Kaunas University of Technology

Faculty of Informatics

Real-Time Facial Expression Recognition in the Wild Using Efficient Hybrid CNN–Transformer Architectures

Master's Final Degree Project

Nadeem Saeed Baloch

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Real-Time Facial Expression Recognition in the Wild Using Efficient Hybrid CNN–Transformer Architectures

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Nadeem Saeed Baloch

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Associate Prof. Dr. Alfonsas Misevičius

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Nadeem Saeed Baloch

Real-Time Facial Expression Recognition in the Wild Using Efficient Hybrid CNN–Transformer Architectures

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Nadeem Saeed Baloch

Confirmed electronically

Kaunas University of Technology

Faculty of Informatics

Master's final degree project

Topic of the project	Real-Time Facial Expression Recognition in the Wild Using Efficient Hybrid CNN–Transformer Architectures
Requirements and conditions	
Supervisor	**Prof. Dr. Armantas Ostreika**

Nadeem Saeed Baloch. Real-Time Facial Expression Recognition in the Wild Using Efficient Hybrid CNN–Transformer Architectures. Master's Final Degree Project / supervisor Prof. Dr. Armantas Ostreika; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer Science, Informatics (B01).
Keywords: facial expression recognition, AffectNet, in-the-wild FER, CNN, Transformer, hybrid architecture, macro-F1, real-time inference, class imbalance, efficiency.

Kaunas, 2026. 67 pages.

Summary

Facial expression recognition in the wild remains a challenging computer vision task because facial images captured in unconstrained conditions are affected by pose variation, illumination changes, blur, occlusion, background clutter, and substantial intra-class variability. In addition, large-scale in-the-wild datasets such as AffectNet-8 exhibit strong class imbalance, which makes it insufficient to evaluate performance using overall accuracy alone. These challenges motivate the development of methods that are not only accurate, but also efficient and stable enough for practical real-time use.

The aim of this thesis is to investigate efficient CNN-only and hybrid CNN–Transformer models for real-time facial expression recognition in the wild and to determine which factors most strongly influence performance on the AffectNet-8 dataset. The study uses the full AffectNet-8 training split and the balanced validation split under a fixed experimental protocol. Two baseline CNN architectures, ResNet50 and ConvNeXt-Tiny, are evaluated first. This is followed by a series of lightweight hybrid ConvNeXt-Tiny + Transformer variants, as well as controlled experiments with imbalance-aware and stability-focused training strategies, including focal loss, class-aware alpha weighting, exponential moving average, warmup scheduling, lower learning rate, and gradient clipping.

The experiments show that the strongest verified gains in this work come primarily from imbalance-aware training rather than from the lightweight Transformer refinement alone. Among the completed single-model experiments, the best overall result is achieved by the ConvNeXt-Tiny baseline trained with focal loss and class-aware alpha, which reaches a validation accuracy of 0.5829 and a validation macro-F1 of 0.5772. The best hybrid single model, based on ConvNeXt-Tiny with a two-block Transformer refinement and stability-focused training, achieves a validation accuracy of 0.5799 and a validation macro-F1 of 0.5708. This shows that the hybrid approach becomes highly competitive, but does not surpass the best CNN-only baseline under the main validation metrics.

The thesis also includes parameter count, latency, and FPS benchmarking under a fixed batch-size-one GPU inference protocol. The results show that ResNet50 is the fastest completed model, while the ConvNeXt-Tiny baseline with focal loss and class-aware alpha provides the best overall accuracy–efficiency trade-off. Additional offline evaluation using checkpoint ensembling and horizontal-flip test-time augmentation brings the hybrid approach very close to the best CNN-only model, but this configuration is treated as an inference enhancement rather than the main real-time deployment setup.

The main conclusion of the thesis is that, in the tested AffectNet-8 setting, the largest improvements in facial expression recognition are achieved through stable imbalance-aware training, while lightweight hybrid CNN–Transformer refinement improves competitiveness but does not provide a superior final single-model result. Therefore, for practical real-time FER under the completed experimental conditions, the strongest recommendation is the CNN-only ConvNeXt-Tiny baseline with focal loss and class-aware alpha.

Nadeem Saeed Baloch. Veido išraiškų atpažinimas realiuoju laiku naudojant hibridines CNN transformerių architektūras. Magistro baigiamasis projektas / vadovas prof. dr. Armantas Ostreika; Informatikos fakultetas, Kauno technologijos universitetas.

Studijų kryptis ir sritis (studijų krypties grupė): Informatikos mokslai, Informatika (B01).

Raktiniai žodžiai: veido išraiškų atpažinimas, „AffectNet“, natūrali FER, CNN, Transformer, hibridinė architektūra, makro-F1, realaus laiko išvados, klasių disbalansas, efektyvumas.

Kaunas, 2026. 67 puslapiai.

Santrauka

Veido išraiškų atpažinimas natūralioje (nekontroliuojamoje) aplinkoje išlieka sudėtinga kompiuterinio matymo užduotimi, nes realiomis sąlygomis užfiksuotus veido vaizdus veikia galvos pozos variacijos, apšvietimo pokyčiai, suliejimas, dalinis uždengimas, fono triukšmas ir didelis tos pačios klasės vaizdų kintamumas. Be to, didelio masto natūralios aplinkos duomenų rinkiniai, tokie kaip „AffectNet-8“, pasižymi ryškiu klasių disbalansu, todėl modelių našumui įvertinti nepakanka vien bendro tikslumo rodiklio. Šios aplinkybės skatina kurti metodus, kurie būtų ne tik tikslūs, bet ir pakankamai efektyvūs bei stabilūs, kad juos būtų galima taikyti praktinėse realaus laiko sistemose.

Šio darbo tikslas – ištirti efektyvius vien tik CNN ir hibridinius CNN–Transformer modelius, skirtus veido išraiškų atpažinimui realiuoju laiku, ir nustatyti, kurie veiksniai labiausiai lemia našumą naudojant „AffectNet-8“ duomenų rinkinį. Tyrime taikomas visas „AffectNet-8“ mokymo rinkinys ir subalansuota validavimo imtis, laikantis fiksuoto eksperimentinio protokolo. Pirmiausia įvertinamos dvi bazinės CNN architektūros – „ResNet50“ ir „ConvNeXt-Tiny“. Vėliau tiriama lengvų hibridinių „ConvNeXt-Tiny + Transformer“ variantų seka ir atliekami kontroliuojami eksperimentai su klasių disbalansą atsižvelgiančiomis bei mokymo stabilumą didinančiomis strategijomis: židinio nuostolių (focal loss) funkcija, klasėms pritaikyti alfa svoriai, eksponentinis slenkamasis vidurkis (EMA), apšilimo (warmup) planavimas, mažesnis mokymosi greitis ir gradiento ribojimas (gradient clipping).

Eksperimentų rezultatai rodo, kad didžiausi patikimai patvirtinti šio darbo pagerėjimai daugiausia susiję su klasių disbalansą atsižvelgiančia mokymo strategija, o ne vien su lengvu „Transformer“ patikslinimu. Iš vieno modelio (single-model) eksperimentų geriausią bendrą rezultatą pasiekė bazinis „ConvNeXt-Tiny“ modelis, apmokytas naudojant židinio nuostolių funkciją ir klasėms pritaikytus alfa svorius: validavimo tikslumas – 0,5829, validavimo makro-F1 – 0,5772. Geriausias hibridinis vieno modelio variantas („ConvNeXt-Tiny“ su dviejų blokų Transformer patikslinimu ir stabilumą didinančiomis mokymo priemonėmis) pasiekė 0,5799 validavimo tikslumą ir 0,5708 validavimo makro-F1. Tai rodo, kad hibridinis metodas tampa labai konkurencingas, tačiau pagal pagrindinius validavimo rodiklius nepranoksta geriausio vien tik CNN bazinio varianto.

Darbe taip pat pateikiamas parametrų skaičiaus, išvadų delsos (latency) ir kadrų per sekundę (FPS) palyginimas, atliktas taikant fiksuotą vieno vaizdo (batch size = 1) GPU išvadų (inferencijos)

matavimo protokolą. Rezultatai rodo, kad „ResNet50“ yra greičiausias iš įvertintų modelių, o „ConvNeXt-Tiny“ bazinis variantas su židinio nuostolių funkcija ir klasėms pritaikytais alfa svoriais užtikrina geriausią bendrą tikslumo ir efektyvumo kompromisą. Papildomas „neprisijungus“ (offline) vertinimas, taikant dviejų kontrolinių taškų ansamblį (checkpoint ensembling) ir horizontalų apvertimą bandymo metu (test-time augmentation), leidžia hibridinį metodą labai priartinti prie geriausio vien tik CNN modelio, tačiau ši konfigūracija vertinama kaip išvadų kokybės pagerinimas, o ne pagrindinė realaus laiko diegimo schema.

Pagrindinė darbo išvada yra tokia: tiriamoje „AffectNet-8“ konfigūracijoje didžiausi veido išraiškų atpažinimo pagerėjimai pasiekiami taikant stabilias klasių disbalansą atsižvelgiančias mokymo strategijas, o lengvas hibridinis CNN–Transformer patikslinimas padidina konkurencingumą, tačiau neužtikrina geresnio galutinio vieno modelio rezultato. Todėl praktinėms realaus laiko FER užduotims, esant šiame darbe taikytoms eksperimentinėms sąlygoms, rekomenduotinas sprendimas yra vien tik CNN pagrįstas „ConvNeXt-Tiny“ modelis su židinio nuostolių funkcija ir klasėms pritaikytais alfa svoriais.

Table of contents

List of figures

List of tables

List of abbreviations and terms

Abbreviations:

CNN – Convolutional Neural Network
CSV – Comma-Separated Values
CUDA – Compute Unified Device Architecture
EMA – Exponential Moving Average
FER – Facial Expression Recognition
FPS – Frames Per Second
GPU – Graphics Processing Unit
JSON – JavaScript Object Notation
LBP – Local Binary Patterns
PyTorch – Open-source deep learning framework used for implementation
TTA – Test-Time Augmentation
ViT – Vision Transformer

Terms

AffectNet-8 – An eight-class configuration of the AffectNet dataset used for in-the-wild facial expression recognition, containing the expression categories neutral, happy, sad, surprise, fear, disgust, anger, and contempt.

Class imbalance – A data distribution problem in which some classes contain substantially more samples than others, which may bias model training toward majority categories.

Confusion matrix – A tabular representation of classification results showing how often each true class is predicted as each possible class.

Cross-entropy loss – A standard loss function for multi-class classification that measures the difference between predicted class probabilities and true labels.

Focal loss – A modified classification loss function that reduces the relative contribution of easy samples and focuses training more strongly on hard or misclassified examples.

Hybrid CNN-Transformer model – A model architecture that integrates convolutional feature extraction with contextual modelling done by a Transformer**.**

In-the-wild conditions- Unconstrained real-world image conditions, characterized by natural variability of pose, illumination, occlusion, blur, background, and facial appearance.

Latency- The mean time it takes to run one input image in an inference, typically in milliseconds.

Macro-F1 - The arithmetic mean of the per-class F1-scores, with the same weight given to each class, regardless of the frequency of its class.

Params Count- The number of trainable model parameters, often used as a measure of model size and complexity.

Per-class F1-score - The F1-score of a single individual class, the balance of precision and recall of a single class.

Single-image inference -In the context of this thesis, single-image inference that uses a single instance of an image to produce a single output value.

Test-Time Augmentation (TTA) - An evaluation method where many transformed versions of the same input are processed in inference and their predictions are added up.

Transformer encoder block – A neural network module, based on self-attention and feed-forward processing, which models the relationship between multiple feature tokens.

Validation accuracy - Proportion of correctly classifying samples in the validation set.

Introduction

Project novelty and relevance

Facial expression recognition is a significant field of affective computing and computer vision, which seeks to automatically recognise the emotional state of a human face, using a computer image as input. The problem has gained more and more relevance in recent years due to the increased demand on intelligent systems that can respond to the natural human affective state and enhance system responsiveness and usability. Unconstrained real-life scenarios still face significant challenges in recognition of facial expressions using reliable methods despite the significant progress in deep learning.

In actual settings, there are many factors that distort facial images and make it difficult to recognize the face, including variations in head pose, changes in light, occlusion, blur, clutter in the background and variations in the intensity of expression and identity. Particularly, these issues become particularly tangible in in-the-wild datasets, where pictures are gathered based on uncontrolled sources of the Internet rather than laboratory. Simultaneously, large-scale benchmarks like AffectNet-8 provide an extra challenge: strong imbalance between classes, in which particular expressions are represented by a large number of additional samples as compared to others. Consequently, even a model that does well with regard to overall accuracy, may still be unhelpful on minority and difficult classes.

This is why the current facial expression recognition studies can not be reduced to the sole aim of the maximization of the accuracy. Balanced quality of recognition, class wise behavior as well as the practical computational efficiency should also be evaluated. The trade-off between recognition performance and inference speed is critical in particular with systems to be used in real-time or near-real-time environments. This is the reason why we study models, which are not only accurate, but also efficient and stable in imbalanced in-the-wild situations.

This thesis is novel in that a controlled and efficiency-conscious study of facial expression is conducted on the complete AffectNet-8 system with CNN-only and lightweight hybrid CNN-Transformer-based systems. The purpose of the work is not only to compare the end recognition scores, but also to identify the factors that participate in the improvement of the results with the help of the definite protocol. In this thesis, special attention is given to the influence of imbalance-aware training, including focal loss and class-aware weighting, as well as to the practical comparison of validation accuracy, macro-F1, per-class behavior, confusion matrices, parameter count, latency, and FPS. The obtained results show not only which completed model performed best overall, but also whether lightweight hybrid CNN–Transformer refinement provides sufficient benefit to justify its additional computational cost.

Aim and objectives

The aim of this thesis is to investigate efficient CNN-only and hybrid CNN–Transformer models for real-time facial expression recognition in the wild and to determine which factors most strongly contribute to robust and efficient performance on the AffectNet-8 dataset.

To achieve this aim, the following objectives were defined:

1. To analyse the problem of facial expression recognition in the wild, the characteristics of AffectNet-8, and recent CNN-based and hybrid CNN–Transformer approaches for facial expression recognition.
2. To design and implement baseline CNN-only models and lightweight hybrid CNN–Transformer models for eight-class facial expression recognition.

3. To examine how the imbalance-conscious and stability-oriented training approaches influence the performance of a model on the AffectNet-8 class distribution.
4. To evaluate the completed models using validation accuracy, macro-F1, per-class results, and confusion matrices.
5. To benchmark the completed single-model configurations using parameter count, latency, and FPS to assess the real-time accuracy–efficiency trade-off.
6. To identify the strongest overall single-model configuration and to determine whether the tested lightweight hybrid refinement provides a practical advantage over strong CNN-only baselines.

Document structure

The thesis consists of three main chapters, conclusions, references, and appendices.

Chapter 1, **In-the-Wild FER Analysis**, reviews facial expression recognition in the wild, AffectNet-8 as a benchmark dataset, CNN-based facial expression recognition methods, Transformer-based and hybrid CNN–Transformer approaches, efficiency-oriented facial expression recognition, and the role of class imbalance and training stability. The chapter concludes with the identification of the research gap addressed in this work.

Chapter 2, **Model and Experimental Design**, defines the problem and scope of the thesis, presents the overall research design, describes the dataset and experimental protocol, data preprocessing and augmentation, baseline and hybrid model designs, training strategies, evaluation metrics, real-time benchmarking protocol, and reproducibility settings.

Chapter 3, **Experimental Results and Efficiency Evaluation**, presents the implementation environment, baseline experiments, hybrid architecture experiments, imbalance-aware training experiments, final hybrid inference-side evaluation, efficiency benchmarking, confusion matrix and per-class analysis, and overall discussion of results.

The thesis ends with **Conclusions**, where the main findings are summarized and the objectives of the work are addressed. Additional material, such as detailed per-class, tables, confusion matrices and information about reproducibility is contained in the appendices.

1. In-the-Wild FER Analysis

1.1. Facial expression recognition in the wild

Facial expression recognition (FER) is a technology that tries to recognize human affective states based on facial features. The first FER studies would frequently be constructed and experimented on in controlled laboratory conditions, in which the subjects were photographed under relatively steady illumination, frontal position and limited background variation. But practical uses of FER are more and more in demand. Performing computations in unconstrained real-life scenarios, where images are subject to pose, lighting, occlusion, blur, identity, scale of face and complexity of background. Consequently, there has been a progressive change in the field of facial expression recognition as an in-the-wild method [1], [2], [3].

Such a change substantially increases the complexity of the recognition task. In-the-wild environments confound expression-related facial expressions with numerous expression-unrelated variations, including head pose, lighting conditions, partial occlusions, age-related variations, identity bias and degradation in image quality. These can decrease the discriminative strength of facial representations, and raise the risk of overfitting, particularly when minority expressions are underrepresented in the training data. The literature on deep FER has focused on the fact that, under unconstrained scenarios, robust recognition faces data insufficiency, data-to-sample ratios, and various nuisance variations that require robust training methods to accommodate these issues [2].

The FER landscape has been altered greatly due to the development of deep learning. It was a neural network (CNNs) that became the most popular due to its capacity to learn directly out of images, hierarchical visual representations. Most recently, Transformer-based and neural network-based architectures based on hybrid CNN-Transformer have been considered in order to supplement powerful local feature extraction using a wider contextual modeling. The rationale of hybrid FER methods is that facial expression recognition is not only dependent on local cues, such as mouth curvature or eyebrow deformation, but also on interactions between multiple facial regions. However, the effectiveness of this extra global modeling hinges on the dataset, the optimization strategy and the efficiency constraints of the target application [2], [4].

