Face emotion recognition by image classification

Project Report for the Course on Machine Learning and Deep Learning (650.025, 25 W), University of Klagenfurt

Albin Morisseau (12532283)
almorisseau@edu.aau.at
07.02.2026, Klagenfurt, Austria

***Abstract*—Facial emotion recognition (FER) aims to automatically infer affective state from facial images and is a key component for next-generation human–computer interaction, mental health support, and safety-critical monitoring systems.**

This work presents a deep learning-based FER pipeline combining robust face preprocessing with transfer learning on modern convolutional architectures. A heterogeneous dataset is constructed by harmonizing three expert-annotated sources (RAF-DB, AffectNet subset, and KDEF) into a unified seven-class label space, followed by face detection, alignment, anomaly filtering, and class-balanced splitting. Several ImageNet- and VGGFace2-pretrained models are evaluated, with particular emphasis on ConvNeXt-Tiny, ResNet-50, and a VGGFace2-based architecture. On an independent test set of 4,515 images, the best ConvNeXt-Tiny configuration, fully fine-tuned on the compiled dataset, achieves 86.6% accuracy and a macro F1-score of 0.853, outperforming the ResNet-50 and VGGFace2-based baselines and remaining competitive with recent state-of-the-art results reported on individual benchmarks.

Grad-CAM analyses show that the model focuses on semantically relevant facial regions, and real-time webcam experiments demonstrate stable predictions under realistic conditions. Ethical implications related to privacy, consent, and fairness are discussed, framing the system as a responsible proof of concept rather than a deployable solution.

I. Introduction

Emotions are a fundamental aspect of human experience and provide crucial insights into a person's mental and affective state. Beyond being occasionally confused with temperament or personality traits, emotions play a central and universal role in social interaction, communication, and overall well-being.

In today's society, where digital interfaces increasingly mediate human relationships, enabling machines to accurately recognize and interpret facial emotions opens up important opportunities for enhancing human–computer interaction, supporting mental health monitoring, enabling adaptive learning technologies, and improving safety systems such as driver monitoring [1]–[4]. Facial emotion recognition (FER) thus emerges as a promising yet challenging research area.

The theoretical grounding of most categorical FER systems can be traced back to the work of Ekman, who proposed that a small set of *basic* emotions are universally expressed and recognized across cultures through specific facial configurations. In its canonical formulation, Ekman's model comprises six primary emotions: anger, fear, disgust, surprise, happiness, and sadness [5], [6].

These categories are associated with distinct patterns of facial muscle activation (Action Units) and have been extensively used to annotate facial expression datasets, providing a practical label space for supervised learning. In this project, we adopt this seven-class taxonomy (we added neutral) as our target label space, which aligns with many widely used FER benchmarks.

Fig. 1: The six basic emotions according to Ekman.

The core hypothesis of this work is that, thanks to recent progress in deep learning, it is now feasible to detect and classify emotions from facial images with a level of performance that is sufficient for real-world assistance and interaction, at least in controlled conditions [2], [4]. A critical technical challenge arises at the very first step: before any emotion can be analyzed, the relevant region of the image: the face, must be accurately detected. Face detection is thus a prerequisite for emotion recognition, as it allows models to focus on the true region of interest, reduces noise and false predictions, and ensures that only meaningful features are extracted for classification. Without this step, confusion and errors in emotion recognition increase drastically, especially in complex, cluttered real-world images.

However, deploying facial emotion recognition also raises profound ethical questions. Because these systems analyze highly personal biometric data, they involve risks related to privacy (collection, storage and potential reuse of facial information), consent (explicit agreement on data collection and usage) and fairness (the risk of biased performance or systematic misclassifications across demographic groups). Several recent ethical reviews further highlight concerns regarding misuse in surveillance, manipulation or discrimination,

and call for strict safeguards and governance frameworks for emotion recognition technologies [1], [7]. In this project, experiments are conducted within strict academic boundaries, with an emphasis on transparency, data minimization and fairness. The system is conceived primarily as a responsible proof of concept rather than as a deployable product.

The goal of this project is to design, train, and evaluate a deep learning–based system capable of recognizing emotions from facial images with sufficient accuracy for practical applications, while explicitly acknowledging its limitations and ethical constraints. More specifically, the work investigates a complete pipeline involving face detection, face preprocessing, and emotion classification using convolutional neural networks and evaluates its quantitative performance in realistic scenarios.

II. Related Work

The application of deep learning to facial emotion recognition (FER) has been extensively studied in recent years. Most supervised approaches formulate FER as a multi-class image classification task, where the goal is to predict one of a set of basic emotions from facial images [8]–[10].

Early CNN-based systems such as FERC [10], which use a two-stage CNN architecture where a first network removes background information and a second network focuses on extracting facial feature vectors, already achieved up to 96% accuracy on controlled datasets (e.g., Extended Cohn–Kanade, Caltech Faces), demonstrating that convolutional networks can automatically learn discriminative facial features without manual feature engineering.

Building on these early results, subsequent work focused on in-the-wild conditions and transfer learning. Bargal et al. [8] combine deep CNN features with a large-scale in-the-wild facial emotion dataset, using ImageNet-pretrained models that are fine-tuned for emotion classification. They report significant improvements over training from scratch, particularly for classes with fewer samples, illustrating the effectiveness of transfer learning for real-world FER. Khorrami et al. [9] explore architectures that emphasize local facial regions such as the eyes and mouth via specialized branches, and show that focusing on these subregions improves recognition of subtle or easily confused expressions, while also revealing that certain neurons implicitly learn to detect specific Facial Action Units.

