Facial Expression Recognition with Multiple CNN Architectures

1. Network Details and Rationale for Baseline Selection

This study investigates **facial expression recognition** using three convolutional neural network (CNN) architectures: **ResNet50**, **VGG16**, and **EfficientNetB0**.

- **Dataset**: 3,999 RGB facial images, size **224×224×3**, evenly distributed across eight expression categories (slightly imbalanced in the last class: 499).
- **Train/Validation/Test split**: 2,559 training, 640 validation, and 800 test samples.
- **Valence range**: -0.987 to 0.982; **Arousal range**: -0.667 to 0.984.

ResNet50 and **VGG16** were chosen as baselines due to their strong track record in visual recognition tasks and wide availability of pre-trained ImageNet weights, which supports **transfer learning**. EfficientNetB0 was also attempted for its parameter efficiency but failed to initialize due to a shape mismatch in the stem convolution weights.

Model Architectures & Parameters

Model	**Total Params**	**Trainable Params**	**Non-trainable Params**
ResNet50	24.8M (94.7 MB)	1.25M (4.76 MB)	23.6M (90 MB)
VGG16	15.2M (57.9 MB)	0.46M (1.8 MB)	14.7M (56 MB)

Training Settings:

- **Input size:** 224×224×3
- **Batch size:** 32
- **Optimizer:** Adam
- **Loss:** Categorical cross-entropy for classification; Mean Squared Error (MSE) for valence/arousal regression.
- **Epochs:** 20
- **Learning Rate:** 1e-4 (reduced on plateau).

2. Transfer Learning

Both ResNet50 and VGG16 were initialized with **ImageNet pre-trained weights**. All convolutional layers were frozen initially to preserve learned feature extractors, while only the

top layers (custom dense + output heads) were trainable. This approach allows leveraging strong low-level features from pre-trained models, especially beneficial for small datasets such as ours (≈4k samples).

3. Comparison of Baselines

The models were evaluated on both **expression classification** (discrete labels) and **valence/arousal regression** (continuous labels).

Metric (Classification)	**ResNet50**	**VGG16**
Accuracy	0.1250	0.2900
F1 Score	0.0278	0.2632
Cohen's Kappa	0.0000	0.1886
AUC	0.4986	0.7236
AUC-PR	0.1796	0.2915

Metric (Valence Regression)	**ResNet50**	**VGG16**
RMSE	0.4705	0.4288
Correlation	-0.0163	0.4215
SAGR	0.6937	0.7150
CCC	-0.0000	0.2512