The other significant feature of the new FER research is an increased focus on evaluation beyond mere accuracy. In in-the-wild datasets that are challenging, overall accuracy can conceal poor performance on challenging or minority classes. Because of this, recent FER research is beginning to report class-sensitive measures, in particular, macro-F1, in addition to confusion matrices and per-class results. This broader assessment perspective is especially significant when the task is large-scale, imbalanced datasets, as it offers a more reliable perspective of whether a model is effective at scale on imbalanced datasets, rather than on dominant classes [2].

In practical terms in-the-wild FER should also be viewed as an accuracy-efficiency issue as well as a pure classification issue. A model with a high recognition quality but with too high latency or computational cost can be of less use in real-time applications. Hence, current FER studies are more inclined to promote efficiency-conscious comparisons, particularly of systems that are to be used in the real world. This renders the investigation of lightweight CNN and hybrid CNN-Transformer models specifically pertinent, as these models are aimed at striking the balance between the quality of representations and the computational capacity [5].

To conclude, recognition of facial expression in the wild is a difficult yet practically significant problem. Its major challenges are due to uncontrolled imaging conditions, variability independent of expression, imbalance in the dataset, and the necessity to maintain a high efficiency level along with recognition quality. These attributes drive the experimental design of this thesis that will

compare CNN-only and lightweight hybrid CNN–Transformer approaches under a fixed protocol and evaluates them using both recognition and efficiency metrics.

1.2. AffectNet-8 as an FER benchmark

One of the most significant benchmarks on in-the-wild affective computing and facial expression recognition is AffectNet. It was proposed to overcome the gap in the large-scale, annotated facial affect datasets that are collected in realistic settings. It was constructed by the Internet images that were retrieved with the help of the large set of emotion-related queries in different languages, and includes approximately one million facial images in total, with a large manually annotated sub-set to analyze expressions, valence, and arousal. Due to its scale and unrestricted nature, AffectNet has become a primary reference point to the current state of FER research.

What is so significant about AffectNet is its scale, but also the diversity of the conditions that are depicted in the images. Because the data are provided by uncontrolled web sources, the benchmark consists of a lot of variance in terms of head pose, light, occlusion, camera quality, intensity of expression, and amount of clutter in the background. In this regard, AffectNet is far more realistic to image conditions in a laboratory setting as compared to many older lab-like FER datasets. It thus offers a more real world scenario to assess whether a model is capable of generalizing beyond clean and highly standardized face images.

The second significant benefit of AffectNet is that it can be utilized by both categorical and dimensional affect analysis. In the dimensional setting, it establishes annotations of valence and arousal, whereas in the categorical setting, it provides facial expression recognition on discrete expression labels. Such flexibility has seen the dataset being extensively used in various affective computing tasks. The eight-class categorical FER configuration is of interest in the context of this thesis, as it is well-aligned with the objective of testing CNN-only and hybrid CNN–Transformer models to classify expressions under real-world conditions.

Meanwhile, AffectNet is a challenging benchmark. Its realism has its challenges that are directly related to this thesis; visually ambiguous expressions, high intra-class variance, overlap between semantically similar categories, and imbalance between classes in the training data. The properties of this dataset are not only that it is suitable to test the raw classification performance, but also to answer more subtle questions, such as the stability of training, sensitivity of the minority-class, and the trade-off between accuracy and efficiency. It is one of the primary reasons why AffectNet should be used as an appropriate benchmark when it comes to testing both CNN-only and hybrid FER models within the framework of a controlled experimental study.

AffectNet is also a useful benchmark from a methodological perspective because it enables comparisons across multiple dimensions of model behavior. Its scale is large enough to support deep learning experiments with modern backbones, yet its complexity is high enough to expose weaknesses in both architecture design and optimization strategy. In particular, it is well suited for studies that aim to distinguish between the contribution of model architecture and the contribution of training methodology, which is one of the central goals of this thesis [1], [6].

For these reasons, AffectNet-8 is used in this thesis as the main benchmark for all experiments. It provides a challenging in-the-wild evaluation environment, supports large-scale controlled comparison, and reflects the class-sensitive and efficiency-sensitive issues that are central to the study. As a result, it is an appropriate benchmark for analysing whether lightweight hybrid CNN–Transformer architectures offer practical benefit over strong CNN-only baselines under a fixed and reproducible protocol.

1.3. CNN-based FER methods

Convolutional neural networks have become the dominant foundation for modern facial expression recognition because they can learn hierarchical representations directly from facial images. In FER, CNNs are particularly effective at capturing local and mid-level visual patterns such as eyebrow deformation, eye closure, mouth curvature, and other expression-related facial structures Survey literature on deep FER indicates that CNN-based systems continue to dominate the field, particularly due to their strong feature-extraction capabilities and viable compatibility with transfer learning based on large-scale image-classification pretraining. Meanwhile, the surveys also underline the idea that FER is still challenging due to the overfitting, identity bias, illumination variation, head pose, and other factors that are not related to the expressions [2], [7].

The CNN-based FER methods have over time evolved to what appears to be a more task oriented design rather than the simpler one which merely adapts the backbone to the task. Most subsequent methods added attention mechanisms, emphasis on local regions, masking modules and feature modulation robustness oriented. The reason why these changes are required is that facial regions can often be represented by spatially localised cues which are not as informative as other facial areas. As a result, CNN-based FER research increasingly focused on improving local discriminative feature extraction and suppressing irrelevant variations [2], [8], [9], [10], [11].

An important example of this line of work is EfficientFace, which was proposed as a lightweight and robust FER network for practical in-the-wild use. EfficientFace combines a local-feature extractor with a channel-spatial modulation mechanism, aiming to preserve robustness under pose variation and occlusion while remaining computationally efficient. The method is especially relevant to this thesis because it illustrates a practical design philosophy similar to the one adopted here: rather than maximizing architectural complexity, it seeks to improve robustness and real-world usability under unconstrained conditions [5].

Another influential attention-oriented FER model is the Distract Your Attention Network (DAN). DAN is motivated by two ideas: first, different facial-expression classes may be visually similar and require stronger class separation; second, multiple facial regions contribute jointly to expression recognition. The model therefore combines a feature clustering mechanism with multi-head attention and attention fusion to encourage the network to attend to multiple relevant facial areas. Even though DAN is not a Transformer model in the traditional ViT sense, it is indicative of the overall change in FER towards more complex spatial reasoning over multiple regions of the face [12].

In spite of these developments, the CNN-based FER techniques are currently limited to the unconstrained large scale setting. Their high local inductive bias is often useful, but purely convolutional processing might be less effective when understanding of expressions relies on more global relationships between distant parts of the face. This has been amongst the factors that have led to later FER studies beginning to look at attention-based and Transformer-based options. However, CNN models will always be highly relevant since they tend to be easier to optimise, more computationally efficient and sometimes strong enough that they can serve as practical deployment baselines. This is as well in line with the results of the present thesis where CNN-only baselines have been a necessity to provide insights into the impacts of subsequent hybrid refinements [2], [12].

To conclude, CNN-based FER techniques continue to be a powerful and essential basis of the field. They offer strong local feature extraction, realistic transfer-learning, and favourable efficiency, with advanced variants providing better sensitivity to subtle, localised facial signals by attention and

feature modulation. These properties justify why CNN baselines are at the center of attention in this thesis and why ConvNeXt-Tiny was chosen as the primary backbone to be used later in this thesis in the form of hybrid experiments [13].

1.4. Transformer-based and hybrid CNN–Transformer FER methods

The increased attention of Transformer-based FER methods is largely inspired by the potential of self-attention to capture long-range dependencies and global context. This is appealing in facial expression recognition where the expressions are not defined by local features that are unique to a single area of the face, but by the coordinated changes in many areas of the face. Transformer-based systems, thus, offer to supplement local facial information with a global contextual interaction. In FER, however, this transition is not as straightforward since the face datasets are often smaller and imbalanced compared to large generic image-classification benchmarks, and pure Transformer models might need strong data support or careful design to be kept efficient and stable [2], [14], [15].

A timely and pertinent example is Visual Transformers with Feature Fusion (VTFF) which was suggested to provide robust recognition of facial expressions in the wild. VTFF is a cross-combination of LBP and CNN features, followed by the modelling of Transformers to enrich the image of visual words with both texture and learned visual features. The importance of this work is that it reflects one of the central ideas in the future, namely, FER hybrid systems, instead of abandoning the convolutional feature extraction, many methods do not abandon the local inductive structure and then they use Transformer-like processing to enhance global interaction [16].

Furthermore, the direction of the local-global fusion that FER studies have been pursuing since then continues. As an example, HFE-Net was suggested to simultaneously capture subtle changes in local expression and whole-face information with a hybrid feature-extraction approach. Other recent reviews of hybrid CNN-Transformer modelling also report that hybrid configurations are appealing due to the strong local pattern-extracting properties of CNNs and the broader contextual modelling of Transformers. Practically, the methods are supposed to address the weakness that purely convolutional FER models might not completely utilize the correlations between distributed facial areas [2], [17], [18].

Meanwhile, hybrid CNNTransformer FER models also present a number of real-life issues. First, they may not be easily optimised compared to CNN-only baselines, particularly on unbalanced and noisy in-the-wild datasets. Second, their further contextual modelling might better improve some than other classes, which implies gains are not always evenly distributed throughout the entire expression set. Third, even lightweight hybrid refinement often costs extra as compared to a powerful CNN backbone only. These issues make hybrid FER an accuracy–efficiency trade-off problem rather than a purely architectural one. This perspective is directly relevant to the present thesis, where the hybrid models became highly competitive but still did not surpass the strongest CNN-only baseline in the final single-model comparison [2], [16], [17], [19], [20], [21].

For this reason, the literature suggests that hybrid FER systems should be evaluated carefully and under controlled conditions. It is not enough to ask whether a hybrid model can improve the final score; one must also determine whether the gain comes from the hybrid architecture itself, from the surrounding training strategy, or from evaluation-time enhancement such as ensembling and test-time augmentation. This is one of the reasons why the present thesis compares lightweight hybrid models against strong CNN-only baselines under a fixed protocol and separates architecture effects from loss-design and optimization effects [2], [17].

In summary, Transformer-based and hybrid CNN–Transformer FER methods are motivated by the need to combine local facial representation with broader contextual modelling. According to the literature, such models can be very competitive, particularly when they are implemented as local-global fusion systems as opposed to implementing them as pure Transformer replacements. Their practical utility however, will be determined by their ability to strike a balance between the quality of recognition, the stability of optimisation and the efficiency of inference. This makes them a suitable and timely focus for the present thesis, which investigates lightweight hybrid refinement under a fixed AffectNet-8 protocol.

1.5. Efficient and real-time FER

Over the last few years the research on facial expression recognition has increasingly shifted beyond the purely accuracy-enhancing research towards more practically-oriented research that is deployment ready. This change is significant since, in many of the possible uses of FER, such as human-computer interaction, driver monitoring, educational interfaces and assistive systems, fast and stable inference is required instead of merely high recognition quality. This has made real-time FER to be an essential sub topic of the larger subtopic. Here, model assessment must extend beyond recognition measures, to include computational measures such as number of parameters, latency and frames per second (FPS).

The biggest issue with efficient FER is that enhancement of recognition quality can be achieved at the expense of model complexity. The larger backbones, multi-branch systems and attention-heavy architectures might be capable of high feature representation, but they also come with increased memory consumption and slower inference. This is especially applicable in in-the-wild FER, where an ideal response of the recognition system is expected to be observed under real-world conditions. Thus, not only the reduction of parameters, but the advantageous balance between recognition performance and computational cost is achieved.

An exemplary case of such a course is EfficientFace, which was, in fact, suggested as a lightweight and robust FER network. EfficientFace was developed to enhance the ability of expression recognition as well as the computational efficiency, demonstrating that practical FER systems need not inherently be very large. It is designed to focus on the compact convolutional feature extraction with feature modulation mechanisms aimed at retaining robustness under unconstrained facial conditions. The fact that such a work exists shows that efficiency-conscious FER is not a minor implementation consideration, but a significant goal of the research.

The concept of real-time FER needs also a clear definition of operations. Practically, there can be various interpretations of the term real-time, depending on the hardware, the input resolution, the amount of data to be included in the batch and the benchmarking process. Thus, the term should not be used vaguely by a thesis or an experimental study. It must instead specify the evaluation hardware, the batch size of inferences, the input resolution and the timing protocol to compute latency and FPS. It is on the basis of such a fixed protocol that fair comparisons of speed-oriented claims can be made. This is particularly significant in research such as the one in question in which there are a number of architectures that have similar recognition quality, and yet differ meaningfully in the practical inference cost.

In terms of the methodology, efficient FER should consequently be examined as an accuracy-efficiency trade-off issue. A model with high recognition score may not be the most resourceful when its inference cost is significantly greater than that of an alternative which is slightly weaker, but orders of magnitude faster. The highest-performing model, on the other hand, might be inappropriate when its overall recognition quality is too low to support minority or challenging classes. This implies that an effective FER research has to consider speed and quality as a single

goal, as opposed to treating them as a goal of their own. This principle is adhered to in this thesis, where the latency, FPS and the number of parameters is reported along with the accuracy, macro-F1 and the results obtained by classes.

Overall, effective and real-time FER is a promising research direction, as successful implementation does not only need high recognition accuracy but also practical implementation. The literature demonstrates that both compact and robust FER architectures are practical and useful with the methodological implication of real-time claims being based on explicit and reproducible measurement protocols. This viewpoint inspires the efficiency analysis that will be done later in the thesis that uses CNN-only and lightweight hybrid CNN-Transformer models as a reference point not only by their recognition performance but also by their practical inference costs [5], [22], [23].

1.6. Class imbalance and stable training in FER

One of the most crucial issues in the large-scale recognition of facial expressions is class imbalance. With unconstrained datasets like AffectNet, certain expressions are much more common than others. This implies that when training, the model is exposed to a disproportionate amount of majority classes, and a proportionately smaller amount of minority classes, which contribute fewer learning signals. Consequently, a typical optimization process can result in a classifier that is robust on common expressions, but weak on infrequent or challenging classes. This issue is especially severe in FER as the most difficult categories, including disgust or contempt, are also the least represented in the training data [2].

Class imbalance has two significant implications on methodology. To begin with, overall accuracy is not a sufficient evaluation measure since it may conceal poor performance on minority classes. Second, the very training strategy becomes a key component in the design of the model. That is, the issue is not just what architecture is in use, but also how the inequity in the distribution of classes is addressed by the optimization process. This is the reason why the need to use class-sensitive assessment particularly macro-F1 and per-class analysis are highlighted in many FER studies and surveys that emphasized the need to use class-sensitive evaluation (especially macro-F1 and per-class analysis) when working with large imbalanced datasets [24].

Focal loss is a commonly used loss function to deal with imbalance. Initially proposed to detect dense objects, focal loss was developed to lower the relative weight of simple, well-classified examples, and to make training more biased towards hard examples. The essence is that standard cross-entropy can lead to the situation where easy samples dominate the learning signal, whereas focal loss rebalances the learning signal to ensure that difficult samples contribute more to the update. The original application area of object detection is directly applicable to imbalanced classification tasks like FER [25].

Practically, focal loss may be used alongside with class-aware weighting to provide minority classes with extra attention when optimizing. It is helpful in cases where the imbalance is pronounced and there are always certain categories of expressions that are underrepresented. Nevertheless, the loss design on imbalance-awareness is not always enough. Instability due to optimization noise, rapid overfitting, or sensitivity to interactions between the pretrained backbone features and newly added modules can also cause large-scale FER training to be affected by instability. This is why not only an appropriate loss function but also a more comprehensive optimization strategy, such as learning-rate scheduling, regularization, early stopping, and in some cases smoothed model-parameter updates are needed to ensure stable training in FER [2].

Training a generalizable architecture can be even more challenging when the underlying architecture is more complex, such as when CNN backbones are used with attention or

Transformer-based refinement. When this happens, optimization can be more fragile, particularly when the distribution of data is either noisy or highly imbalanced. This implies that comparisons of architecture must be carefully interpreted: a weaker result does not necessarily imply that the architectural concept is less sensitive to the training configuration; in fact, in some cases, it may imply that the given model is more sensitive to the training configuration. This observation can be directly applied to the current thesis where imbalance-aware and stability-focused training approaches become highly competitive only after the introduction of such training methods [26].

As a more broad methodological approach, class imbalance and training stability should then be considered first-class research issues in FER and not as secondary issues related to implementation. A model that is found to be weak through a generic training setup can be competitive through a loss function and optimization procedure that is more appropriate to the data distribution. On the other hand, a strong backbone can not be allowed to achieve its full potential when it is trained on a loss that overemphasizes the majority classes. Therefore, the literature is very helpful in supporting the idea that to provide fair FER comparison the architecture evaluation and training-strategy evaluation need to be conducted [25].

To sum up, class imbalance and training on stables are the basic problems with modern facial expression recognition, particularly on large in-the-wild datasets. They influence model choice and interpretation of results, and can help explain why macro-F1, per-class behavior, and the design of losses are critical in FER research. This knowledge directly informs the experimental design of the current thesis, where the focal loss, the weighting based on classes, and the optimization oriented on stability are explored as the key determiners in enhancing the recognition performance in the AffectNet-8 setup [27], [28], [29], [30].

1.7. Summary of the analysis and research gap

The discussion included in this chapter reveals that the problem of recognition of facial expressions in the wild is a complicated and practically significant research problem. Contemporary FER systems are required to work in uncontrolled conditions that encompass pose variation, change of illumination, occlusion, blur, identity differentiation and background complexity. Moreover, large-scale in-the-wild datasets like AffectNet-8 are characterized by a significant imbalance in the number of classes representing different emotions. This is why trustworthy FER assessment should not only look at the accuracy, but also macro-F1, per-class behaviour and error structure [31].