More recent surveys and benchmarks on large in-the-wild datasets such as RAF-DB and AffectNet confirm that modern CNN backbones (e.g., ResNet variants) remain strong baselines, while transformer-based architectures (e.g., Vision Transformers) currently reach state-of-the-art performance, with weighted average recall around 92–93% on RAF-DB but only about 65–67% on AffectNet, reflecting the increased difficulty of noisy, real-world data. These studies also emphasize the importance of robust preprocessing (face detection and alignment, standardized resizing), careful label harmonization, and data augmentation for cross-dataset generalization.

Overall, existing research confirms the feasibility of deep learning–based facial emotion recognition and provides clear guidance on preprocessing, model selection, and training strategies. The approach proposed in this project follows these best practices while focusing specifically on the challenges of combining multiple datasets and seeking robust performance under realistic, in-the-wild conditions.

III. Data Collection and Data Preprocessing

A. Dataset Sources

The facial emotion recognition dataset used in this study is an aggregation of three publicly available, expert-annotated collections: RAF-DB [11], a subset of AffectNet [12] and KDEF [13]. Combining these sources introduces a wide range of recording conditions, subject demographics and expression intensities, which is essential for training robust deep learning models.

- **RAF-DB.** The Real-world Affective Faces Database (RAF-DB) contains 15,339 in-the-wild facial images collected from the web. Each image is annotated with basic emotion labels and exhibits large variability in subject age, gender, ethnicity, head pose, lighting and accessories, making RAF-DB a challenging and realistic benchmark for FER.
- **AffectNet (subset).** A subset of AffectNet is used to further increase the diversity of in-the-wild samples. It contributes 25,262 annotated facial images with a broader range of emotion intensity and valence. This dataset introduces more ambiguous and subtle expressions, which are important for assessing the robustness of models beyond clearly prototypical faces.
- **KDEF.** The Karolinska Directed Emotional Faces (KDEF) dataset provides 4,900 high-control images captured under studio conditions. Subjects are photographed with standardized poses, lighting, backgrounds, and obvious confounding factors such as beards, mustaches, earrings and eyeglasses are excluded. KDEF thus offers clean, well-controlled examples of prototypical facial expressions.

By leveraging around 45,000 images from both in-the-wild and controlled recording conditions, the combined dataset is expected to improve model generalization performance compared to training on any single source alone. Finally, a key underlying assumption of this study is the correctness of the expert annotations provided with the original datasets.

B. Data Preprocessing and Partitioning

1) Data aggregation and Label Normalization: The three original datasets use different label conventions and structural organizations. To enable joint training, all annotations were mapped to a common set of seven basic emotions: **surprised, fear, disgust, happy, sad, angry** and **neutral**. Each emotion was associated with a stable integer identifier between 0 and 6, defining a canonical label space used throughout the project.

- **RAF-DB**: Labels provided in the original CSV files and folder structure were converted into this shared label space.

- **AffectNet**: Only images whose YOLO-format annotations could be mapped to one of the seven target emotions were retained.
- **KDEF**: Emotion labels were extracted directly from the image filenames using the standardized KDEF emotion codes (e.g., `SU`, `AF`, `HA`). Only images whose emotion codes matched the predefined target classes were retained. In addition, pose information encoded in the filenames was used to filter the dataset, keeping only near-frontal views (straight, half-left and half-right) while discarding images with more extreme profile angles.

This harmonization ensures that the same semantic emotion category is treated consistently across all three datasets, which is essential when learning from heterogeneous sources.

2) Splitting of data: After label mapping, the data were then divided into training, validation and test sets using a stratified 80/10/10 split, so that the class and origin datasets distribution remains comparable across splits. A fixed random seed was used during shuffling to guarantee reproducibility of the experiments.

TABLE I: Distribution of the sources dataset after cleaning and splitting

Source	Train (80%)	Val (10%)	Test (10%)	Total
AffectNet	18,612	2,695	2,690	23,997
KDEF	1,409	274	285	1,968
RAF-DB	12,269	1,530	1,540	15,339
Total	**32,290**	**4,499**	**4,515**	**41,304**

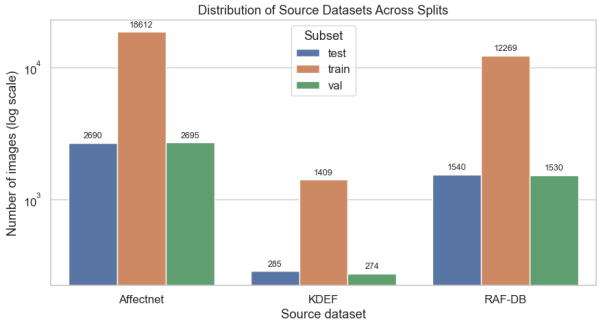

Fig. 2: Distribution of the emotion classes after cleaning and splitting.

TABLE II: Distribution of the dataset after cleaning and splitting by emotion.

Class	Train (80%)	Val (10%)	Test (10%)	Total
Angry	3,720	512	515	4,747
Disgust	3,576	503	504	4,583
Fear	2,877	384	387	3,648
Happy	7,796	1,070	1,071	9,937
Neutral	4,885	714	717	6,316
Sad	4,489	614	616	5,719
Surprised	4,947	702	705	6,354
Total	**32,290**	**4,499**	**4,515**	**41,304**

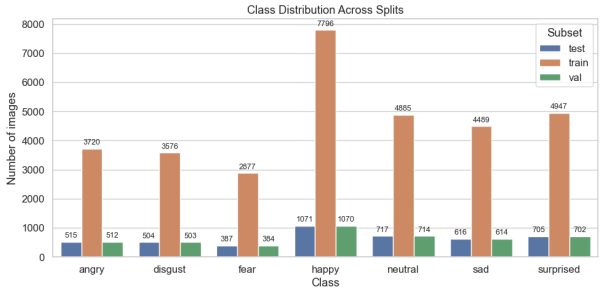

Fig. 3: Class distribution across the training, validation, and test splits.

