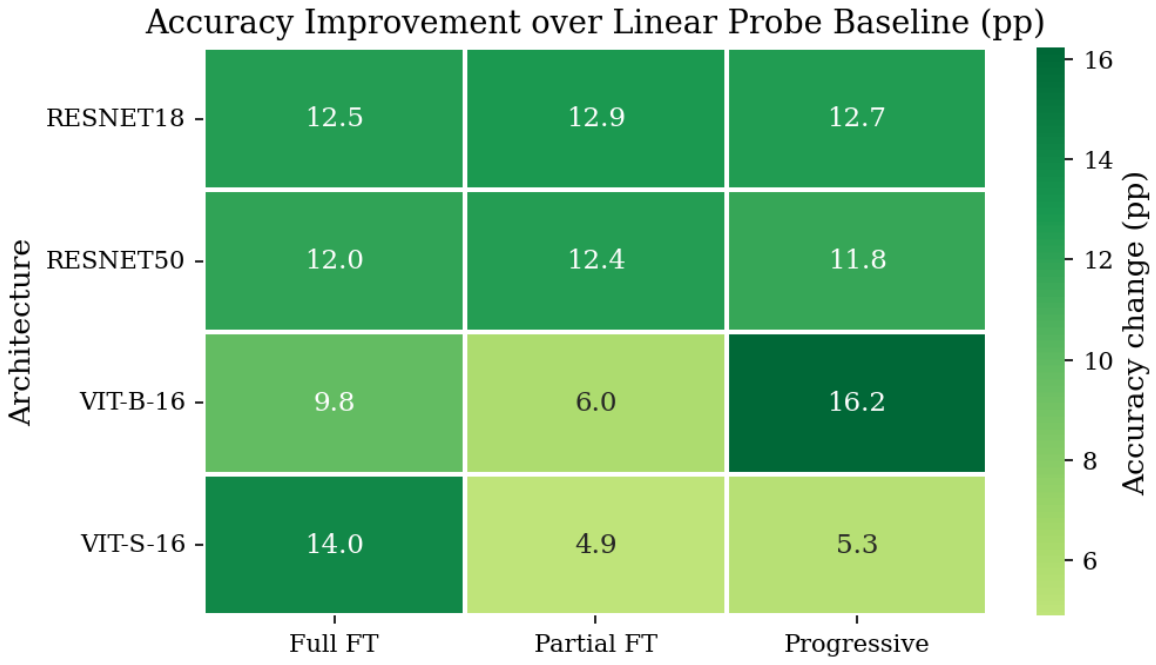


Figure 9. Heatmap describing the Accuracy Improvement over Linear Probe baseline.

6.5. Limitations

This study has several limitations. First, the dataset, while large for the anime/manga FER domain, is small by deep learning standards; the three-way split further reduces the training partition to 2,100 images, which may not be sufficient for data-hungry architectures such as Vision Transformers to reach their full potential. Second, annotation was performed by a single annotator, introducing potential subjective bias, particularly for visually ambiguous categories such as Embarrassed and Happy, where inter-annotator disagreement is likely. No inter-annotator agreement metric (e.g., Cohen's kappa) was computed, so the extent of label noise in the ground truth cannot be quantified. Third, the use of Google Colab's free tier imposed hardware constraints that limited hyperparameter search; in

 doi:10.20944/preprints202604.0729.v2

particular, no systematic grid or random search was conducted over learning rates, batch sizes, or regularization strategies. Fourth, while we evaluated four fine-tuning strategies, the hyperparameters governing each strategy (e.g., the unfreezing schedule for progressive fine-tuning, the number of unfrozen blocks for partial fine-tuning) were held constant across architectures rather than tuned per-model. The anomalous performance of ViT-S/16 under progressive unfreezing (66.67%, compared to 76.22% under full fine-tuning) suggests that architecture-specific tuning of these parameters could yield different results. Fifth, the test-set accuracies were on average 4.2 percentage points higher than the corresponding validation accuracies, an unusual pattern that may reflect an uneven difficulty distribution across the stratified partitions; future work should investigate this through repeated random splits or cross-validation.