The literature review also demonstrates that CNN-based methods are still a powerful backbone to FER because of their capability to learn strong local facial representations and their realistic computational efficiency. Simultaneously, more recent Transformer-based and hybrid CNN + Transformer methods have been suggested to supplement local feature extraction with a more comprehensive contextual modelling across the several face regions. The driving factor behind these hybrid approaches is the fact that facial expressions are not solely determined by isolated local factors, but are also jointly conditioned by multiple parts of the face.

Nevertheless, the literature review shows that the practical advantages of hybrid FER models have not been completely determined. Even though hybrid methods can be very competitive, not all hybrid methods will be effective in comparison with robust CNN baselines. Specifically, the reviewed studies indicate that architecture refinement might not be enough in case the training set is not suitable to the data distribution. This problem is particularly acute with large, unbalanced in-the-wild datasets, where loss design, class-sensitive optimisation, training stability can play a large part in the final outcome. Thus, the literature indicates that a significant methodological gap exists: FER models are to be compared in a controlled environment that will isolate the effect of architecture on training-strategy effects.

The other crucial gap that was noted in the analysis is in respect to efficiency-oriented evaluation. Most studies in FER focus on recognition accuracy, but fewer offer a clear, practically-founded comparison and include latency, FPS, and number of parameters, as well as class-sensitive measures. Because this thesis is explicitly stated in terms of real-time or near-real-time FER, this efficiency-conscious assessment is paramount. It would be hard without it to determine whether a more robust model is also practically appropriate to deploy it.

Resting on this analysis, the research gap that will be tackled in the thesis can be put in the following way. A controlled, reproducible study to compare CNN-only and lightweight hybrid CNN transformer FER models on the AffectNet-8 benchmark under a single fixed protocol, as well as the effect of imbalance-sensitive and stability-sensitive training regimes. Moreover, such a study must evaluate not only the quality of recognition, but also the practical accuracy-efficiency trade-off of the number of parameters, latency, and FPS.

This is what directly drives the work that is conducted in the subsequent chapters. The thesis therefore investigates whether lightweight hybrid CNN–Transformer refinement can improve in-the-wild facial expression recognition beyond strong CNN baselines, whether the main gains come from architecture or from training strategy, and which completed model provides the strongest balance between recognition quality and real-time efficiency under the AffectNet-8 setting.

2. Model and Experimental Design

2.1. Problem definition and project scope

Facial expression recognition in unconstrained conditions remains a challenging computer vision problem. In real-world scenarios, facial images are affected by substantial variations in pose, illumination, occlusion, blur, background clutter, facial scale, and expression intensity. In addition, large in-the-wild datasets such as AffectNet-8 introduce another important difficulty: strong class imbalance, where majority classes such as happy and neutral are represented much more frequently than minority classes such as disgust and contempt. These factors make it difficult to build a recognition system that is not only accurate, but also stable, efficient, and suitable for practical real-time use [1], [2].

The problem addressed in this thesis is the development and evaluation of facial expression recognition models for eight-class expression classification under in-the-wild conditions, using AffectNet-8 as the main benchmark. The study focuses on the recognition of the following expression categories: neutral, happy, sad, surprise, fear, disgust, anger, and contempt. The task is formulated as a supervised image classification problem in which a cropped facial image is mapped to one of the predefined expression classes.

A second important aspect of the problem is the tension between recognition quality and computational efficiency. A model intended for real-time or near-real-time use should not only provide strong validation accuracy and macro-F1, but should also maintain low latency, sufficiently high frames per second, and moderate parameter count. Therefore, this thesis treats facial expression recognition not only as a classification problem, but also as an accuracy–efficiency trade-off problem. Here real-time is understood operationally to refer to inference of low-latency single-image under a fixed benchmarking configuration and the detailed measurement protocol is given in Section 2.9.

The thesis is also narrowly scoped so as to provide a focused and controlled study. To begin with, it is confined to image-based facial expression recognition, and does not take into account temporal modelling of video. Second, the research is founded on one benchmark dataset, AffectNet-8, without the need to conduct cross-dataset generalisation experiments. Third, large-scale ensemble architectures are not even considered as the main deployment models, but only single-model CNN baselines and lightweight hybrid CNNTransformer variants. Fourth, it is not covered in the work how multimodal affect recognition, facial action unit detection, valence-arousal regression as the primary task, and landmark-only recognition pipelines could be done. These were made in order to maintain the study on a technical scale as well as to enable a rigorous comparison between architecture design, training strategy, and computational efficiency.

Under this umbrella, the thesis is based on three closely intertwined questions. The initial question is whether a lightweight hybrid CNN-Transformer architecture can achieve strong performance on in-the-wild facial expression recognition. The second question is the gains of the main performance are by architecture refinement, or by stable imbalance-aware training strategies. Third, in a given experimental situation, what model will result in the most realistically balanced inference efficiency and recognition quality.

In this respect, the primary issue explored in this thesis is how to design and test a practical facial expression recognition system on AffectNet-8 that can still be competitive in terms of severely imbalanced classes without compromising on practical inference speed. It is not the purpose of the study to purport a new state-of-the-art result on all published protocols. Rather it intends to give a controlled, reproducible analysis of the factors that have the most significant effect on performance

in this environment, and in particular on macro-F1, difficult minority classes, and the real-time effectiveness of the final models.

2.2. Overall research design

The comprehensive research design of this thesis was developed into a controlled comparative research study. The main idea of the experimental design was that the evaluation protocol was held constant, and only one major factor of the experimental design was changed at a time, e.g. the choice of backbone, the refinement of hybrids, the loss function, or the training stabilisation strategy. This allowed differentiating improvements that resulted due to architectural changes and improvements that resulted due to the training strategy. The research was thus formulated to not only to compare the final scores, but also to uncover the real source of performance changes.

The research process was organised into four main stages. The initial step had robust CNN-only baselines. To achieve these goals, two baseline architectures were chosen ResNet50 and ConvNeXt-Tiny. ResNet50 was used as a popular and well understood reference architecture whereas ConvNeXt-Tiny was used as a more powerful modern Convolutional backbone with a better balance between representational strength and computational efficiency. This stage was to decide which CNN backbone to consider as the primary guide with which to proceed with the hybrid experiments.

The second phase was concerned with lightweight hybrid CNN-Transformer modelling. Upon finding the more robust CNN backbone, more lightweight versions were created by combining ConvNeXt-Tiny feature extractor with a simple Transformer refinement unit. The hybrid design was deliberately made small to be of relevance to the real time framing of the thesis. The various levels of hybrid depths were experimented in a controlled manner starting with a simple single block Transformer refinement and extending to a two block variant. This step was intended to assess whether the global contextual modelling with Transformer layers can be quantified as a benefit over the CNN-only backbone.

The third phase analyzed the impact of the imbalance-conscious and stability-oriented training methods.Since AffectNet-8 is strongly imbalanced, a simple comparison of architectures under standard cross-entropy loss would not be sufficient. For this reason, the study investigated focal loss, class-aware alpha weighting, and additional training-stability mechanisms such as exponential moving average, lower learning rate, warmup scheduling, and gradient clipping. The purpose of this stage was to determine whether the largest gains in recognition performance come from model architecture or from the training procedure used under class imbalance.

The fourth stage addressed practical deployment-oriented evaluation. After the main training experiments were completed, the strongest models were benchmarked in terms of parameter count, latency per image, and frames per second. In addition, a final inference-side evaluation was carried out for the strongest hybrid checkpoints using checkpoint ensembling and horizontal-flip test-time augmentation. This was done to estimate the best achievable hybrid performance without changing the training process, while still keeping the main real-time conclusions grounded in single-model inference.

The experimental protocol adopted a fixed experimental protocol in all the core experiments. The training and validation configuration used by the AffectNet-8 training and validation model, the input resolution, the evaluation measures, and the logic used to benchmark its results were the same across the models under comparison unless a particular experimental factor was deliberately manipulated. This made internal comparisons to be fair and interpretable. Validation accuracy and macro-F1 were taken as the main recognition metrics and the detailed class-wise interpretation was

based on the results provided in per-class format and confusion matrix. The efficiency was measured in terms of the number of parameters, the latency and the FPS with a batch-size-one GPU inference protocol.

The ultimate model test in this thesis was thus an evaluation of two complementary views. The recognition quality was the first view, which was assessed with the help of validation accuracy, macro-F1, and per-class behaviour. The second view was practical efficiency, which was measured by the number of parameters and real-time measurements of inference. Not a single number was used as the final conclusions of the study, but rather the combined analysis of these two perspectives under a consistent and reproducible design.

The general workflow of the study can be simplified as follows: setting the base, refining the hybrid, training analysis based on the imbalance and practical efficiency benchmarking. This design assists the primary objective of the thesis, which is to assess efficient CNN-only and hybrid CNN-Transformer algorithms to recognize facial expressions in-the-wild in a controlled environment, as well as to be able to identify which factors have the greatest influence on the strong and efficient performance.

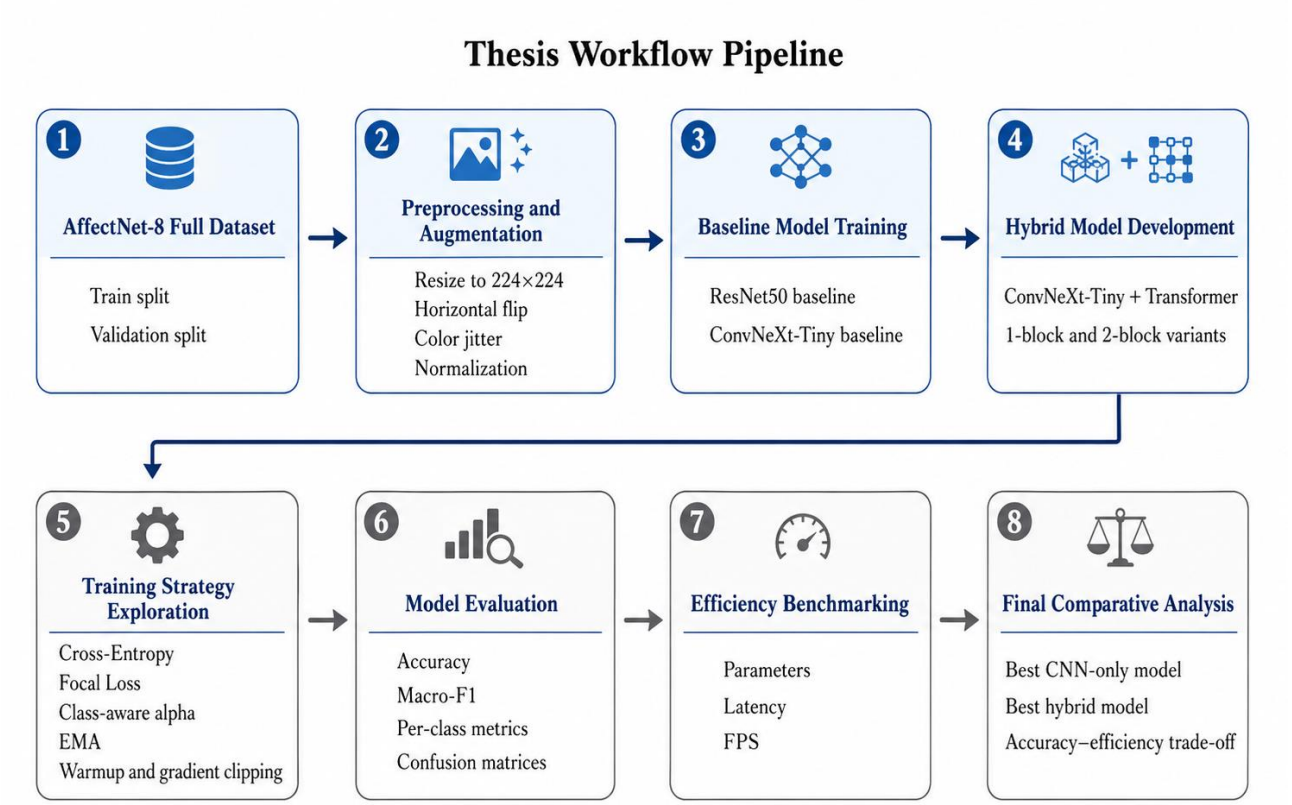

Fig. 1. Thesis Workflow Pipeline

The general workflow of the thesis is shown in Fig. 1. The paper starts with the AffectNet-8 full dataset, preceded by preprocessing and data augmentation. Then controlled experiments are carried out with respect to baseline CNN models and proposed hybrid CNN-Transformer models. Various training approaches are tested with class imbalance and the trained models are compared using classification and efficiency measures.

2.3. Dataset and experimental protocol

This thesis has adopted AffectNet-8 as the primary reference dataset in terms of the recognition of facial expression in the wild. One of the most publicly available datasets on affective computing and analyzing facial expressions in unconstrained, real-world scenarios is AffectNet. Its images were gathered on the Internet and thus will have significant changes in illumination, pose, occlusion, blur, facial scale, expression intensity and background complexity. These characteristics render the dataset suitable to assess practical face expression recognition models that can be used in non-laboratory settings.

In this work, the eight-class configuration of AffectNet was used. The target expression classes were as shown below: neutral, happy, sad, surprise, fear, disgust, anger, and contempt. The mapping of the class index, which was used throughout the thesis, was: 0 -neutral, 1 - happy, 2 - sad, 3 - surprise, 4 - fear, 5 - disgust, 6 - anger, and 7 - contempt. All training, evaluation, logging and benchmarking scripts were applied to this fixed label mapping of the entire experimental workflow so as to achieve consistency across all training, evaluation, logging, and benchmarking scripts.

The dataset setup of this thesis included a complete training split, and a balanced validation split. The training set had 287,651 samples, and the validation set had 3,999 samples. The validation split was approximately balanced across the eight categories, with about 500 samples per class, making it particularly useful in testing recognition quality not just based on the majority-class dominance. In comparison, the training set was very skewed. The largest class was the happy with 134,415 samples, and the smallest was the contempt with only 3,750 samples. This extreme imbalance was among the main reasons why the emphasis on macro-F1 and per-class assessment were prioritised throughout the research.

Split	Number of samples	Notes
Training	287,651	Full AffectNet-8 training split
Validation	3,999	Balanced validation split

Table 1. AffectNet-8 train and validation split summary

The verified class distribution of the training set used in this work was as follows: neutral – 74,874, happy – 134,415, sad – 25,459, surprise – 14,090, fear – 6,378, disgust – 3,803, anger – 24,882, and contempt – 3,750. The validation split consisted of 500 images per class, with the exception of contempt which had 499 images. These statistics established that the experimental environment is highly affected by the imbalance of classes and reporting an overall accuracy would not be sufficient to provide reliable interpretation of model behaviour. Due to this reason, accuracy was always reported along with macro-F1 and class-wise results [1].

Class index	Expression	Number of samples
0	Neutral	74,874
1	Happy	134,415
2	Sad	25,459
3	Surprise	14,090
4	Fear	6,378
5	Disgust	3,803
6	Anger	24,882
7	Contempt	3,750

Table 2. AffectNet-8 training class distribution

All the main experiments in this thesis were done on a single internal protocol. The training and validation splits, input size, metric definitions, and benchmark methodology were the same across the models under comparison unless there was a specific experimental factor that is purposely altered. This protocol was followed so that they could do comparisons that were fair between CNN-only baselines, lightweight hybrid CNNTransformer variants, and various training strategies. This would imply in practice that the internal comparisons in the thesis are controlled comparisons as opposed to loosely collected independent runs.

The data was accessed by indexing files (CSV) which contained image identifiers, image paths, expression labels and other metadata. Indexed CSV files were also used to facilitate reliable, reproducible, and uniform loading of the same data into all experiments. It also made it easier to maintain a consistent training and validation procedure during the project. The training and validation dataset were loaded in distinct CSV files in the implementation, avoiding the accidental mixing of the data splits and minimizing the risk of evaluation leakage.

This thesis has an experimental protocol which can thus be summarised as follows. To begin with, all of the single-model experiments were trained and validated by using the identical AffectNet-8 training and validation set-up. Second, all models were trained and assessed through the identical map of classes and the resolution of inputs. Third, consistent metrics (validation accuracy, macro-F1, class-wise measures, confusion matrixes and, later, parameter counts, latency, and FPS) were used to evaluate performance. Fourth, every experimental finding was also interpreted in the realms of the quality of recognition and the practical efficiency. This protocol allowed to draw meaningful conclusions regarding the relative effectiveness of the choice of architecture, hybrid refinement, imbalance-aware training, and inference efficiency in a common evaluation environment.

To clarify this, the main role of AffectNet-8 in this thesis is not only to offer a benchmark of the classification accuracy, but also to reveal how various models behave in the real-life scenario of imbalance and in-the-wild visual variability. The dataset, therefore, is the main experimental focal point where the major research questions in the thesis are researched.

2.4. Data preprocessing and augmentation

The preprocessing and augmentation pipeline, which was employed in this thesis, was crafted to meet two conflicting needs. On the one hand, the input images needed to be normalised and resized uniformly to allow the modern deep learning models to be trained steadily. Conversely the augmentation approach had to be conservative enough so as not to overly distort delicate facial-expression information. Given that facial expression recognition depends on the local features that are fine-grained, and particularly around the eyes, eyebrows, and mouth, over-aggressive augmentation can distort the very information that the model should learn to differentiate.