3) Face detection and alignment: Raw facial images collected from heterogeneous sources often exhibit substantial variability in background content, head pose and facial orientation. Such factors may introduce spurious correlations and distract learning algorithms from expression-relevant facial cues. To mitigate these sources of variability, a face detection and alignment preprocessing stage was applied to all samples prior to training.

1) **Face Detection and Landmark Localization:** Each image was processed using the Multi-task Cascaded Convolutional Neural Network (MTCNN) [14], which jointly performs face detection and facial landmark localization. For each detected face, the corresponding bounding box and five key landmarks, including the left and right eye centers, were extracted.
2) **Pose Normalization via Eye Alignment:** To reduce in-plane rotation and head pose variability, the angle between the left and right eye landmarks was computed. A rigid rotation was then applied to the image such that the eye centers were horizontally aligned. This alignment step enforces a consistent facial orientation across samples and improves invariance to head tilt. (Figure 4)
3) **Region of Interest (ROI) Extraction:** Following alignment, a facial region of interest was extracted by cropping the detected bounding box. To preserve relevant contextual information (e.g., jawline and hairline) while limiting background clutter, the bounding box was symmetrically expanded by 30% in each direction before cropping.
4) **Spatial Normalization:** The extracted facial ROI was resized to a fixed resolution of 224×224 pixels, ensuring compatibility with standard convolutional neural network backbones and enabling efficient batch processing during training.

This preprocessing was applied consistently to the training, validation and test splits and to all seven emotion classes. By standardizing face location, scale and orientation, it allows the subsequent deep learning models to focus on expression-related variations rather than large geometric differences between images.

Fig. 4: Example of face detection and eye-based alignment using MTCNN on an image from our dataset.

4) Exploratory Data Analysis (EDA): An Exploratory Data Analysis was conducted to assess the quality, balance and consistency of the processed dataset.

a) Class Distribution and Stratification: As reported in Table I-II and illustrated in Figure 2 and 3, the dataset exhibits a degree of class imbalance. The *happy* class is significantly over-represented, with approximately 10,000 samples, whereas emotions such as *fear* and *disgust* contain substantially fewer instances.

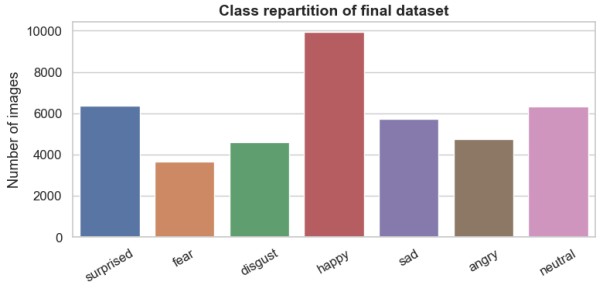

Fig. 5: Illustration of the class imbalance of the dataset.

This imbalance largely reflects the inherent bias of in-the-wild facial expression datasets, where positive or easily recognizable expressions (e.g., happiness) are more frequently captured and annotated than subtle or less frequent affective states such as fear or disgust.

Consequently, this imbalance was explicitly taken into account during dataset splitting and model training through the use of a weighted random sampler, ensuring that each class is presented to the model with equal probability.

b) Qualitative Assessment and anomalies removal: To assess data quality before model training, a visual inspection of randomly selected training samples was done to ensure the robustness of the preprocessing pipeline.

Given the large number of images, an automatic anomaly detection step was implemented to complement manual inspection. The objective of this step is to identify potentially problematic images that could negatively impact learning (e.g., corrupted files, almost blank images or extreme illumination).

Each image is first converted to grayscale and global statistics such as mean intensity and variance are computed. Several types of anomalies are flagged:

- **Corrupted images**: Images that cannot be loaded or processed correctly.
- **Low-variance images**: Nearly uniform or heavily degraded images with minimal intensity variation.
- **Overly dark images**: Images whose average intensity falls below a predefined threshold.
- **Overly bright images**: Images whose average intensity exceeds a predefined threshold.

Images detected as anomalies were excluded from the compiled dataset. This automated screening reduces the need for manual inspection and increases the reliability of the preprocessing pipeline, ensuring that the training data predominantly consist of informative and visually consistent facial expressions.

Fig. 6: Examples of images removed from the dataset.

IV. Model Selection and Transfer Learning

A. Transfer Learning on Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have become the reference paradigm for image classification tasks, including facial emotion recognition (FER). Their effectiveness relies on their ability to exploit local spatial correlations: neighboring pixels in a face image exhibit strong dependencies (e.g., around the eyes, mouth and eyebrows). However, training deep CNN architectures from scratch requires very large annotated datasets in order to prevent overfitting and is highly sensitive to initialization. In FER, publicly available datasets such as RAF-DB and AffectNet remain relatively modest in size compared to large-scale natural image collections like ImageNet. Consequently, transfer learning has emerged as a highly effective and widely adopted strategy in this field.

Transfer learning consists in leveraging models pre-trained on large datasets and adapting them to a specific target task. In this work, CNN backbones pre-trained on ImageNet-1K are used as initialization. Early convolutional layers generally learn low-level and generic features (edges, gradients, simple textures) that are largely task-independent, whereas deeper layers encode higher-level semantic representations that can be adapted to facial structures and expressions. Reusing these learned features significantly reduces training time and computational cost while improving generalization performance.