7. Conclusions and Future Work

We presented a systematic comparison of CNN (ResNet-18, ResNet-50) and Vision Transformer (ViT-B/16, ViT-S/16) architectures for facial expression recognition in anime and manga characters, evaluated across four transfer learning strategies: linear probing, partial fine-tuning, full fine-tuning, and progressive unfreezing. Using a balanced dataset of 3,000 images across six emotion categories, with a stratified train/validation/test split ensuring unbiased final evaluation, we demonstrated that:

1. **Fine-tuning strategy is the dominant performance factor.** Under linear probing, all models cluster between 61% and 65% test accuracy. With appropriate fine-tuning, every model exceeds 73%, and the best configuration (ViT-B/16 with progressive unfreezing) reaches 80.89%. The 19-percentage-point gap between the worst and best configurations across the full experiment matrix exceeds any inter-architecture difference within a single strategy, establishing fine-tuning depth as the primary determinant of performance on this task.
2. **Vision Transformers benefit disproportionately from fine-tuning.** ViTs improve by an average of 15.56 percentage points from linear probing to their best strategy, compared to 12.67 percentage points for CNNs. Under fine-tuning, ViT-B/16 achieves the highest accuracy overall (80.89%) and ViT-S/16 the second highest (76.22%), reversing the near-parity observed under linear probing. This confirms that frozen-backbone evaluations underestimate the potential of Transformers in domain-shifted settings.
3. **CNNs and ViTs favor different adaptation regimes.** ResNets achieve their best performance under partial fine-tuning, where the frozen early layers act as an effective regularizer. ViTs require deeper adaptation, full fine-tuning or progressive unfreezing, to realize their representational advantage.
4. **Highly distinctive emotions are robustly recognized across all configurations**, with Psycho-Crazy achieving F1 scores of 0.77–0.91 regardless of architecture or fine-tuning strategy. In contrast, visually similar emotions (Happy/Embarrassed) remain challenging even at higher overall accuracy levels, suggesting that label ambiguity, rather than model capacity, is the binding constraint for these categories.

These results carry a practical implication: for resource-constrained deployment with small non-photorealistic datasets, ResNet-18 with partial fine-tuning offers a strong speed–accuracy trade-off (75.11% accuracy at 22.7 s per epoch). When computational cost is less constrained, ViT-B/16 with progressive unfreezing achieves substantially higher accuracy (80.89%) but requires approximately 3.5× longer training time per epoch.

Future work should focus on: (1) expanding the dataset to 10,000+ images to further leverage the capacity advantage of Transformers; (2) introducing inter-annotator agreement metrics to quantify label noise, particularly for the Embarrassed and Happy categories; (3) incorporating object detection frameworks for joint face localization and emotion classification; (4) investigating architecture-specific unfreezing schedules, given the sensitivity observed for ViT-S/16; and (5) extending the approach to mixed-domain datasets combining manga, anime, and Western cartoon styles.

 doi:10.20944/preprints202604.0729.v2

References

1. W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophysics*, vol. 5, pp. 115–133, 1943.
2. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.
3. P. Ekman, *Emotion in the Human Face*. Cambridge University Press, 1982.
4. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
5. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, pp. 886–893.
6. T. Ojala, M. Pietikäinen, and M. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. PAMI*, vol. 24, no. 7, pp. 971–987, 2002.
7. M. Parrillo, L. Laura, and A. Manna, "Transfer learning strategies for comic character recognition in low-data regimes: A comparative study," *Future Internet*, vol. 18, no. 4, p. 192, 2026. https://doi.org/10.3390/fi18040192
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.
9. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
10. A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
11. A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
12. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
13. S. Li *et al.*, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Processing*, vol. 26, no. 4, pp. 1943–1956, 2017.
14. I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2013.
15. J. W. Hill, "Deep learning for emotion recognition in cartoons," M.S. thesis, Univ. of Lincoln, 2017.
16. M. Köklü, "Manga facial expression classification using convolutional neural networks," PyTorch-based academic project, 2020.
17. L. F. Barrett *et al.*, "The theory of constructed emotion: An active inference account of interoception and categorization," *Social Cognitive and Affective Neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.
18. J. A. Russell, "A circumplex model of affect," *J. Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
19. Roboflow Inc., "Roboflow: Dataset management and preprocessing for computer vision," [Online]. Available: https://roboflow.com/
20. S. Evan, "Face of Pixiv top daily illustration 2018," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/stevenevan99/face-of-pixiv-top-daily-illustration-2018
21. G. Camerlingo, P. Fantozzi, L. Laura, and M. Parrillo, "Teaching neural networks using comic strips," in Proc. MIS4TEL 2024, Lecture Notes in Networks and Systems, vol. 1171, Springer, Cham, 2024, pp. 1–10.
22. S. Elia, M. Parrillo, "Code GitHub Repository," [Online]. Available: https://github.com/eliasantorodevengineer-dotcom/Facial_Expression_Recognition_in_Anime_and_Manga