All the models in the thesis were input resolutions of 224 × 224 pixels. In both training and validation, all images were resized to this constant spatial resolution, and then sent to the network. Three reasons were behind this decision. First, the 224 224 is a standard resolution, which is widely supported by ImageNet-pretrained backbones like ResNet50 and ConvNeXt-Tiny. Second, controlled comparison of models and training configurations is simplified by using a fixed input size. Third, this resolution provides a reasonable compromise between the quality of the representation and the efficiency of the computation which is crucial to the real-time framing of the thesis.

The preprocessing pipeline of validation was deliberately a simple pipeline. All validation images were resized to 224 × 224, converted to the form of tensors, and normalised with values of ImageNet mean and standard deviation. The purpose of the validation pipeline was to provide a

stable and deterministic evaluation setting. None of the stochastic augmentation was performed during validation since such manipulations would introduce the unnecessary randomness to the reported results and makes the comparison between the experiments less reliable.

The training preprocessing pipeline was similar to the basic resizing and normalisation steps, but with a small number of augmentations to enhance generalisation. In particular, the training images were resized to 224 × 224, randomly flipped horizontally with probability 0.5, and subjected to light colour jittering, converted to tensor format, and normalised using ImageNet statistics. Horizontal flipping was also added since most facial expressions are semantically equivalent when flipped left-right, and thus, the operation provides increased data variation but does not change the class label. Limited variability in appearance in terms of brightness, contrast, saturation, and hue, was introduced by light colour jittering, where the model is required not just to be insensitive to changes in illumination and colour variation, but also to maintain the facial structure unchanged.

The colour jitter parameters were purposely maintained moderate. This decision is an indication that facial expression recognition is not parallel to coarse object recognition. In FER, the discriminative information is frequently contained in subtle configurations of facial muscles as opposed to strong object-level shifts in shape. Thus, severe affine distortion, massive occlusions, extreme blurring or excessive aggressive erasing were not part of the core protocol of controlled procedures. The objective was to enhance robustness without any synthetic distortions that can corrupt any expression-related cues or may bias the comparison between variants of the model.

The preprocessing approach was maintained constant over the large experiments in order to maintain internal comparability. It implies that any observed differences in the main results should not be attributed to differences in augmentation recipes, but to the intended experimental factors, like choice of backbone, hybrid refinement, focal loss, class-aware alpha weighting or stability-focused training mechanisms. A fixed preprocessing pipeline was thus a significant aspect of the controlled experimental design outlined in the preceding section.

The next factor that makes it reasonable to maintain the augmentation moderate is the very type of the AffectNet-8 data. The training data is already rich in large natural variability in pose, lighting, scale, and occlusion, which inherently gives the training data a lot of diversity. The purpose of augmentation in such a setting is not to make extreme artificial conditions, but to add little additional variation, which helps reduce overfitting but preserves the realism of the appearance of a face. This is the reason that the final preprocessing design of the thesis can be interpreted as the carefully conservative augmentation strategy designed to recognize expressions rather than the maximal augmentation strategy that is designed to be generic in recognizing images.

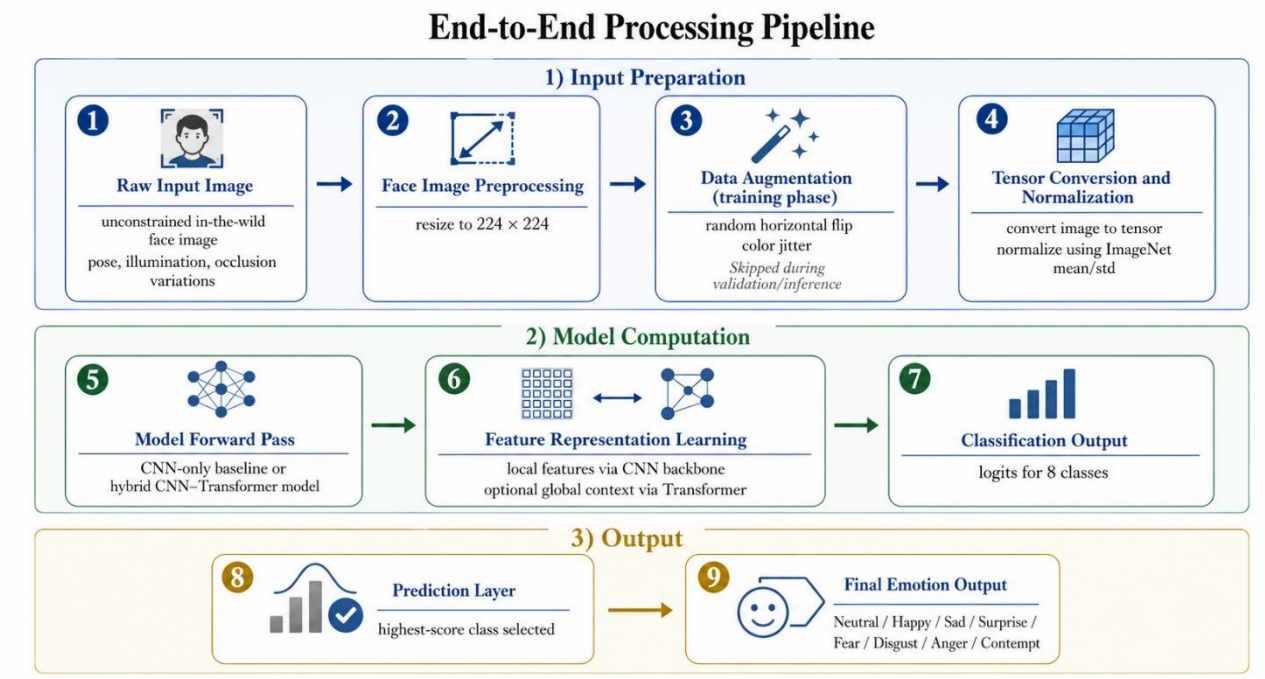

Fig. 2. End to End Processing Pipeline

The processing pipeline end-to-end processing pipeline utilized in this work is shown in Fig. 2. The first step is to resize and normalise a raw facial image followed by data augmentation, which is only performed during training. The processed image is subsequently sent through the chosen model whereby discriminative features are extracted and converted into logits of the classes. Lastly, the label of the predicted facial expression is taken out of the highest-confidence output.

To conclude, the preprocessing and augmentation pipeline utilized in this thesis was constructed on the principles of consistency, efficiency and task-appropriateness. All images were standardised to a common input size; validation was deterministic; and training used only limited augmentation that is unlikely to corrupt subtle signals of facial-expressions. This design not only facilitated a fair comparison of models but also allowed the latter analysis to focus on the actual effects of architecture and training strategy, as opposed to uncontrollable changes in data preparation.

2.5. Baseline model designs

The models used as the baseline in this thesis were chosen to provide an effective reference point on which to assess more advanced training strategies and hybrid CNN Transformer refinements. The goal of the baseline stage was not just to find the initial levels of recognition and computational performance to be used in further experiments. The baseline architectures were ResNet50 and ConvNeXt-Tiny.

ResNet50 was used as the first baseline. The reason why this model was selected is that, residual convolutional networks continue to be one of the most widely used and most understood reference architectures in visual recognition. ResNet50 provides a good baseline to compare with when it comes to controlled comparison because it has an established training behavior, a moderate cost in terms of computational power, and has been widely used in previous FER and general image classification experiments. In this thesis, ResNet50 was adapted for eight-class facial expression recognition by replacing the original final classification layer with a task-specific output layer corresponding to the AffectNet-8 label set. The baseline therefore retained the convolutional feature

extraction capacity of the original network while aligning the classifier head with the target problem [7].

The second baseline was based on ConvNeXt-Tiny. This model was selected as the stronger modern convolutional alternative in the study. Compared with classical residual CNNs, ConvNeXt-Tiny offers a more recent convolutional design that has shown strong representational capacity while remaining relatively efficient in practice. In the context of this thesis, ConvNeXt-Tiny was important because it allowed the evaluation to move beyond a classical CNN reference and test whether a stronger convolutional backbone alone could yield better balanced performance under the challenging in-the-wild and imbalanced AffectNet-8 setting. As in the ResNet50 baseline, the final classification head of ConvNeXt-Tiny was adapted to produce predictions for the eight target expression categories [13].

Both baseline models were implemented as single-image facial expression classifiers. Their task was to take an input facial image of fixed size 224 × 224 and assign it to one of the eight predefined expression classes. In both cases, pretrained backbone weights were used as the starting point of the model initialization. This decision was made to improve optimization stability and to provide a fair and practical transfer-learning setting, which is common in image-based deep learning tasks where the target dataset is challenging and the backbone architectures are originally designed for large-scale visual recognition.

In order to maintain a fair comparison, the two baselines were integrated into the same training and evaluation framework. They used the same dataset protocol, the same input resolution, the same basic preprocessing logic, and the same performance metrics. This made it possible to interpret differences between the baselines primarily in terms of backbone design rather than in terms of unrelated implementation factors. The baseline stage therefore served as the foundation for the later hybrid experiments and for the subsequent ablation studies on imbalance-aware training.

A further reason for including both ResNet50 and ConvNeXt-Tiny baselines was methodological. ResNet50 was a powerful classical CNN reference with more advantageous speed properties, whereas ConvNeXt-Tiny was a more modern convolutional architecture with more favorable speed characteristics. This combination allowed two types of comparisons in the latter parts of the thesis. First, it allowed the study to know which CNN backbone to use as the primary reference when refining hybrids. Second, it allowed the analysis of an accuracy-efficiency trade-off between a faster classical baseline and a more heavy but powerful modern CNN model.

The base models were also very crucial in the interpretation of the contribution of subsequent enhancements. It would not be possible to ascertain whether gains occurred due to hybrid modeling, due to loss functions that are imbalance-aware, or due to training stabilization methods. Thus, the designs of the baseline models were purposely reduced to simple and standardized designs. They were not to achieve maximum complexity, but to provide strong and interpretable reference points to the entire thesis workflow.

To sum up, the baseline phase of the thesis was comprised of two CNN-only models, the ResNet50 and the ConvNeXt-Tiny. ResNet50 was used as a classical, efficiency-oriented baseline, whereas ConvNeXt-Tiny was a more effective modern convolutional baseline. Both models were modified to the eight-class task of AffectNet-8 and trained under the same controlled condition. The outcome of this step subsequently explained the decision to use ConvNeXt-Tiny as the primary backbone to the hybrid experiments.

2.6. Hybrid model designs

The second step of the study was to examine whether a lightweight hybrid CNN + Transformer design would lead to an improvement in facial expression recognition performance under the same controlled protocol as the CNN-only baselines. The hybrid model design was built on top of the ConvNeXt-Tiny backbone, which had already shown stronger balanced performance than the ResNet50 baseline. The goal of the hybrid stage was not to build a very large or highly specialized Transformer architecture, but to introduce a compact contextual refinement module while preserving practical efficiency and relevance to the real-time thesis framing.

The general idea of the hybrid design was to combine the local hierarchical feature extraction strength of a convolutional network with the contextual modeling capability of Transformer encoder layers. In facial expression recognition, convolutional backbones are effective at extracting local visual patterns such as edges, textures, facial contours, and localized deformations around the mouth, eyes, and eyebrows. However, the interpretation of facial expressions may also depend on the interaction between multiple facial regions. The hybrid design was therefore motivated by the assumption that a lightweight Transformer refinement stage could provide useful global context after the convolutional backbone had already extracted strong visual features [4], [16].

In the implemented hybrid architecture, ConvNeXt-Tiny was used as the primary feature extractor. Instead of performing classification directly from the backbone output, the final spatial feature representation of ConvNeXt-Tiny was converted into a sequence of tokens. These tokens were then projected into a representation suitable for Transformer-based processing. After this projection stage, one or more Transformer encoder blocks were applied to refine the token sequence through contextual interaction. The output token representation was then aggregated using mean pooling and passed to the final classifier head to produce the eight-class prediction.

A key design principle of the hybrid model was efficiency. The purpose of the Transformer component being lightweight was to ensure it did not impose too much computational overhead. Instead of scaling up the Transformer stack depth or scaling up the size of a vision-transformer branch the thesis tested compact refinement modules with limited depth. This allowed examining whether any amount of Transformer-based contextual modeling could enhance the convolutional baseline. It also made sure that the resultant hybrid models were consistent with the real-time and efficiency-conscious goals of the thesis.

The experiments of hybrid were arranged in the controlled order. The earliest hybrid form featured only one Transformer encoder block, which was the most simple hybrid refinement. This model was the first experiment of whether architecture alone can yield quantifiable benefits beyond the CNN-only ConvNeXt-Tiny baseline. Following this preliminary finding, a second variant, with two blocks of Transformer encoders, was tested. The rationale behind this extension was to find out whether the first hybrid had been too shallow and whether a slightly deeper contextual polishing would help increase performance without necessarily increasing the weight. Critically, the change between one block to two blocks was viewed as an explicit check of fairness and not the entire architecture redesign.

In terms of the experiment, the hybrid model design was to provide the answer to three questions. First, does the introduction of Transformer-based contextual refinement enhance the quality of recognition above that of CNN-only baseline? Second, to what extent can the hybrid design be trained to respond to training strategy in the strongly imbalanced AffectNet-8 condition? Third, are the extra contextual modeling worth the extra parameter count and inference cost that the Transformer blocks introduce? The reasons why the thesis did not find a single hybrid

implementation, but instead discussed several hybrid variants and stabilization strategies within the same broader framework can be explained through these questions.

Another critical observation is that the hybrid models in this thesis are to be understood as lightweight experimental architectures as opposed to maximal-complexity state-of-the-art systems. The objective was to run an experiment to determine whether a useful and efficient hybrid refinement could enhance the backbone in a controlled environment. This difference is significant in terms of the interpretation of the final results. In a scenario where the hybrid is competitive, but not higher than the strongest CNN-only model, the outcome is also informative as it enlightens about the effect size of the lightweight Transformer refinement compared to the training strategy and the computational cost.

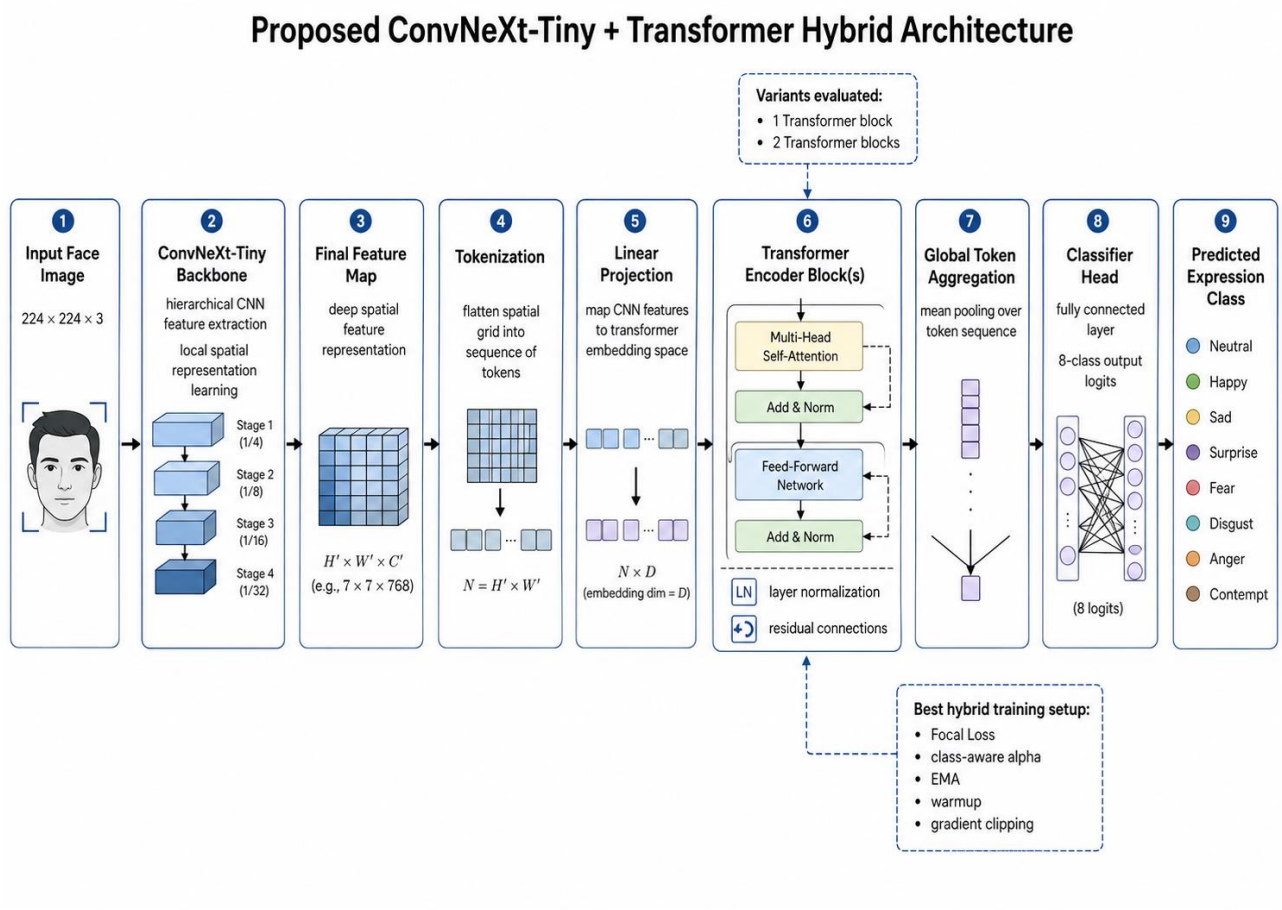

Fig. 3. Hybrid Model Architecture

The proposed hybrid architecture as proposed in this study is shown in Fig. 3. The model uses an efficient local feature extraction via a ConvNeXt-Tiny backbone with a lightweight Transformer encoder to provide global contextual modeling. The resulting spatial feature map of the CNN backbone is transformed into a sequence of tokens, which are projected into an embedding space, and processed by one or more Transformer encoder blocks, before being classified into eight categories of facial expression.