1) Model Architectures Selection:

a) ConvNeXt.: ConvNeXt is a modern convolutional architecture that revises classical ConvNet design in light of advances from Vision Transformers while remaining fully convolutional. It employs deep stages with large kernels, inverted bottlenecks and layer normalization, achieving high computational efficiency and competitive ImageNet performance (top-1

accuracy above 80%). When fine-tuned on FER, ConvNeXt captures both global facial structure and fine-grained local patterns.

b) ResNet-50.: ResNet-50 consists of 50 layers organized into residual blocks with skip connections. Skip connections mitigate the vanishing gradient problem and allow stable training of deep networks. In this project, a ResNet-50 backbone pretrained on ImageNet was used. Both full fine-tuning and partial fine-tuning (freezing layers 1–2 and updating layers 3–4) were explored, offering a good trade-off between computational cost and performance.

c) VGGFace2-based model: VGGFace2 is a large-scale face recognition dataset containing over 3.3 million images of 9,131 identities with variations in pose, age, illumination and ethnicity. Models pretrained on VGGFace2 learn features highly specialized for human faces, making them particularly suitable for FER. In this work, a VGGFace2-pretrained backbone was used as a feature extractor, with a task-specific classification head added. Fine-tuning the deeper layers allows the network to shift from identity to expression discrimination while reusing robust facial representations.

B. Training Strategy and Implementation

1) Data Augmentation Strategy: To improve robustness to natural variations in facial appearance and capture conditions, data augmentation was applied during training. The final augmentation pipeline was designed to remain visually plausible for facial images while increasing variability in pose, scale and illumination.

The following transformations were used:

- **RandomResizedCrop** with a scale in $[0.8, 1.0]$: randomly crops and rescales the image while preserving most of the face, simulating variations in camera distance and framing and encouraging the model to focus on locally informative regions rather than a fixed global layout.
- **RandomHorizontalFlip**: leverages the approximate left–right symmetry of human faces, since expressions such as happiness or sadness remain semantically valid under horizontal mirroring.
- **RandAugment** (2 operations, magnitude 6): applies a small number of randomly selected geometric or photometric perturbations with moderate intensity, increasing diversity in a controlled way without creating unrealistic distortions of facial structure.
- **ColorJitter** (brightness and contrast): introduces mild changes in brightness and contrast to mimic illumination differences between studio images (KDEF) and in-the-wild images (RAF-DB, AffectNet), reducing sensitivity to lighting conditions.
- **ImageNet normalization**: input images are converted to tensors and normalized with the standard ImageNet mean and standard deviation, matching the distribution expected by the pretrained CNN backbones used for transfer learning.

This augmentation strategy aims to expose the model to realistic variations commonly encountered in FER (pose, framing, and lighting), while preserving the underlying facial expression so that labels remain valid.

2) Handling Class Imbalance: The compiled FER dataset exhibits a notable class imbalance, with emotions such as *happy* and *neutral* being much more frequent than expressions like *disgust* or *fear*, especially in in-the-wild datasets. If left unaddressed, this imbalance would bias the model toward majority classes and lead to poor recognition performance on underrepresented emotions.

To mitigate this effect, a *weighted random sampler* was employed during training instead of a standard uniform sampler. Class weights were computed inversely proportional to class frequencies,

$$w_c = \frac{N_{\text{total}}}{N_{\text{classes}} \times N_c},$$

where N_c denotes the number of samples belonging to class c. Each training image then received a sampling weight corresponding to the weight of its class, so that minority classes are drawn more frequently into mini-batches.

This sampling strategy effectively rebalances the contribution of each emotion during optimization, without modifying the loss function itself. It encourages the model to maintain higher sensitivity to rare but semantically important expressions (e.g., *disgust*, *fear*), which is crucial for realistic facial emotion recognition in imbalanced real-world data.

3) Transfer Learning combined with Fine Tuning: For all architectures, the original classification head was replaced with a task-specific linear layer mapping backbone features to the seven target emotions, followed by a softmax activation. Networks were initialized with pretrained weights (ImageNet-1K for ConvNeXt-Tiny, ResNet-50, and Inception, and VGGFace2 for the face-specialized model), so that generic visual or facial features learned on large-scale datasets could be reused for FER.

For the ConvNeXt-Tiny backbone, several transfer learning regimes were compared (Figure 7):

- **Pure transfer learning (frozen backbone):** The pretrained convolutional backbone was used as a fixed feature extractor and only the newly added classification head was trained.
- **Partial fine-tuning:** The last ConvNeXt stage (block 7) was unfrozen and fine-tuned jointly with the classification head, while earlier layers remained frozen to preserve generic low-level representations.
- **Full fine-tuning:** All convolutional layers were unfrozen and updated, allowing both low and high-level features to adapt to FER.

For ResNet-50 and the VGGFace2-based model, only the last convolutional blocks were fine-tuned, as is common in FER, since low-level filters (edges, textures) transfer well from large face or image datasets, whereas adapting only high-level facial features is usually sufficient to specialize the model to emotion recognition.