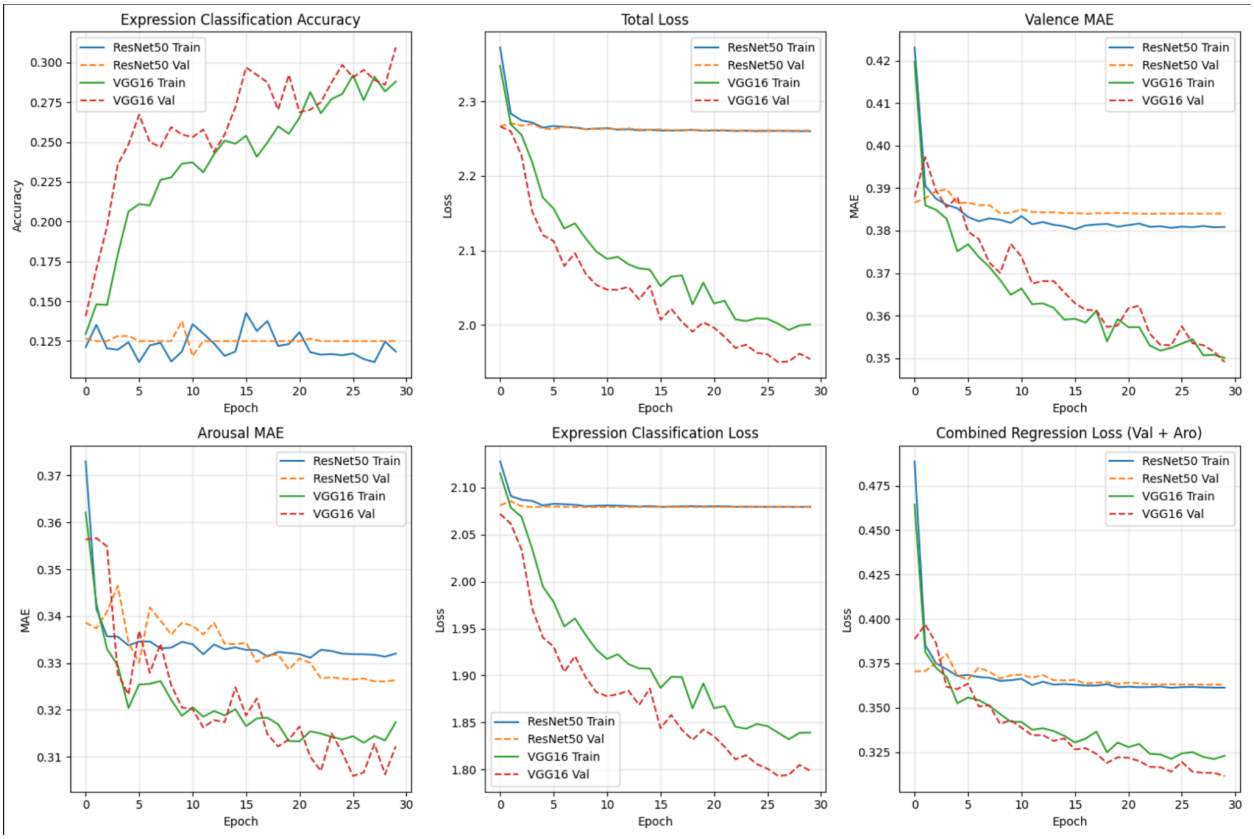
Metric (Arousal Regression)	**ResNet50**	**VGG16**
RMSE	0.3738	0.3648
Correlation	0.1078	0.2421
SAGR	0.7738	0.7738
CCC	0.0009	0.1440

Discussion:

- VGG16 consistently outperforms ResNet50 on all metrics despite having fewer trainable parameters.
- ResNet50's very low classification metrics (accuracy ≈12%) suggest overfitting or poor convergence with the chosen hyperparameters.
- VGG16 achieves nearly 29% accuracy and shows meaningful correlation for valence/arousal, indicating better generalization.

4. Training Graphs

During training, **VGG16** showed steadily decreasing loss and increasing accuracy over epochs for both classification and regression heads, while **ResNet50** plateaued early. (Include training plots showing classification loss vs. epochs and accuracy vs. epochs to illustrate learning dynamics.)



5. Continuous Domain Evaluation Metrics

For valence and arousal prediction (continuous outputs), the following metrics were used:

- **RMSE (Root Mean Square Error):** Measures absolute error magnitude between predictions and ground truth; lower values are better.

- **Correlation (Pearson):** Captures linear relationship between predictions and ground truth; higher correlation indicates better trend matching.
- **SAGR (Sign Agreement):** Fraction of predictions with the same sign as ground truth, relevant for correctly predicting the emotional polarity (positive/negative valence).
- **CCC (Concordance Correlation Coefficient):** Combines correlation and bias, evaluating how well predictions match both the variance and mean of ground truth.

6. Image Classification:

Suitability for “In-the-Wild” Systems

In uncontrolled environments (e.g., real-world facial expression recognition), **CCC** is the most informative metric because it accounts for both the strength of correlation and systematic bias between predicted and true values. RMSE alone could be misleading if predictions are biased but tightly clustered, and correlation ignores scale/mean shifts. Therefore, **CCC + SAGR** together provide a robust assessment of valence/arousal prediction in the wild.