Overall, the design of the hybrid model design in this thesis was grounded on ConvNeXt-Tiny feature extraction and token conversion, projection, Transformer-based contextual refinement, and token aggregation with final eight-class classification. It was deliberately made lightweight, and was tested in both one-block and two-block versions. This architecture family formed the core of the hybrid investigation in the thesis and enabled a controlled assessment of whether compact

Transformer refinement can improve in-the-wild facial expression recognition under strict efficiency and reproducibility constraints.

2.7. Training strategies

The training strategies used in this thesis were designed to support two main goals: first, to provide fair and controlled comparisons between model architectures, and second, to improve recognition stability under the strong class imbalance of AffectNet-8. Since the dataset contains a large difference in class frequencies between majority and minority expressions, the choice of loss function and optimization procedure plays an important role in the final performance. For this reason, the study did not rely on a single training configuration, but instead examined a sequence of training strategies in a controlled manner.

The starting point for the experiments was a standard supervised classification setup based on cross-entropy loss. Cross-entropy was used in the initial baseline experiments because it provides a well-understood and widely adopted reference point for image classification. Using this loss in the first stage of the study made it possible to evaluate the effect of the backbone and the lightweight hybrid refinement before introducing more specialized imbalance-aware modifications In this respect, the concept of cross-entropy was used as the neutral reference point that subsequent improvements could be measured.

Nonetheless, the gross imbalance of classes in AffectNet-8 renders the cross-entropy inadequate to conduct a complete examination of the issue. With standard cross-entropy optimization, many of the majority-expressions are often dominant in the learning process, which can result in weaker recognition of the underrepresented expressions like disgust and contempt. To solve this problem, the thesis presented focal loss as the alternative that takes into account the imbalance. The focal loss decreases the relative contribution of examples that are well-classified and increases the focus on harder or misclassified examples. This is what makes it especially well-suited to datasets in which the classifier might otherwise over-fit to the majority-class patterns. It is that which makes it especially well-suited to datasets in which the pattern of majority-class might otherwise over-adapt to the classifier [25].

Besides focal modulation, the thesis also used the class-aware alpha weighting. The alpha values were obtained by observing training set distribution using an inverse-square-root frequency plan and then normalizing the resultant frequencies so that the average of the resulting frequencies would be near 1. This method maximized the input of minority classes whilst not overly extreme weights of classes. The combination of focal loss and class-aware alpha was used as the main imbalance-aware training strategy in the later experiments and became one of the strongest factors contributing to performance improvement in the study [25].

The training experiments also investigated stability-focused optimization strategies. In the first experiments of hybrid Transformers, it was found that certain Transformer-based variants would achieve competitive validation performance early in training, but then would degrade or become less stable as training progressed. This action led to the establishment of more stabilization measures The exponential moving average (EMA) of model parameters was the most important of these. EMA maintains a smoothed version of the model weights over training iterations and can improve generalization by reducing sensitivity to short-term optimization noise. In the strongest hybrid training configuration, the EMA model rather than the raw training weights was used for final evaluation.

A second stabilization mechanism was learning-rate refinement. In the later training stages, the learning rate was reduced compared to the earlier experiments in order to make optimization gentler

and reduce the risk of rapid overfitting or unstable updates. This was intertwined with brief linear warmup period prior to cosine decay. The rationale behind the warmup was to prevent the early-optimization dynamics, especially in hybrid models, where the interaction between pretrained convolutional features and newly trained Transformer components can cause the early stages of training to be more sensitive. After the warmup period, cosine decay was used to gradually reduce the learning rate during the rest of training.

Gradient clipping was introduced to the stability-based training package as well. It was designed to avoid excessively large updates in the course of backpropagation and to enhance numerical stability, particularly in experiments with the hybrid CNNTransformer model. The last configuration of stability-enhanced was a clipping threshold of 1.0. This change was maintained in a simple form and not intended to be a separate contribution, but as a practical protection that augmented EMA and warmup scheduling.

Throughout the experiments, the family of optimizers was the same. All key runs of training were done with AdamW since it offers useful adaptive optimization and decouples the weight decay and gradient update. This enabled it to be used in ensuring fair comparison between various backbones, losses and hybrid variations. Early stopping was used with a fixed patience parameter to minimize unnecessary training when performance is decreasing and to identify the most robust validation checkpoints without using visually selected epochs. The implementation also used mixed-precision training, which would help enhance the computational efficiency without modifying the core model design [32].

Experimental discipline was an important consideration in the design of the training strategy. The research did not necessarily set out and at the same time, put many new variables into the research. Rather the training strategy was developed in phases. The models were initially compared in cross-entropy. Second, the focal loss and class-aware alpha were added to evaluate the impact of imbalance-aware training. Third, EMA, a reduced learning rate, using warmup schedules, and clipping gradients were introduced to evaluate whether better training stability could be used to further improve the most robust hybrid variants. This simulated development was critical towards segregating the effects of architecture versus the effects of optimization.

Overall, training strategies in this thesis included a gradual increase of standard cross-entropy training to focal loss with class-aware alpha, and finally more stability-oriented optimization with EMA, warmup, a lower learning rate, and gradient clipping. This design was not only practical, but also analytically significant in that it facilitated the demonstration that the greatest verified gains in the study were due largely to imbalance-conscious and stability-oriented training as opposed to to architecture refinement alone.

2.8. Evaluation metrics

The evaluation plan of this thesis has been crafted to show the quality of recognition and applicability to real-life deployment. Given that the task is facial expression recognition on an imbalanced in-the-wild dataset, no one metric would suffice to make a meaningful interpretation of model behavior. This is why the thesis relies on a system of complementary measures that reflect the overall correctness, balanced multi-class performance, class-specific performance, and computational efficiency.

The initial main measure is classification accuracy. Accuracy is the ratio of the number of correctly identified images in the validation set, and can be used to provide an intuitive global picture of recognition performance. It is an effective overall measure of the quality of the model and it is one of the most widely reported measures within the FER literature. But in the case of AffectNet-8,

accuracy is not sufficient since the distribution of classes in the training set is highly skewed. A model can have a reasonably high accuracy by focusing on majority classes and yet a very low accuracy on minority classes. Consequently, despite the steadiness in the accuracy reports in this thesis, it is not applied as the sole criterion in selecting or interpreting a model.

This is why the second key measure is the macro- F1. The F1-score is a harmonic-mean combination of precision and recall, and macro-F1 is the harmonic-mean average of the F1-score across all the classes. This implies that the weight of each expression class to the final value is the same irrespective of the frequency of occurrence in the training data. Macro-F1 is thus especially suitable to AffectNet-8, where the frequency of classes like disgust and contempt is much lower than that of happy or neutral. In this thesis, macro-F1 is considered as one of the primary model selection criteria since it reflects more reliably balanced recognition quality than accuracy in the presence of severe class imbalance.

These main summary measures are supplemented by class-wise measures to give a more detailed interpretation of model behavior. Precision, recall, and F1-score can be calculated based on the confusion matrix, by class. Of these, the per-class F1 is especially useful, as it shows which expression categories can be recognized reliably and which ones are hard to recognize. This is of particular concern in AffectNet-8, where certain classes are already known to be substantially harder than other classes. The thesis employs the results per-class to interpret the behavior of both the best CNN-only model and the best hybrid model, with special focus on minority and difficult classes such as fear, disgust and contempt.

Another significant aspect of the evaluation is confusion matrices. A confusion matrix captures the frequency with which each true class is classified as each possible class, and thus allows the structure of classification errors to be viewed. As part of this thesis, raw and normalized confusion matrices were created as part of the evaluation. The raw confusions matrix retains absolute counts but the normalized version is much more useful in the visual comparison between classes due to its ability to express each row relative to the number of samples in that class. Normalized confusion matrices are especially appropriate to discuss the most appropriate CNN-only and the best hybrid models in the final thesis presentation.

In addition to the quality of recognition, the thesis also measures the computational efficiency. Because the work is expressly framed in terms of real-time or near-real-time recognition of facial expression, model comparison should not only cover measures of accuracy, but also inference cost. To this end, three measures of efficiency are employed; the number of parameters, latency per image, and frames per second (FPS). The number of parameters is a succinct measure of the size of the model and the complexity of its architecture. Latency per image (time in milliseconds) is the time spent by a single forward pass of a fixed batch-size-one benchmarking protocol. FPS is a throughput measure, and is derived using the same latency value. Combined, these measures aid in the analysis of the practical trade-off between the recognition quality and the inference efficiency.

The efficiency measures are viewed in a fixed benchmarking environment, instead of being regarded as universal device-independent properties. Measurement of both reported values of latency and FPS were measured with identical input resolution, batch size, hardware type, and repeated inference procedure. Therefore, their main role in the thesis is comparative: they make it possible to compare the relative efficiency of the completed models under one controlled environment. This is especially relevant when comparing CNN-only and hybrid architectures, since the latter may introduce a modest increase in parameter count but a more noticeable increase in inference latency.

Finally, the thesis also distinguishes between single-model evaluation and inference-side enhancement. The main real-time conclusions are based on single-model inference, because this is the fairest representation of a deployable real-time model. However, an additional offline evaluation was later performed using checkpoint ensembling and horizontal-flip test-time augmentation for the strongest hybrid models. This additional evaluation was useful for estimating the upper practical potential of the hybrid approach, but it is not treated as the main deployment-oriented result because ensemble and TTA configurations increase inference complexity.

In summary, the evaluation framework of this thesis combines overall accuracy, macro-F1, per-class behavior, confusion matrices, parameter count, latency, and FPS. This mix was chosen to make sure that the quality of classification as well as the practical efficiency are reflected in the analysis. In this context, the primary recognition metrics are accuracy and macro-F1, whereas the per-class recognition results and confusion matrices are used to enable detailed interpretation, and the count of parameters, latency and FPS are used to support the claims of the study about real-time and efficiency.

2.9. Real-time benchmarking protocol

The focus of the thesis is clearly stated as being in the area of practical real-time or near real-time facial expression recognition. Thus, in the present study, evaluation of the models is not based solely on recognition quality, but it also involves computational efficiency based on a fixed inference protocol. The term real-time here is not intended to be misinterpreted and does not mean a vague notion of latency in general, but rather low latency single image inference as measured in a controlled GPU benchmarking setup.

Main training experiments completed prior to benchmarking procedure. For each experiment, the best checkpoint was loaded, the model architecture was reconstructed and the model was put in evaluation mode. For the main real-time comparison, only single-model checkpoints were used as this stage was designed to simulate real-time practical deployable inference. Ensemble and test-time augmentation configurations were not considered as primary real-time configuration, but as offline inference improvements.

For all benchmark measurements the KTU Jupyter GPU environment with CUDA support was used as the computational environment, similar to that used for the experiments in the thesis. The models were tested on the NVIDIA H100 NVL GPU. Mixed-precision inference was used to represent a realistic deployment scenario. The input resolution in all the models compared was 224×224 pixels, which is the same resolution as was used during the main experiments. The batch size was set to 1 for the thesis, which deals with inference based on a single image in real time rather than high throughput batch operation.

Benchmarking was comprised of two phases: warm up and timed measurement. The model carried out 50 forward passes in the warmup period without any timing results. This decreased the initial instability with GPU calls and guaranteed that measured times were pertaining to the steady-state inference behavior. Following warmup, 200 passes were timed forward with the same conditions. During timing, GPU synchronization was used to ensure that wall clock time measured was due to completed CUDA inference operations and not the queued asynchronously.

Latency was measured as the mean wall-clock time for each forward pass of the batch of size 1 and reported in milliseconds per image. The same protocol was used to obtain the FPS, which indicates the number of images that can be processed per second in the same set-up used for the benchmarking. The latency and FPS numbers reported in this thesis must thus be seen mostly as

relative to the model in question and the environment in which it was measured and not as general deployment guarantees for the platform.

The number of trainable parameters was another value that was benchmarked. The number of parameters (in both absolute and millions) was reported. While the number of parameters is not necessarily sufficient to characterize the practical inference speed of a model, it is a second indicator of the size of the model and how complex its architecture is. This thesis is very helpful for understanding the extra cost added by lightweight Transformer refinement, relative to CNN-only baselines.

Main benchmark comparisons are only for the completed single-model experiments: ResNet50 baseline, ConvNeXt-Tiny baseline, various ConvNeXt-Tiny plus Transformer models, and the top models that were stabilized during training. We discuss an offline enhancement of best hybrid inference-side rescue setup based on horizontal-flip test-time augmentation and ensembling at checkpoints separately. It is not considered as the primary real-time configuration due to the ensembling and the TTA, which make the inference more complex.

This benchmarking protocol provides for the real-time interpretation of thesis in two ways. First, it allows to make a meaningful comparison of CNN-only and hybrid architectures in the same practical conditions. Second, it enables the study to answer the question of whether there is an acceptable or excessive inference overhead cost. This is particularly relevant since the hybrid models had only a modest number of additional parameters; however, a more significant latency penalty compared to the best CNN-only baseline.

To sum up, real-time evaluation, used in this thesis, is achieved with a fixed single-image GPU inference protocol, with 224 × 224 input resolution, a batch size of one, running with mixed precision, 50 warmup iterations and 200 iterations of the forward pass in the same hardware environment. To offer a practical and repeatable sense of model efficiency, latency, FPS and parameter count are reported concurrently. This protocol will help ensure that the claims made in the thesis are real-time and based on measurable evidence instead of qualitative assumptions.

2.10. Reproducibility settings

The concept of reproducibility was addressed as a significant need in the experimental process of this thesis. This was not only to achieve competitive results but also to make sure that the experiments were conducted in a fixed, well documented and repeatable set up. It is due to this fact that the study has had a stable training and validation protocol, random seed control, common preprocessing logic and a common implementation framework across the major experiments. In this section, the main reproducibility settings in which the work was created are summarized.

The dataset protocol was determined at the start of the research and maintained throughout all the main experiments. Experiments were done on the complete AffectNet-8 training split and the balanced validation split with CSV-based indexing. The training data were loaded using train index full.csv, and the validation data using val index full.csv. The class mapping was fixed over the course of the work: 0 - neutral, 1 - happy, 2 - sad, 3 - surprise, 4 - fear, 5 - disgust, 6 - anger, and 7 - contempt. The study maintained the same split definition and label mapping across all of the runs in order to ensure that internal model comparisons were done under the same, consistent protocol.

Random seed control was also applied in the implementation. The seed value that was used in the main training scripts was 42. The seed was used to initialize Python-level, NumPy-level, and PyTorch-level randomness in the training pipeline. This seed control significantly minimized unnecessary run-to-run variability and made the experimental process more predictable and

understandable, although still possibly dependent on the behavior of low-level libraries and the particular characteristics of the hardware.

The input setting was made uniform across the experiments. All the primary models accepted 224 x 224 pixels as input. The identical basic preprocessing and normalization pipeline was used across all major experiments and the augmentation strategy was held constant during controlled comparisons unless a change in augmentation was the desired experimental factor. All core experiments were deterministic with regard to the validation pipeline. One of the primary reasons why the results in the thesis can be viewed as controlled comparisons as opposed to loosely related runs is this fixed data pipeline.

The family of optimizers was always maintained through the core experiments. AdamW was taken as optimization procedure during the core training process. The same general framework was also used in the baseline experiment and the hybrid experiment in terms of checkpointing, logging and early stopping. Early stopping patience was to be 4 validation tests in the main training scripts. This avoided excessive overtraining and made sure that the best checkpoints selected were those that corresponded to observed validation performance as opposed to arbitrary final epochs [32].

The loss functions and training settings were implemented in a progressive and trackable way. Experiments in initial reference employed cross-entropy loss. Focal loss and class-aware weighting of alpha were added later to deal with AffectNet-8 imbalance in classes. The experiments on the focal loss took a gamma value of 2.0. The class-sensitive alpha values were calculated using the observed distribution of the classes with inverse-square-root frequency weighting and normalized so as to maintain a mean close to 1. The final alpha values used in the main focal-loss runs were: 0.403369, 0.301054, 0.691747, 0.929849, 1.382056, 1.789801, 0.699721, and 1.802404 for classes 0 through 7, respectively.

The CNN-only model used in the thesis, with focal loss and class-aware alpha, and based on the ConvNeXt-Tiny baseline architecture with image size 224, batch size 128, optimization with AdamW, learning rate 1e-4, weight decay 1e-4, cosine annealing scheduling, mixed precision training, early stopping and the focal-plus-alpha loss setting described above. The most powerful hybrid single model, i.e. the ConvNeXt-Tiny + lightweight Transformer refinement stage with EMA, used the same dataset protocol and loss settings, except that it added a two-block Transformer refinement stage with EMA, a reduced learning rate of 5e-5, a short linear warmup then followed by cosine decay, and gradient clipping with a threshold of 1.0. These settings are clearly recorded since they are related to the overall final models that will be addressed in the thesis.

To further facilitate reproducibility, implementation was structured into reusable Python modules to load dataset, define models, loss functions, evaluation, confusion matrix generation, training and benchmarking. Systematic storage of model checkpoints, validation metrics, predictions and benchmark summaries were stored in the project results directories. The organizational structure enabled verification of important results post training and minimized the chances of unrecorded manual processes when evaluating results.

It is also important to note that the primary benchmark protocol of latency and FPS was fixed without training. All the benchmarked models were run with an input resolution of 224 x 224, batch size of 1, 50 warmups, and 200 timed forward passes using mixed-precision inference in the same CUDA-based environment. The fact that a common benchmark protocol is used is also an important aspect of reproducibility since it also means that efficiency comparisons across models are based on the same measurement logic.