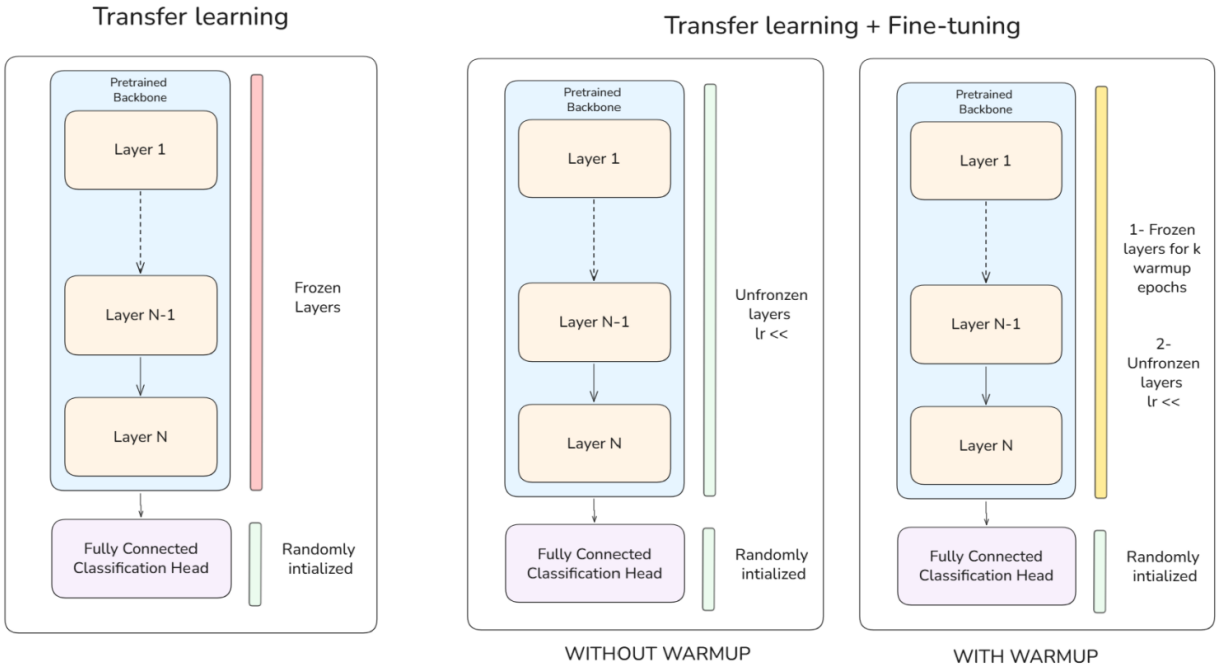

Fig. 7: Schemas of different training strategies tested during the project.

Table III summarizes the quantitative results obtained with the three main configurations on the test set.

TABLE III: Impact of transfer learning and fine-tuning strategy for ConvNeXt-Tiny on FER.

Strategy	Accuracy	F1-score	Test loss
Transfer learning	0.533	0.5074	1.259
Partial fine-tuning	0.796	0.7756	0.6579
Full fine-tuning	**0.864**	**0.8510**	**0.4596**

Full fine-tuning clearly outperformed both pure transfer learning and partial fine-tuning across all metrics, achieving the highest accuracy, top-3 accuracy, F1-score and the lowest test loss. This configuration was therefore selected as the default training strategy for ConvNeXt-Tiny in the remainder of the experiments.

In addition, a short *warm-up* configuration was tested for the full fine-tuning regime, where only the classification head was trained during the first 5 epochs before unfreezing the backbone. This strategy was intended to stabilize optimization when updating all layers, but in practice it did not bring a clear improvement (Table IV) over direct full fine-tuning and mainly resulted in a shifted learning curve (Figure 8).

TABLE IV: Impact of warmup for ConvNeXt-Tiny training.

Strategy	Accuracy	F1-score	Test loss
With warmup	0.866	0.8531	0.4578
Without warmup	0.864	0.8510	0.4596

The overall training strategy can be summarized as follows: 1) **Initialization:** Build the chosen backbone (ConvNeXt-Tiny, ResNet-50 or VGGFace2) with ImageNet/VGGFace2 pre-trained weights and replace the original head with a linear layer over the seven emotions. 2) **Transfer mode:** Set the backbone to frozen, partially fine-tuned (last block only) or fully fine-tuned according to the configuration. 3) **Optimization:** Train using class-balanced mini-batches with data augmentation, Mixup and EMA regularization, Adam optimizer and a learning-rate scheduler. 4) **Validation and selection:** Monitor validation loss and macro F1, save the best checkpoint and apply early stopping when no further improvement is observed.

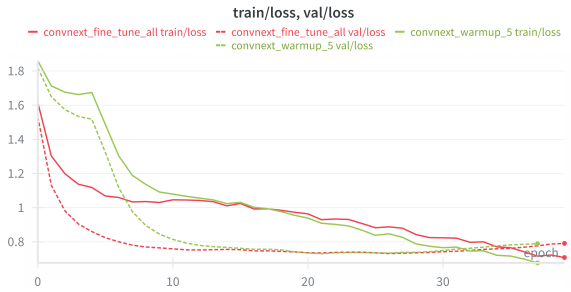

Fig. 8: Training and validation loss evolution with and without warmup strategy.

V. Experiments and Results

1) Implementation details and Environment: All experiments were conducted using the **PyTorch** framework on a dedicated **NVIDIA RTX 4070 GPU** with 8 GB of VRAM. A local setup was chosen over cloud alternatives to avoid runtime interruptions, ensure consistent training sessions and provide

immediate access to local storage for debugging and dataset management.

Training metrics, including loss, accuracy and learning rate progression were logged using the **Weights & Biases (W&B)** platform. This enabled systematic monitoring of model performance, reproducibility and side-by-side comparison of different architectures and fine-tuning strategies.

2) Model Training Protocol and HyperParameter Optimization:

3) Training protocol: A standardized training protocol was applied across all models to ensure fair comparison. Key components included:

- **Loss Function:** Cross-Entropy with **Label Smoothing** to mitigate overconfident predictions. This is particularly important in facial emotion recognition, where certain classes (e.g., Fear vs. Surprise) are visually similar.
- **Optimizer:** Adam optimizer, chosen for its adaptive learning rate properties and reliable convergence.
- **Learning Rate Schedule:** A **Cosine Annealing** scheduler gradually reduces the learning rate, promoting convergence to flatter minima and better generalization.
- **Batch Size:** 32, balancing memory usage and gradient stability.
- **Early Stopping:** Training was halted if validation loss did not improve for a fixed number of epochs. The model weights corresponding to the highest validation accuracy were restored.

Three complementary regularization techniques were used to improve generalization and stabilize training.

a) Dropout in the classification head: To further limit overfitting on the relatively small FER dataset, a dropout layer was inserted in the classification head before the final linear layer. By randomly zeroing a fraction of the head activations at each training step, dropout discourages co-adaptation between neurons and encourages the model to learn more robust, distributed representations of facial expressions.

b) Mixup: During training, Mixup is applied to a subset of mini-batches by linearly combining pairs of input images and their corresponding labels with a mixing coefficient sampled from a Beta distribution. This creates interpolated facial expressions and soft targets, reducing overfitting and encouraging the model to behave linearly between training samples.

c) Exponential Moving Average (EMA): An Exponential Moving Average of the model parameters is maintained throughout training and used at validation time. By smoothing rapid parameter updates, EMA yields a more stable estimator of the network and typically improves validation performance compared to evaluating the raw, non-averaged weights. It was only used for ConvNext training.

4) Hyper parameters optimization: For ConvNeXt-Tiny and ResNet-50, hyperparameters were explored using Bayesian sweeps in Weights & Biases rather than exhaustive search: short runs of 15–20 epochs were used to probe the effect of key parameters, mainly learning rates and to identify promising value ranges, as illustrated in Figure 9, where the best runs clearly concentrate in specific backbone and head learning-rate regions while other hyperparameters show weaker impact.

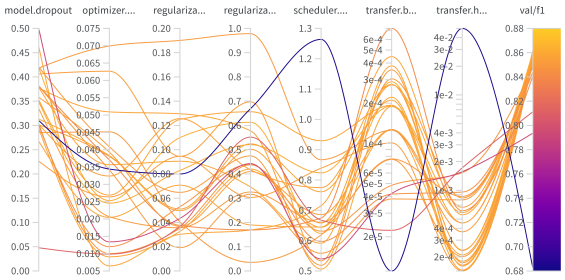

Fig. 9: Resume of the bayesian hyper parameter optimization using W&B sweep.

The following table presents the final training configurations retained for each backbone, summarizing the main transfer-learning settings and optimization hyperparameters used in the best-performing models.

5) Training dynammics: Figure 10 compares the training and validation loss trajectories of the three best configurations (VGGFace2, ResNet-50, ConvNeXt-Tiny). All models show a monotonic decrease of training loss, while validation loss remains stable without late divergence, indicating controlled overfitting. ConvNeXt-Tiny converges faster and reaches the lowest validation loss, whereas VGGFace2 decreases more slowly and stabilizes at a higher loss level, consistent with the quantitative results reported in Section V-6.

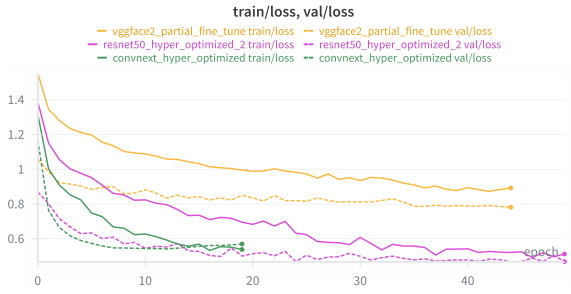

Fig. 10: Training and validation loss curves for the three best-performing models. Faster convergence and lower validation loss are observed for ConvNeXt-Tiny compared to ResNet-50 and VGGFace2.

6) Quantitative performance: All models were assessed on an independent test set of 4,]515 images. Metrics included overall accuracy, macro-averaged precision, recall and F1-score. Table VI presents the results.

ConvNeXt-Tiny achieves the best performance on the test set in terms of accuracy and macro-averaged precision, recall and F1-score, which is consistent with its faster convergence and lower validation loss observed in the training curves of Figure 10.

TABLE V: Final training configurations for the three best-performing models.

Model	Transfer mode	Epochs	Dropout	Label sm.	Backbone LR	Head LR
VGGFace2 (partial)	Freeze, partial	45	0.50	0.10	1.0×10^{-4}	1.0×10^{-4}
ResNet-50 (sweep)	Freeze, partial	100	0.19	0.001	5.45×10^{-5}	9.37×10^{-4}
ConvNeXt-Tiny (sweep)	Full fine-tune	20	0.29	0.033	1.26×10^{-4}	2.10×10^{-4}

TABLE VI: Performance comparison of evaluated models on the FER test set.

Model	Acc.	Prec.	Recall	F1-Score
ConvNeXt	**0.866**	**0.853**	**0.853**	**0.853**
ResNet50	0.856	0.845	0.842	0.843
VGGFace2	0.85	0.836	0.838	0.837

7) Class-wise Performance analysis: The model's ability to differentiate expressions was further analyzed via the confusion matrix (Figure 11). High recognition rates were observed for distinct emotions such as *Happiness*, while subtle expressions like *Fear* and *Sadness* and *Disgust* were more frequently misclassified, often as *Surprise*, *Neutral* or *Anger*.