Overall, this thesis exhibits reproducibility through the use of fixed data splits, fixed label mappings, seed control, and standardized preprocessing, a consistent optimizer and evaluation framework, explicit documentation of final hyperparameters, and systematic checkpoint and result storage. These measures do not give a complete elimination of the sources of variation in the deep learning experiments, but they provide a clear and sufficiently rigorous basis to reproduce the main findings of the study in the same computational environment.

3. Experimental Results and Efficiency Evaluation

3.1. Implementation environment

The experiments and implementation mentioned in this thesis were conducted in the KTU Jupyter computing platform. The experiments were implemented using deep learning libraries (accelerated in a graphical processing unit, or GPU) to be trained and evaluated on both CNN-only and CNN+Transformer models using the full AffectNet-8 data set within a realistic time constraint. PyTorch, along with torchvision, was the primary software framework used to develop the models. These libraries were chosen as they offer a stable support to pretrained convolutional backbones, flexible model customization, mixed precision training, and reproducible evaluation workflows.

The core experiment hardware environment consisted of NVIDIA H100 NVL GPU resources that were supported by CUDA. The model training was conducted on this environment, as well as the following efficiency benchmarking step. The thesis needed to be based on GPU acceleration since the project involved repetitions of experiments on a large, imbalanced training set and measured latency and FPS in a realistic inference configuration. Single-image inference measurements were performed on the GPU using mixed precision to indicate a realistic, high-performance test environment.

It was implemented in the form of a modular research codebase. Data loading, definition of baseline models and hybrid models, loss functions, evaluation logic, confusion matrix generation, training scripts, and benchmarking scripts were in separate python modules. It provided a simpler control of the experimental process and minimized the possibility of discrepancies between runs. It also enabled the same evaluation and logging logic to be reused in many experiments, which was significant in terms of maintaining the fixed protocol in the previous chapter.

The pipeline of loading the dataset was founded on CSV indexing files of the training and validation splits. Each sample was indexed with an image path and an expression label and the same indexing logic was reused across all the major runs. This would simplify control of the AffectNet-8 data splits and allow easier control of maintenance of a reproducible workflow between baseline experiments and the final benchmarking stage. Use of CSV based indexing also enhanced transparency in checking class distributions and in checking whether the train and validation settings had been held constant.

In the primary experiments, mixed precision was used to train a model. This decreased memory consumption and enhanced efficiency of training without compromising the overall framework of the optimization process. Checkpoint saving, logging and evaluation have been directly incorporated into the training workflow. Each experiment had the latest checkpoint stored in the framework, the best checkpoint (in terms of validation performance) stored in the framework, and the validation predictions, the JSON summaries of the results, the confusion matrices in both numeric and graphical formats, and the log files containing the epoch-by-epoch evolution of the key metrics. Such a systematic storage of outputs proved especially handy in the future when verifying the reported results and putting the final thesis tables together.

The training strategies which were based on stability were also supported by the environment of implementation. In the subsequent experiments, this framework was expanded to include focal loss, class-aware alpha weighting, exponential moving average (EMA), gradient clipping, and learning-rate scheduling with warmup and cosine decay. All these features of training were introduced to maintain comparability of experiments and to permit controlled transitions of simpler to more complex training setups. To achieve the best combination between the two models, analysis was done using the EMA-smoothed weights that were stored in the course of training.

It was also found necessary to develop a separate evaluation-side script that is then used during the final inference-stage hybrid rescue attempt. It is a script that facilitates checkpoint ensembling and horizontal-flip test-time augmentation without re-training the model. Moreover, a special benchmarking script was developed to re-create the trained models based on their saved checkpoints and to measure the number of parameters, latency and FPS under a fixed single image inference protocol. This enabled the thesis to be able to support not just recognition-oriented conclusions, but also efficiency-oriented conclusions based on measured evidence.

In general, the implementation environment of the thesis may be characterized as a structured and experiment-oriented deep learning workflow implemented on PyTorch, torchvision, CUDA, and executing on a GPU. A notable aspect of the system was its modular design, reproducibility, rigorous evaluation controls and well organized storage of both training and inference outputs. These properties were significant in making sure that the final conclusions of the thesis were based upon the verified and consistently produced results as opposed to the isolated exploratory runs.

3.2. Baseline experiments

The initial step to the empirical part of the thesis was the baseline experiments. They were meant to achieve robust CNN-only reference points prior to introducing hybrid CNNrefinement and Transformerrefinement, and even more advanced imbalance-aware training strategies. The two baseline architectures were tested: ResNet50 and ConvNeXt-Tiny. They were conducted under identical training and validation conditions, allowing direct comparison of the behavior of the backbone in the AffectNet-8 setting.

The original baseline experiment involved the use of ResNet50 which was trained using the standard cross-entropy loss. The classical convolutional methods used in the study used this model as the reference. ResNet50 had a validation accuracy of 0.5326 and a validation macro-F1 of 0.5026. Efficiency wise it was also the fastest single model to complete in the study with 23.524 million parameters, average latency of 3.4479 ms per image and a measured throughput of 290.03 FPS. These findings confirmed that ResNet50 is an effective and practical baseline, yet they also indicated that there are limitations to balanced recognition performance with ResNet50 under the strong AffectNet-8 class imbalance.

Further examination of the ResNet50 baseline revealed that the model was performing fairly well on most of the classes and weak on the most challenging minority classes. Specifically, the weakest type of class during the first stage of the baseline was contempt. This was a significant observation as it was early evidence that the study could not be solely based on accuracy and that imbalance-conscious evaluation and training would most likely be required. Thus, the ResNet50 baseline served two purposes in the thesis: to provide an efficiency-oriented CNN baseline and to emphasize the shortcomings of standard cross-entropy training in the presence of an extremely high imbalance between classes.

The second baseline model used ConvNeXt-Tiny that was also initially trained using cross-entropy loss. This model had a validation accuracy of 0.5259 and validation macro-F1 of 0.5048. Though its accuracy was a bit lower than the ResNet50 baseline, its macro-F1 was a bit higher. This distinction is significant since one of the primary model selection criteria in the thesis was macro-F1 due to the imbalance between classes in AffectNet-8. Efficiency wise, ConvNeXt-Tiny had 27.826 million parameters, average latency of 4.0567 ms per image and a FPS of 246.51.

The baseline comparison showed that ConvNeXt-Tiny offered a stronger balance of class-sensitive performance even though it was somewhat slower and larger than ResNet50. In particular,

ConvNeXt-Tiny showed stronger behavior on several difficult classes, including fear, disgust, and contempt, while ResNet50 retained a speed advantage. This result was important for the later stages of the thesis because it justified selecting ConvNeXt-Tiny as the primary backbone for hybrid experiments. If ResNet50 had been clearly superior on both accuracy and macro-F1, it would have remained the main backbone candidate. However, the baseline results indicated that ConvNeXt-Tiny was the more suitable choice for the subsequent architecture and training investigations.

The first insight of the thesis into the accuracy-efficiency trade-off was also provided by the baseline experiments. ResNet50 showed that a small, faster model can serve as a powerful efficiency-oriented baseline, but the ConvNeXt-Tiny baseline indicated that a modest increase in computation cost can be used to scale-up balanced recognition performance. This trade-off would be a common motif in the remainder of the study. The baseline stage found the two comparison poles that would be significant later; not only a faster CNN efficiency backbone; but also a stronger balanced recognition backbone.

The other significant purpose of the base experiments was a methodological one. They established the reference points which would be used in the correct interpretation of subsequent improvements. It could not have been determined whether later improvements were as a result of the hybrid architecture, use of focal loss, class-sensitive alpha weighting or training stabilization without baseline results. The baseline stage thus formed the empirical basis to the subsequent ablation analysis of the thesis.

Overall, the baseline experiments led to three important conclusions. First, ResNet50 is the fastest of the completed single-model configurations and therefore serves as a strong efficiency-focused reference. Second, ConvNeXt-Tiny provides slightly stronger balanced performance than ResNet50 in the initial CNN-only comparison and is therefore the more suitable backbone for later refinement. Third, the baseline results already reveal the practical importance of evaluating minority-class behavior and macro-F1 rather than relying on accuracy alone. These conclusions directly motivated the next stage of the thesis, namely the evaluation of lightweight hybrid CNN–Transformer variants and the investigation of imbalance-aware training strategies.

Experiment	Model	Accuracy	Macro-F1	Params (M)	Latency (ms)	FPS
baseline_resnet50_ce_run1	ResNet50 baseline	0.5326	0.5026	23.524	3.4479	290.03
convnext_tiny_baseline_ce_run1	ConvNeXt-Tiny baseline	0.5259	0.5048	27.826	4.0567	246.51

Table 3. Baseline CNN model comparison

3.3. Hybrid architecture experiments

After establishing the CNN-only baselines, the next stage of the study investigated whether lightweight hybrid CNN–Transformer refinement could improve facial expression recognition performance under the same controlled protocol. The hybrid experiments were based on ConvNeXt-Tiny, since the baseline comparison had shown that this backbone provided slightly stronger balanced recognition performance than ResNet50. The purpose of the hybrid stage was to test whether adding Transformer-based contextual modeling on top of the stronger CNN backbone would lead to measurable gains in validation accuracy and macro-F1.

The initial hybrid experiment had a lightweight ConvNeXt-Tiny + Transformer architecture that was trained with standard cross-entropy loss. This model was a simplest hybrid refinement in the thesis. The last spatial features of the ConvNeXt-Tiny backbone were transformed into tokens, projected to an appropriate representation, processed by a Transformer encoder, and then aggregated followed by final classification. In this original training setup, the hybrid model obtained the following validation accuracy of 0.5239 and validation macro- F1 of 0.5009. These values put it just under both completed CNN-only baselines, so only a lightweight hybrid architecture could provide a definite improvement.

The fact that the initial hybrid outcome was not better as a whole, did not make it any less informative. It indicated that merely putting a Transformer refinement block on top of a robust convolutional backbone does not necessarily enhance expression recognition in the AffectNet-8 context. Meanwhile, the behavior at the level of the classes showed that the hybrid model was not a completely unpromising one. Specifically, it demonstrated indicators of a better performance on challenging minority-class recognition tasks, which inspired further research. Consequently, the fact that the first hybrid experiment did not yield any results should not be understood as the fact that the refinement of architecture alone is not the key driver of the improvement of performance in this context.

A second architecture fairness test was subsequently done by doubling the depth of the Transformer by adding one to two encoder blocks. The design of this experiment was intentional to be a minimal architecture change without changing the training setup otherwise. The resulting two block hybrid model had a validation accuracy of 0.5761 and a validation macro-F1 of 0.5681. This was a slight improvement over the previous one-block hybrid when trained under the same focal-plus-alpha training protocol. Nevertheless, the hybrid model did not even outperform the most powerful CNN-only ConvNeXt-Tiny baseline trained with focal loss and class-aware alpha.

A final best-potential hybrid single-model run was then performed using the same two-block hybrid architecture together with a stronger stability-focused optimization setting. This configuration included focal loss, class-aware alpha, EMA, reduced learning rate, warmup, and gradient clipping. The resulting model, identified as the strongest verified hybrid single model in the thesis, achieved a validation accuracy of 0.5799 and a validation macro-F1 of 0.5708. This brought the hybrid very close to the best CNN-only model, but it still remained slightly behind on both primary validation metrics. Therefore, the final single-model hybrid conclusion was that the hybrid could become highly competitive, but did not officially surpass the strongest CNN-only ConvNeXt-Tiny configuration.

In addition to the training-time hybrid results, an inference-side hybrid rescue evaluation was performed to estimate the upper practical potential of the hybrid approach without retraining. Two strong hybrid checkpoints were combined using checkpoint ensembling, and horizontal-flip test-time augmentation was also evaluated. Among the evaluated inference settings, the best result was obtained by the two-checkpoint hybrid ensemble with horizontal-flip TTA, which achieved a validation accuracy of 0.5821 and a validation macro-F1 of 0.5741. This configuration nearly closed the gap to the best CNN-only model. However, because it relies on multiple checkpoints and augmented inference, it is treated in the thesis as an offline inference enhancement rather than as the main real-time deployment model.

The hybrid architecture experiments therefore support several important observations. First, architecture refinement alone did not produce the strongest gains. Second, a slightly deeper hybrid became more competitive than the initial lightweight variant. Third, stability-focused training improved the hybrid further and produced the strongest hybrid single-model result. Fourth, despite this development, the highest CNN-only ConvNeXt-Tiny model was still better on the primary

validation metrics. Lastly, ensembling at the inference side and TTA demonstrated that the hybrid methodology has a higher practical ceiling than the outcome alone of the single-checkpoint methodology would indicate, at the cost of increased inference complexity.

Regarding efficiency, significant yet unambiguous computational cost was also brought by the hybrid models. The highest hybrid single model had 29.073 million parameters, as compared to 27.826 million parameters in the best CNN-only ConvNeXt-Tiny baseline. The optimal hybrid single model also had increased latency and reduced FPS as compared to the best CNN-only model. This implies that the hybrid was not only slightly lagging behind in recognition quality but also required a slightly higher inference cost. It has a significant role in the final interpretation since the thesis is expressly focused on the trade-off between the recognition performance and the practical efficiency.

In short, the hybrid architecture experiments reveal that lightweight CNNTransformer refinement is a promising and competitive direction to in-the-wild facial expression recognition, but in the forms tested, it was not the dominant source of improvement. The best hybrid model came close to the best CNN-only model and improved substantially over the initial hybrid variant, but the final evidence indicates that the main gains in this thesis came more strongly from training strategy than from the lightweight Transformer refinement itself.

Experiment	**Hybrid configuration**	**Accuracy**	**Macro-F1**	**Params (M)**	**Latency (ms)**	**FPS**
convnext_tiny_transformer_ce_run1	ConvNeXt-Tiny + Transformer (1 block), CE	0.5239	0.5009	28.546	4.5061	221.92
convnext_tiny_transformer_focal_alpha_run1	ConvNeXt-Tiny + Transformer (1 block), Focal + alpha	0.5706	0.5634	28.546	4.5224	221.12
convnext_tiny_transformer_focal_alpha_run2	ConvNeXt-Tiny + Transformer (2 blocks), Focal + alpha	0.5761	0.5681	29.073	4.8656	205.53

convnext_tiny_transformer_focal_alpha_ema_run3	ConvNeXt-Tiny + Transformer (2 blocks), Focal + alpha + EMA	0.5799	0.5708	29.073	4.665	214.36

Table 4. Hybrid model progression results

3.4. Imbalance-aware training experiments

One of the central goals of this thesis was to determine whether the main performance gains in AffectNet-8 come from model architecture or from training strategy. This question was particularly important because AffectNet-8 has a strongly imbalanced class distribution, and therefore a model may behave very differently depending on how the optimization process handles minority classes. For this reason, the thesis included a set of controlled experiments focused on imbalance-aware and stability-focused training strategies, especially focal loss, class-aware alpha weighting, and later optimization stabilizers such as EMA, warmup, and gradient clipping.

The first reference point for this analysis was the comparison between cross-entropy and focal-loss-based training. Under standard cross-entropy training, the ConvNeXt-Tiny baseline achieved a validation macro-F1 of 0.5048. When the same baseline architecture was trained with focal loss and class-aware alpha, its validation macro-F1 increased to 0.5772, while validation accuracy increased from 0.5259 to 0.5829. This was one of the strongest improvements observed in the entire study. Since the architecture itself remained the same, the gain can be attributed primarily to the change in training strategy rather than to model design.

A similar pattern was observed for the hybrid architecture. The first lightweight hybrid trained with cross-entropy achieved a validation macro-F1 of 0.5009. When focal loss and class-aware alpha were applied to the same general hybrid design, validation macro-F1 increased to 0.5634 and validation accuracy increased to 0.5706. This confirmed that focal loss and class-aware weighting were beneficial not only for the CNN-only baseline, but also for the hybrid models. Thus, the performance improvement linked to the imbalance-conscious training was not related to a particular architecture family; rather, it was a consistent effect across the major model types that were taken into consideration in the thesis.

The most successful controlled ablation of the best CNN-only and hybrid focal-plus-alpha configurations provided one of the clearest study findings. The resulting model, despite being even worse than the corresponding hybrid variants on the main validation metrics, still outperformed the corresponding hybrid variants on the main validation metrics. This implies that whilst the hybrid was able to enjoy the benefits of imbalance-conscious training, the biggest gains that have been verified cannot be solely attributed to the Transformer refinement. Rather, the evidence has shown that the training configuration on the basis of focal losses was the most dominant source of improvement in the experiments that were conducted.

In the perspective of classes, the imbalance-conscious training experiments were especially significant in the case of challenging and underrepresented classes. The total macro-F1 increases were added with the better performance of recognition of such categories as fear, disgust, and

contempt. This is exactly the type of effect expected from a loss function that reduces the influence of easy majority-class samples and increases the contribution of hard or minority-class examples. The detailed per-class analyses later in the chapter confirm that the strongest performance gains were not only numerical improvements in overall macro-F1, but also meaningful improvements in difficult-class recognition.

After establishing the value of focal loss and class-aware alpha, the study also examined whether the strongest hybrid model could be further improved through stability-focused training. This led to the final hybrid configuration with EMA, lower learning rate, warmup, and gradient clipping. Compared with the earlier two-block hybrid trained with focal loss and alpha alone, this stability-focused hybrid improved from 0.5681 to 0.5708 in macro-F1 and from 0.5761 to 0.5799 in validation accuracy. These gains were real but smaller than the earlier gains obtained from the transition from cross-entropy to focal-plus-alpha training. This again suggests that the primary breakthrough in the thesis came from imbalance-aware loss design, while stabilization refinements provided a secondary improvement.