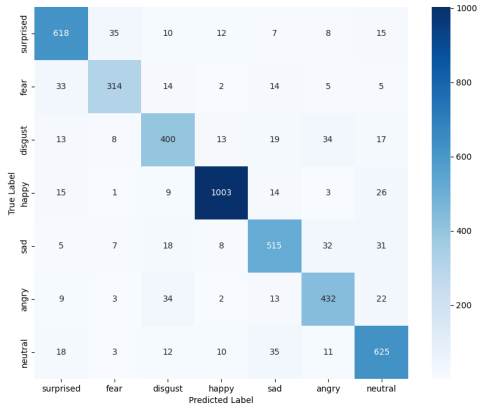

Fig. 11: Confusion matrix for the best-performing model on the test set : ConvNext tiny.

These results highlight the intrinsic difficulty of static FER, especially for subtle and overlapping facial movements. As illustrated in Figure 12, several samples exhibit highly ambiguous expressions, for which even human observers may struggle to distinguish, for example, *fear* from *surprise* or *anger* from *disgust*, since these emotions recruit partially similar facial muscle configurations. In addition, a fraction of AffectNet is known to suffer from noisy or uncertain annotations, particularly for high-arousal, high-valence expressions, which further complicates the learning of fine-grained emotion boundaries for our models.

To better assess performance on difficult classes, precision–recall (PR) curves were computed for each emotion and each backbone. Figure 13 shows the PR curves for the *Anger* class. ConvNeXt-Tiny consistently dominates ResNet-50 and VGGFace2 over most of the recall range, indicating higher precision at comparable recall levels and confirming its superiority on one of the most challenging categories.

(a) Fear/Surprise (b) Anger/Disgust (c) Sad/Neutral

Fig. 12: Examples of errors made by the model (Prediction/Ground-truth).

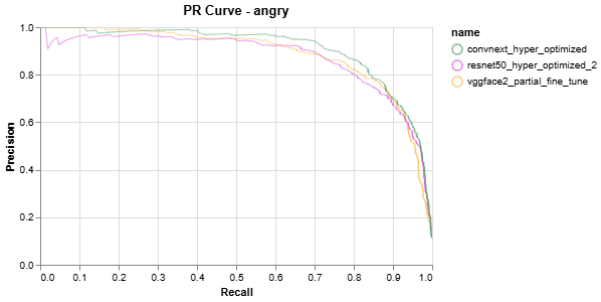

Fig. 13: Precision–recall curves for the *Anger* class comparing ConvNeXt-Tiny, ResNet-50, and VGGFace2.

Overall, ConvNeXt achieves the strongest results when examining the PR curves for the most difficult emotions (Anger, Sadness, Disgust and Fear), indicating a superior ability to discriminate between these closely related classes.

8) Models comparison: Taken together, the training dynamics, quantitative metrics and class-wise analyses consistently indicate that ConvNeXt-Tiny is the most effective backbone for the considered FER setting. First, ConvNeXt-Tiny converges faster and reaches the lowest validation loss throughout training (Figure 10), suggesting a more favorable optimization landscape under the adopted data augmentation and regularization strategy. Second, on the independent test set it achieves the highest accuracy and macro-averaged precision, recall, and F1-score (Table VI), with a non-trivial margin over ResNet-50 and VGGFace2 despite a comparable parameter budget.

ResNet-50 remains a strong baseline, with stable training curves and competitive performance, but it systematically lags behind ConvNeXt-Tiny on all global metrics. The VGGFace2-based model benefits from face-specialized pretraining and performs well on clear, prototypical expressions, yet it ex-

hibits slower convergence and higher residual loss, which is consistent with its slightly lower test performance. Finally, ConvNeXt-Tiny also achieves the strongest results when examining the PR curves for the most difficult emotions (Anger, Sadness, Disgust and Fear), indicating a superior ability to discriminate between these closely related classes.

VI. Evaluation of the Selected Model

A. Performance metrics and comparison with SOTA

The ConvNeXt-Tiny model, trained on the compiled RAF-DB + AffectNet subset + KDEF dataset, reaches an accuracy of 86.6% and a macro-averaged F1-score of 0.853 on the test set (Table VI). These results are in line with recent state-of-the-art FER systems trained on single datasets, which typically report accuracies around 92–93% on RAF-DB and 65–67% weighted average recall on AffectNet. Although our model is trained on a heterogeneous combination of sources rather than on each benchmark in isolation, its performance lies between these reported values, suggesting that it remains competitive while targeting a more challenging multi-dataset scenario.

B. Qualitative evaluation with Grad-CAM

To gain insight into the decision process of the ConvNeXt-Tiny model, Grad-CAM visualizations were generated for representative test images. For correctly classified samples, the heatmaps align well with human intuition: the model focuses on the mouth region for *Happiness* and *Disgust*, on the eyes and eyebrows for *Anger* and *Fear*, and on a broader facial area for *Neutral* expressions.

(a) Happy (b) Angry (c) Disgust

Fig. 14: Grad-CAM visualizations for selected correctly classified test images.

Failure cases also provide useful information. Figure 15 shows an example where an *Anger* expression is misclassified as *Disgust*. The Grad-CAM map reveals that the model concentrates on the lower part of the face (nose and mouth wrinkles), which are indeed highly indicative of disgust, while giving less weight to the eye region that better characterizes anger. This suggests that some confusions stem from the model over-emphasizing specific action units that are shared across emotions.

Overall, Grad-CAM analyses indicate that the model relies on semantically meaningful facial regions, even when it makes mistakes, which is encouraging from an interpretability standpoint.

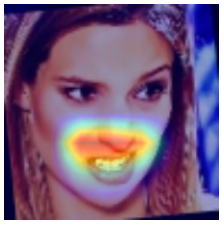

Fig. 15: Example of misclassification: an angry face predicted as disgust. The model focuses on the mouth/nose region, which is strongly associated with disgust.