The strongest evidence for the role of training strategy comes from the overall progression of the experiments. The best completed CNN-only model and the best completed hybrid single model were both obtained only after the introduction of focal loss and class-aware alpha. Only in case of these changes in training, the hybrid architecture became highly competitive. Similarly, the optimal hybrid single model did not only entail focal-plus-alpha training but also extra stabilization. Thus, the conclusions that the most significant gains were training-based, not architecture-based are supported by the sequence of the experiment.

The scientific significance of this finding is due to two reasons. To begin with, it is more specific in explaining what enhanced performance in this project. The gains would have been easily over-attributed to the hybrid design itself, without the controlled training ablations. Second, it results in a practical conclusion to facial expression recognition under the class imbalance: before it advances architectural complexity, it is necessary to make sure that the training process is suitable to the data distribution. In the case of AffectNet-8, this implies that loss design and training stability could be at least equally important as, and in this case more so important than, lightweight Transformer refinement.

Conclusively, the imbalance-sensitive training experiments can be considered one of the strongest aspects of the thesis. They demonstrate that class-aware alpha with focal loss gained the most verified gains across the experiments carried out, significantly enhancing both CNN-only and hybrid models. They further demonstrate that refinements aimed at stability like EMA, warmup, and clipping could make the hybrid more competitive, but such subsequent refinements remain minor compared to the impact of the underlying imbalance-sensitive loss design. This results in one of the key conclusions of the thesis: in the experimented AffectNet-8 configuration, the most important of the strong performance gains was the training strategy and not the lightweight hybrid refinement on its own.

Architecture	Training setting	Accuracy	Macro-F1	Accuracy gain	Macro-F1 gain
ConvNeXt-Tiny baseline	Cross-Entropy	0.5259	0.5048	—	—
ConvNeXt-Tiny baseline	Focal + alpha	0.5829	0.5772	0.057	0.0724

ConvNeXt-Tiny + Transformer (1 block)	Cross-Entropy	0.5239	0.5009	—	—
ConvNeXt-Tiny + Transformer (1 block)	Focal + alpha	0.5706	0.5634	0.0467	0.0625

Table 5. Effect of imbalance-aware training on CNN-only and hybrid models

3.5. Final hybrid inference-side evaluation

Following the main single-model training experimentations, another inference-side evaluation was done to approximate the practical upper bound of the hybrid approach without undergoing another training cycle. The motivation for this step was that the strongest hybrid single model had already become highly competitive, but still remained slightly below the best CNN-only ConvNeXt-Tiny baseline on the main validation metrics. Therefore, before finalizing the conclusions of the thesis, a final inference-side rescue attempt was conducted to determine whether the hybrid could be improved further through evaluation-only techniques.

This stage did not involve retraining or architecture redesign. Instead, it focused on three practical inference-level strategies: test-time augmentation (TTA) applied to the strongest hybrid checkpoint, checkpoint ensembling of the two strongest hybrid models, and checkpoint ensembling combined with horizontal-flip TTA. The evaluated hybrid checkpoints were the best saved models from the two-block focal-plus-alpha hybrid and the stability-enhanced hybrid with EMA. These checkpoints represented the strongest hybrid configurations obtained during the main training phase, and therefore provided a suitable basis for an inference-side upper-bound evaluation.

The first evaluated configuration used the strongest hybrid single checkpoint without any additional inference enhancement. This environment simulated the performance of the optimum hybrid single model on an independent evaluation program which was used as the benchmark to compare with the others later. The second device used horizontal-flip TTA of a single checkpoint. Here, every validation image was tested in its original and horizontally flipped versions, and the resulting logits were averaged and the final prediction was made. The aim was to test the hypothesis that simple, inference-time facial-symmetry-based augmentation can be used to enhance robustness.

The third setup involved a two checkpoint ensemble in the absence of TTA. The validation image was in this case processed separately by the two most powerful hybrid models and the final prediction was obtained by averaging the resultant logits. This strategy was intended to reduce checkpoint-specific variance and to combine complementary decision behavior learned during different training runs. The fourth and final configuration combined the two-checkpoint ensemble with horizontal-flip TTA. This represented the strongest inference-side hybrid setting evaluated in the thesis.

The results showed that TTA alone did not improve the strongest hybrid single checkpoint. In fact, the single-model TTA configuration produced slightly lower overall performance than the single-model no-TTA reference. This indicates that horizontal-flip augmentation by itself was not sufficient to improve the hybrid in a meaningful way under the tested setup. By contrast, checkpoint ensembling produced a small but consistent improvement over the best single hybrid model. When

the two strong hybrid checkpoints were combined without TTA, the validation results improved slightly, showing that the two checkpoints carried partially complementary information.

The strongest inference-side hybrid result was obtained when checkpoint ensembling and horizontal-flip TTA were used together. This configuration achieved a validation accuracy of 0.5821 and a validation macro-F1 of 0.5741. Compared with the strongest hybrid single model, this represented a measurable improvement and brought the hybrid extremely close to the best CNN-only ConvNeXt-Tiny baseline. The remaining gap to the best CNN-only model was very small: approximately 0.0008 in validation accuracy and 0.0031 in macro-F1. Therefore, the final inference-side rescue attempt showed that the hybrid approach has stronger practical potential than is visible from the best single checkpoint alone.

Despite this encouraging result, the thesis does not treat the ensemble-plus-TTA configuration as the main real-time model. The reason is that checkpoint ensembling and test-time augmentation both increase inference complexity and reduce deployment simplicity. Since the real-time framing of the thesis is based on efficient single-model inference, the ensemble-plus-TTA configuration is more appropriately interpreted as an offline evaluation enhancement or upper-bound hybrid inference setting. It shows the practical best the tested hybrid family could achieve without further training, but does not supplant the single-model findings of the research.

This difference matters towards the ultimate interpretation. An all-purpose single model that was the best in the thesis was the ConvNeXt-Tiny baseline with focal loss and a class-aware alpha. The strongest hybrid single model remained the two-block hybrid trained with focal loss, alpha, EMA, warmup, reduced learning rate, and gradient clipping. The inference-side rescue result demonstrated that the hybrid architecture could be pushed even closer to the strongest CNN-only model, but the thesis maintains a clear separation between real-time single-model deployment conclusions and offline inference-side enhancement.

In summary, the final hybrid inference-side evaluation showed that checkpoint ensembling was more beneficial than TTA alone and that the combination of ensembling and horizontal-flip TTA produced the best hybrid inference result in the thesis. This result did not officially surpass the strongest CNN-only single model, but it nearly closed the gap and therefore strengthens the conclusion that the tested hybrid approach was highly competitive even if it did not become the strongest overall model under the final single-model evaluation criteria.

Configuration	Accuracy	Macro-F1
Single best hybrid without TTA	0.5796	0.5704
Single best hybrid with h-flip TTA	0.5776	0.5677
Two-checkpoint hybrid ensemble	0.5799	0.5714

Two-checkpoint hybrid ensemble + h-flip TTA	0.5821	0.5741

Table 6. Hybrid inference-side rescue results with TTA and ensemble

The best hybrid inference-side result was obtained by two-checkpoint ensembling with horizontal-flip TTA.

3.6. Efficiency benchmarking

Efficiency benchmarking was carried out after the model-training phase in order to support the real-time framing of the thesis. The purpose of this stage was to compare the practical inference cost of the completed single-model experiments under one fixed benchmarking protocol. In addition to recognition metrics, the study therefore included total parameter count, latency per image, and frames per second (FPS). These measures allowed the thesis to evaluate not only which model performed best in terms of validation accuracy and macro-F1, but also which model offered the strongest practical accuracy–efficiency trade-off.

The benchmark included seven completed single-model experiments: the ResNet50 baseline, the ConvNeXt-Tiny baseline, the first lightweight hybrid trained with cross-entropy, the first focal-plus-alpha hybrid, the strongest CNN-only ConvNeXt-Tiny model, the stronger two-block hybrid, and the strongest hybrid single model with EMA. Each model was reconstructed from its saved best checkpoint, transferred to the GPU, and evaluated under the same single-image inference conditions. The benchmark protocol used 224 × 224 input resolution, batch size one, mixed precision inference, 50 warmup iterations, and 200 timed forward passes.

The fastest model among all benchmarked configurations was the ResNet50 baseline. It contained 23.524 million trainable parameters, achieved an average latency of 3.4479 milliseconds per image, and reached 290.03 FPS. These results confirm that ResNet50 provides the strongest speed-oriented baseline in the thesis. However, as discussed earlier, this speed advantage was accompanied by weaker macro-F1 and weaker minority-class behavior compared with the strongest ConvNeXt-based models.

The ConvNeXt-Tiny baseline trained with cross-entropy contained 27.826 million parameters, achieved 4.0567 milliseconds latency, and reached 246.51 FPS. Although this model was slower than the ResNet50 baseline, it provided slightly stronger macro-F1 and therefore represented a more balanced convolutional reference. More importantly, the strongest CNN-only configuration in the thesis, namely ConvNeXt-Tiny with focal loss and class-aware alpha, retained the same parameter count of 27.826 million while improving performance substantially. This best single-model configuration achieved 3.9226 milliseconds latency and 254.94 FPS, alongside validation accuracy of 0.5829 and macro-F1 of 0.5772. These results make it the strongest overall single-model accuracy–efficiency trade-off in the completed experiments.

The hybrid models consistently introduced a modest increase in parameter count relative to the CNN-only ConvNeXt-Tiny baseline. The one-block hybrid variants contained 28.546 million parameters, while the two-block hybrid variants contained 29.073 million parameters. Although this parameter increase was not extremely large, the corresponding latency penalty was more visible than the parameter growth alone would suggest. The first hybrid with cross-entropy achieved

4.5061 milliseconds latency and 221.92 FPS, while the strongest hybrid single model achieved 4.6650 milliseconds latency and 214.36 FPS. The two-block focal-plus-alpha hybrid without EMA was even slightly slower at 4.8656 milliseconds and 205.53 FPS.

These results demonstrate an important point: in practical inference, the computational cost of hybrid refinement is influenced not only by parameter count but also by the nature of Transformer-related operations. Consequently, the added cost of the hybrid is more clearly reflected in latency and FPS than in total parameters alone. From the best CNN-only model to the best hybrid single model, the parameter count increased by approximately 1.25 million, while latency increased from 3.9226 to 4.6650 milliseconds and throughput decreased from 254.94 to 214.36 FPS. This means that the hybrid remained computationally efficient in absolute terms, but introduced a noticeable speed penalty relative to the strongest CNN-only baseline.

The benchmark results are therefore important for the final trade-off discussion of the thesis. If one focuses strictly on speed, ResNet50 is the strongest model. If one focuses on the best balance between validation performance and efficiency, the strongest result is given by the ConvNeXt-Tiny baseline with focal loss and class-aware alpha. If one focuses specifically on hybrid modeling, the strongest single hybrid is the EMA-stabilized two-block ConvNeXt-Tiny + Transformer model, which is highly competitive but less efficient than the best CNN-only model and still slightly weaker on the main validation metrics.

The benchmark also helps explain why the offline hybrid rescue configuration is not treated as the main real-time solution. Even though the ensemble-plus-TTA hybrid nearly matched the best CNN-only model in recognition quality, it relies on multiple checkpoints and augmented inference, which would further increase effective runtime cost. Therefore, the thesis reserves the real-time comparison for the single-model benchmarked configurations and presents the ensemble-plus-TTA result only as an upper-bound hybrid inference enhancement.

In summary, the efficiency benchmarking phase confirmed three main findings. First, ResNet50 is the fastest completed single-model configuration. Second, the ConvNeXt-Tiny baseline with focal loss and class-aware alpha provides the best overall accuracy–efficiency trade-off in the thesis. Third, the hybrid models are both computationally feasible and experience a definite inference penalty compared to the strongest CNN-only model without strong validation performance or macro-F1. These results form the core of the final real time interpretation of the study, and support the conclusion that the most suitable model choice is one that does not just depend on the quality of recognition, but also on the acceptable level of computational overhead.

Experiment	Accuracy	Macro-F1	Params (M)	Latency (ms)	FPS
baseline_resnet50_ce_run1	0.5326	0.5026	23.524	3.4479	290.03
convnext_tiny_baseline_ce_run1	0.5259	0.5048	27.826	4.0567	246.51
convnext_tiny_transformer_ce_run1	0.5239	0.5009	28.546	4.5061	221.92

convnext_tiny_transformer_focal_alpha_run1	0.5706	0.5634	28.546	4.5224	221.12
convnext_tiny_baseline_focal_alpha_run1	0.5829	0.5772	27.826	3.9226	254.94
convnext_tiny_transformer_focal_alpha_run2	0.5761	0.5681	29.073	4.8656	205.53
convnext_tiny_transformer_focal_alpha_ema_run3	0.5799	0.5708	29.073	4.665	214.36

Table 7. Final parameter, latency, and FPS benchmarking results

3.7. Confusion matrices and per-class analysis

Although validation accuracy and macro-F1 give a comprehensive overview of recognition performance, they do not give a complete picture of how the models behave on each expression class. This is why normalized confusion matrices and per-class F1 results of the best CNN-only model and the best hybrid single model are also included in the final analysis of the thesis. This extra class-wise perspective is particularly crucial in AffectNet-8, where the imbalance of classes and the ambiguity of expressions can have a strong influence on individual categories.

The CNN-only model with the best results in the thesis was the ConvNeXt-Tiny baseline trained using focal loss and class-aware alpha. The validation macro-F1 and validation accuracy of this model are 0.5772 and 0.5829 respectively. Its best-epoch per-class F1 results were as follows: neutral – 0.5131, happy – 0.6795, sad – 0.6176, surprise – 0.5508, fear – 0.6173, disgust – 0.6093, anger – 0.5843, and contempt – 0.4460. These values mean that not only the best CNN-only model, but also the relatively best balanced over the entire set of classes. Specifically, it scored well on disgust and a fairly high contempt score than the previous baseline models, a significant result given that the data used was of minority-class difficulty.

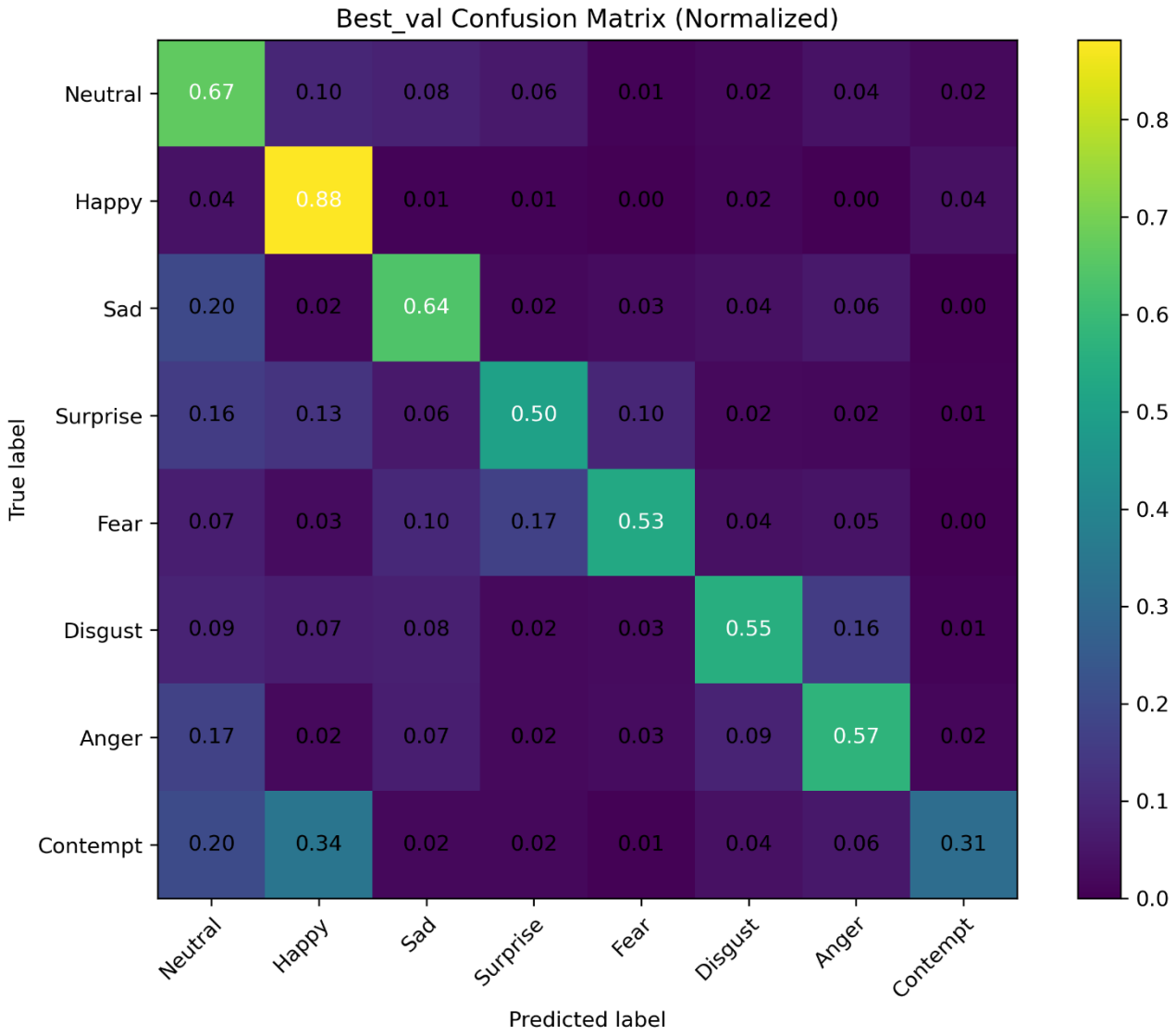

Fig. 4. Normalized confusion matrix of the best CNN-only model

The best hybrid single model in the thesis was the ConvNeXt-Tiny + lightweight Transformer model with two blocks of Transferring, focal loss, class-aware alpha, EMA, reduced learning rate, warmup and gradient clipping. The validation macro-F1 and the validation accuracy of this model were 0.5708 and 0.5799, respectively. Its best-epoch per-class F1 results were: neutral – 0.5149, happy – 0.6894, sad – 0.6325, surprise – 0.5668, fear – 0.6438, disgust – 0.5608, anger – 0.5726, and contempt – 0.3852. These values indicate that the most competitive hybrid model was a bit stronger, in several classes, than the most competitive CNN-only model.