C. Evaluation in "real-time" conditions

To assess the practical usability of the model, ConvNeXt-Tiny was integrated into a simple real-time pipeline using a standard laptop webcam. Face detection was performed on each frame using a Hamming-based sliding window (Haar-like) detector, followed by our pre-processing and FER model inference. The system was able to detect faces and output stable emotion predictions at interactive frame rates.

As illustrated in Figure 16, the model provides coherent predictions across a variety of conditions, including different head poses, lighting changes, the presence of eyeglasses and partial occlusions (e.g., hood or covered lower face). While quantitative latency measurements are beyond the scope of this project, these experiments suggest that the trained model can be deployed in basic real-time scenarios without major degradation in recognition quality.

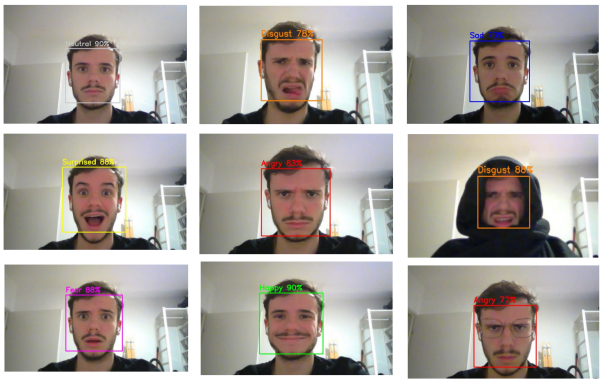

Fig. 16: Real-time inference examples with the ConvNeXt-Tiny model on webcam frames. The system remains robust to glasses, moderate occlusions and pose variations.

VII. Conclusions

This work presented an end-to-end facial emotion recognition (FER) pipeline based on modern deep learning and transfer learning techniques. Three complementary datasets (RAF-DB, an AffectNet subset and KDEF) were harmonized into a unified seven-class label space, together with a rigorous preprocessing pipeline including face detection, alignment, anomaly removal, and stratified splitting. This provided a robust foundation for training and fairly comparing multiple convolutional architectures under heterogeneous conditions.

A systematic evaluation of ConvNeXt-Tiny, ResNet-50, and a VGGFace2-based model was conducted using a shared training protocol with advanced data augmentation, class-imbalance handling, and regularization techniques. ConvNeXt-Tiny consistently outperformed the other models, achieving 86.6% accuracy and a macro F1-score of 0.853 on an independent test set of 4,515 images. These results are competitive with recent state-of-the-art FER methods, despite the increased difficulty of a multi-dataset training setup.

Class-wise analysis shows that emotions such as *Happiness* are recognized reliably, whereas more subtle expressions (*Fear*, *Sadness*, *Disgust*) remain challenging and are often confused with related categories. Grad-CAM visualizations confirm that the model focuses on semantically relevant facial regions, and a real-time webcam prototype demonstrates robustness to moderate variations in pose, illumination and occlusion.

Despite these strengths, several limitations remain. The model operates on static images and does not leverage temporal information, performance on underrepresented classes is still limited, and potential demographic biases were not explicitly evaluated. Consequently, while the proposed system is effective and practical, further work is required to improve fairness, generalization and robustness in real-world deployments.

VIII. Supplementary Material

Supplementary material belonging to this project can be accessed as follows:

- **Datasets:**
  - KDEF: https://www.kaggle.com/datasets/chenrich/kdef-database?select=neutral
  - RAF-DB: https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset/
  - AFFECTNET (subset): https://www.kaggle.com/datasets/fatihkgg/affectnet-yolo-format
- **Code and outputs:** https://github.com/AlbinMorisseau/FER
- **Model weights:** Google Drive folder
- **Training and evaluation logs (Weights & Biases):** https://wandb.ai/almorisseau-universit-t-klagenfurt/projects

References

[1] D. Bryant and A. Howard, "Ethical considerations of facial expression recognition ai in human-robot interactions," in *Proceedings of the ECSARA Workshop at RO-MAN 2024*, Paris, France, Aug 2024, pp. 1–10. [Online]. Available: https://perso.ensta-paris.fr/~shangguan/Ro-manWS/files/ECSARA_2024_paper.pdf

[2] "Aratek: How facial emotion recognition expresses feelings," https://www.aratek.co/news/how-does-facial-emotion-recognition-express-your-feelings, 2024.

[3] D. Mekinec, "Visage technologies: Facial emotion recognition guide," https://visagetechnologies.com/facial-emotion-recognition-guide/, Jul 2024.

[4] "Edps tech dispatch: Facial emotion recognition," https://www.edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf, 2021.

[5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[6] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] "Ssrn paper on ethics and ai," https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5200140, 2024.

[8] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*, Tokyo, Japan, Oct 2016, pp. 433–436.

[9] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Santiago, Chile, Dec 2015, pp. 19–27.

[10] N. Mehendale, "Facial emotion recognition using convolutional neural networks (ferc)," *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, Mar 2020. [Online]. Available: https://link.springer.com/article/10.1007/s42452-020-2234-1

[11] S. Alok, "RAF-DB DATASET," https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset, 2023, kaggle.

[12] F. Karakus, "Affectnet yolo-format facial expression dataset," https://www.kaggle.com/datasets/fatihkgg/affectnet-yolo-format, 2023, accessed: 2026-01-26.

[13] R. Chen, "Kdef (karolinska directed emotional faces) dataset," https://www.kaggle.com/datasets/chenrich/kdef-databaset, 2020, accessed: 2026-01-26.

[14] I. de Paz Centeno, "ipazc/mtcnn: v1.0.0," Oct. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13901378