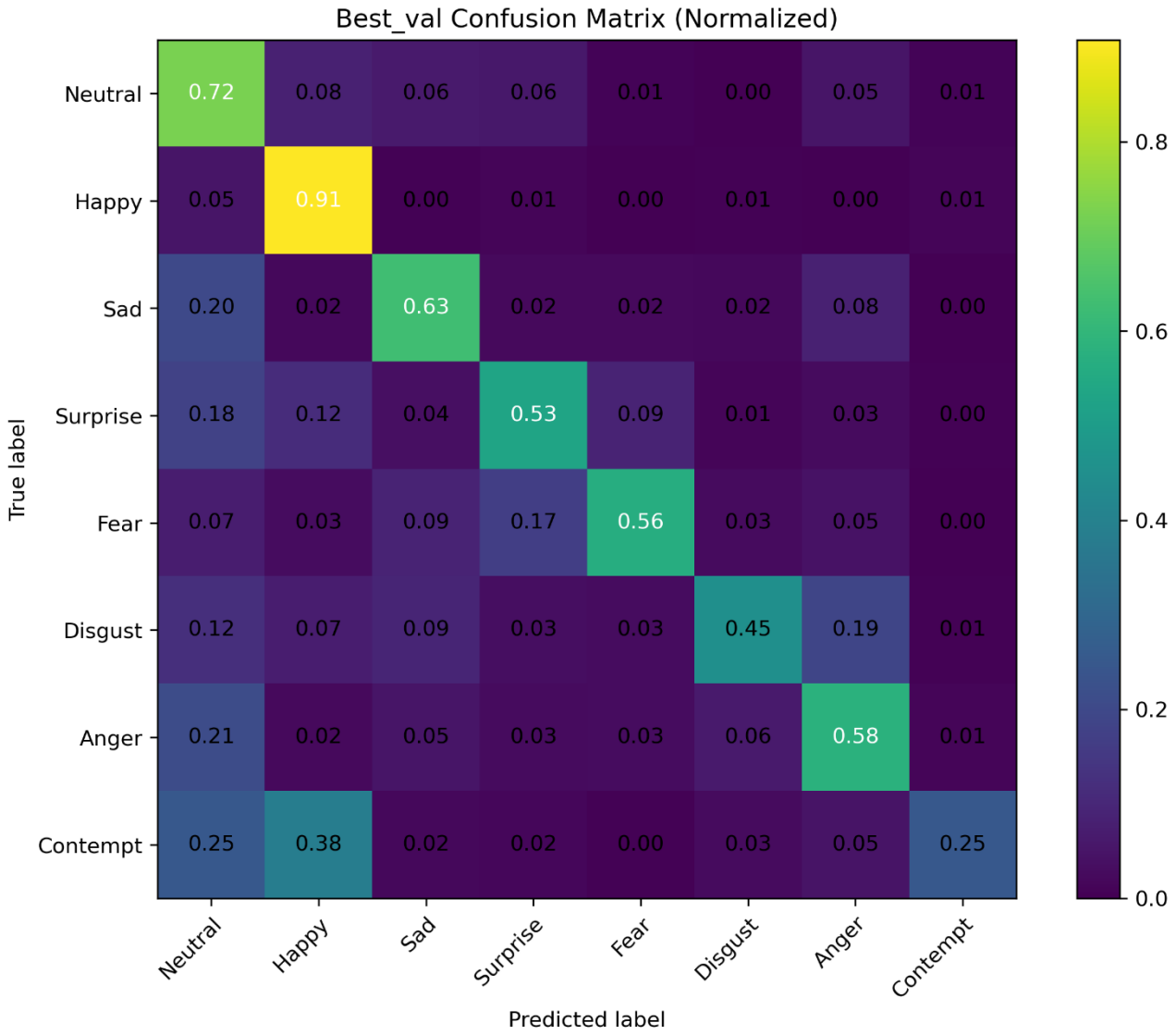

Fig. 5. Normalized confusion matrix of the best hybrid single model

Overall accuracy and macro-F1 measures do not provide a comprehensive picture of the direct class-wise comparison between the best CNN-only model and the best hybrid single model. The hybrid model achieved slightly higher F1-scores on neutral (0.5149 vs. 0.5131), happy (0.6894 vs. 0.6795), sad (0.6325 vs. 0.6176), surprise (0.5668 vs. 0.5508), and fear (0.6438 vs. 0.6173). However, the CNN-only model remained stronger on disgust (0.6093 vs. 0.5608), anger (0.5843 vs. 0.5726), and especially contempt (0.4460 vs. 0.3852). This class-level difference is the reason why the overall advantage for the hybrid model was very close to that of the CNN-only model.

Nevertheless, CNN-only model was still more robust in disgust, anger, and contempt. This is a supportive finding since these categories comprise some of the most challenging and imbalance-sensitive groups in AffectNet-8. The widest gap was recorded in contempt with CNN-only model recording 0.4460 as opposed to the best hybrid single model 0.3852. As contempt is both infrequent and difficult, this difference added to the ultimate macro-F1 benefit of the CNN-only model. In this way, although the hybrid was made very competitive overall, its poorer performance on several important classes did not allow it to officially be considered more competitive than the best CNN-only configuration.

This interpretation is supported by the normalized confusion matrices shown in **Fig. 4** and **Fig. 5**. In both models, the matrices indicate that the occurrence of errors in expression recognition is not evenly distributed, but tends to occur between semantically or visually similar classes. The CNN-only model with the highest score in terms of the proportion of correct predictions in the class set, and in particular in the minority-sensitive categories. The strongest hybrid model, in its turn, demonstrates a slightly improved concentration of correct predictions of some of the classes of the middle level of difficulty, but also more problematic confusion at the classes that are structurally challenging, in particular, the ones that have lower sample frequency. This proves that the variation between the two best models is not merely one of world quality, but of that of classes.

The analysis by class also contributes to the explanation of why macro-F1 had been such a critical measure in this thesis. Assuming that the best CNN-only and the best hybrid models had been compared based only on their accuracy, the difference between the two might have looked very small and can be readily dismissed. But on a per-class analysis we find that this tiny difference in the world is also due to substantial differences in the minority-class behavior. Put another way, the CNN-only version, which has the macro-F1 advantage, is not coincidental. It shows a better performance balance in all eight categories of expression under the ultimate validation procedure.

The confusion matrices also explain the choice of treating the ensemble-plus-TTA hybrid result as a separate result rather than a part of the main single-model comparison. Even though the offline hybrid rescue setup almost bridged the gap in overall validation metrics, the overall thesis concluding about real-time and deployment suitability must be based on the best single models. Thus, the discussion of the main confusion-matrix of the thesis is concerned with the best CNN-only model and the best hybrid single model instead of the ensemble configuration.

In general, the confusion matrices and the results per-class reinforce the ultimate interpretation of the thesis. They demonstrate that the best CNN-only and hybrid models are similar in overall quality, but use slightly different class-wise behavior to achieve this competitiveness. The hybrid model has shown to be of benefit in various classes indicating that lightweight Transformer refinement could be beneficial in contextual facial representation. Simultaneously, the CNN-only model is more reliable in a range of challenging, minority-sensitive categories and thus achieves the best final macro-F1 score among the experiments that have been completed successfully by the single model.

The detailed per-class metric tables for the best CNN-only model and the best hybrid single model are given in Appendix A, Tables 9 and 10. Additional normalized confusion matrices for the ResNet50 baseline and the two-block hybrid model without EMA are provided in Appendix B (Figs. 6 and 7) to support secondary comparison of error structure beyond the main best-model analysis.

Class	Best CNN-only F1	Best hybrid single-model F1
Neutral	0.5131	0.5149
Happy	0.6795	0.6894
Sad	0.6176	0.6325
Surprise	0.5508	0.5668
Fear	0.6173	0.6438
Disgust	0.6093	0.5608
Anger	0.5843	0.5726

Contempt	0.446	0.3852

Table 8. Per-class F1 comparison of the best CNN-only and best hybrid single model

3.8. Discussion of results

The findings of this thesis can make a number of valuable contributions to the in-the-wild recognitions of facial expressions in the presence of class imbalance. The initial significant conclusion is that the largest confirmed gains in the final experiments were achieved through imbalance-conscious training as opposed to architecture refinement by itself. The staged development of the experiments supports this conclusion. Training the ConvNeXt-Tiny baseline and the initial hybrid model using standard cross-entropy did not produce a significant improvement over the initial baselines in macro-F1. Nonetheless, when both the focal loss and class-conscious alpha weighting were added, both the CNN-only and hybrid models significantly improved. The scale of these improvements was larger than the subsequent gains due to adding more Transformer depth or other methods of stabilization like EMA and warmup.

This finding is significant since it elucidates the actual basis of increase in the research. Initially, the hypothesis that hybrid CNN-Transformer refinement could be the main factor that drives the improved performance was reasonable. But the ablations done demonstrate that this would be an imperfect interpretation. The largest gains were realized only following the training process was scaled to the skewed data distribution of AffectNet-8. Thus, the thesis serves not only to provide a comparison of architectures, but also to provide a more specific explanation why the best models got better. Here the most significant was the training policy, particularly, focal loss with class-aware alpha.

The second key finding is that the lightweight hybrid CNN-Transformer models were competitive, but did not officially outperform the strongest CNN-only ConvNeXt-Tiny baseline. This is an academic result that is of academic use although the hybrid was not the most convenient overall model. Still informative is a negative or non- winning architectural outcome that is the result of a controlled, reproducible process. The hybrid models in this thesis were tested on the same protocol of data, preprocessing, primary metrics and benchmarking conditions. Therefore, the conclusion that the tested lightweight hybrid refinement did not beat the best CNN-only model is a meaningful finding rather than an absence of contribution.

Several evidence-based explanations can be proposed for this outcome. First, the hybrid models showed signs of optimization difficulty. The early hybrid variants often peaked quickly and then degraded, which indicates that the added contextual modeling component made training less stable. This interpretation is supported by the later observation that EMA, lower learning rate, warmup, and gradient clipping improved the hybrid more than they improved already stable CNN-only baselines. Second, the additional global context provided by a lightweight Transformer may have been beneficial for some classes, but not sufficient to dominate the stronger convolutional representation learned by ConvNeXt-Tiny. In other words, global contextual refinement appears useful, but its effect size in the tested lightweight form was smaller than the effect size of the imbalance-aware loss design.

A third important explanation concerns the computational trade-off. The hybrid models only slightly increased the number of parameters, however, they also led to a significant increase in latency and a decrease in FPS. This meant that the hybrid variants were required to not only justify their increased architectural complexity, but their increased inference cost as well. As the best hybrid single model was a bit lower than the best CNN-only model with focal loss and class-aware

alpha, the final accuracy-efficiency trade-off favored the CNN-only ConvNeXt-Tiny baseline with focal loss and class-aware alpha. This finding is especially crucial with the real-time framing of the thesis.

This interpretation is supported by the benchmark results. The ResNet50 baseline was the fastest completed model and thus it is the speed-oriented reference. The best overall balance of recognition quality and efficiency was achieved by the strongest CNN-only ConvNeXt-Tiny model, and the strongest hybrid single model was highly competitive but was slower and slightly weaker on the main validation metrics. This implies that the most appropriate model decision will be determined by the desired priority. When the objective is to achieve a greater speed the reference is ResNet50. The focal loss and class-aware alpha ConvNeXt-Tiny baseline is the most appropriate between the completed single-model configurations.

The last rescue assessment, which is placed on the inference side, provided a valuable twist to the discussion. With checkpoint ensembling and horizontal-flip TTA, the hybrid family achieved a validation accuracy of 0.5821 and a macro-F1 of 0.5741, almost bridging the gap between the best CNN-only model and the hybrid family. This demonstrates that the hybrid method possesses greater practical potential not just when individual best checkpoints are considered. Nonetheless, since such an outcome is based on a series of checkpoints and augmented inference, this should be treated not as the fundamental real-time deployment outcome, but as the offline upper-bound enhancement of the same. This is a key distinction to a strict interpretation of the study.

In a more global view, a significant point of the thesis as well is that an adequate choice of metrics should be taken. In case only accuracy has been reported, the difference between the strongest CNN-only and the strongest hybrid single models would seem extremely small and can be interpreted as negligible. But with the addition of macro-F1, per-class result and confusion matrices, the study demonstrates that these little global differences capture significant class-level trade-offs. This is particularly applicable to AffectNet-8 where minority and challenging classes may highly affect the practical usefulness of a model.

Put collectively, the findings are in favor of a moderate overall interpretation. The test lightweight hybrid CNNTransformer was not a dead end. Quite the contrary, it turned extremely competitive as the focal loss, class-aware alpha and stability-focused optimization were introduced. It also demonstrated class-specific strength and was very close to the strongest CNN-only model in case of inference-side improvement. However, the ultimate controlled evidence shows that the optimum overall single-model result in the present study was obtained by CNN-only ConvNeXt-Tiny baseline with focal loss and class-aware alpha. Thus, the greatest improvements in this AffectNet-8 setup were stronger due to proper imbalance-aware training than those of the tested lightweight Transformer refinement.

This scientific conclusion is practical and has a scientific use. Scientifically it isolates effects of architecture and effects of training in a controlled fashion. In practice, it implies that when the in-the-wild FER is used under severe imbalance, one must first ensure that the robust loss design and training stability are first met before assuming that hybrid contextual refinement will automatically lead to an improvement in the final result. In this respect, the thesis will not only provide final numbers, but also a better idea of what is the most important when using efficient and reliable FER under the following conditions.

Conclusions

1. Related work analysis indicated that the problem of facial expression recognition in the wild is a challenging one because of the variation of poses and illumination changes, occlusion, blur, background complexity, and extreme class imbalance. The literature further revealed that contemporary FER assessment must be based on the balanced measurements and efficiency in practice.
2. The implemented baseline and hybrid model study showed that ConvNeXt-Tiny is a stronger backbone than the tested ResNet50 baseline in terms of balanced recognition performance on AffectNet-8. Although ResNet50 remained the fastest completed model, ConvNeXt-Tiny provided a better basis for achieving higher macro-F1 and stronger overall recognition quality.
3. The controlled training experiments demonstrated that the largest verified gains in this thesis were produced by imbalance-aware and stability-focused training rather than by lightweight hybrid refinement alone. In particular, focal loss with class-aware alpha substantially improved both validation accuracy and macro-F1, showing that loss design is a critical factor for facial expression recognition under severe class imbalance.
4. The completed evaluation results showed that the strongest single-model result was achieved by the ConvNeXt-Tiny baseline with focal loss and class-aware alpha, which reached a validation accuracy of 0.5829 and a validation macro-F1 of 0.5772. The strongest hybrid single model, based on ConvNeXt-Tiny with two Transformer blocks and stability-focused training, achieved a validation accuracy of 0.5799 and a validation macro-F1 of 0.5708. This confirms that the tested hybrid approach became highly competitive, but did not surpass the best CNN-only model under the main validation metrics.
5. The class-wise and confusion-matrix analysis showed that the hybrid model performed slightly better on several classes, including neutral, happy, sad, surprise, and fear, while the strongest CNN-only model remained more reliable on difficult minority-sensitive classes, especially disgust, anger, and contempt. This explains why the final difference between the best models cannot be understood from accuracy alone and confirms the importance of macro-F1 and per-class evaluation in AffectNet-8.
6. The efficiency benchmark showed that the best overall accuracy–efficiency trade-off among the completed single-model experiments was achieved by the ConvNeXt-Tiny baseline with focal loss and class-aware alpha. The hybrid models introduced only a moderate increase in parameter count, but also a clear latency penalty and lower FPS. Therefore, under the completed experimental conditions of this thesis, the strongest practical recommendation for real-time facial expression recognition is the CNN-only ConvNeXt-Tiny baseline with focal loss and class-aware alpha.

List of References

[1] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” Oct. 2017, doi: 10.1109/TAFFC.2017.2740923.

[2] Y. Wang *et al.*, “A Survey on Facial Expression Recognition of Static and Dynamic Emotions,” Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.15777

[3] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” Oct. 2018, doi: 10.1109/TAFFC.2020.2981446.

[4] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929

[5] G. Wang, J. Li, Z. Wu, J. Xu, J. Shen, and W. Yang, “EfficientFace: An Efficient Deep Network with Feature Enhancement for Accurate Face Detection,” Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.11816

[6] A. P. Fard, M. M. Hosseini, T. D. Sweeny, and M. H. Mahoor, “AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels,” Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.22506

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[8] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, “Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition,” Sep. 2019, [Online]. Available: http://arxiv.org/abs/1905.04075

[9] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, “Recurrent neural networks for emotion recognition in video,” in *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Nov. 2015, pp. 467–474. doi: 10.1145/2818346.2830596.

[10] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, “Suppressing Uncertainties for Large-Scale Facial Expression Recognition.” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_Suppressing_Uncertainties_for_Large-Scale_Facial_Expression_Recognition_CVPR_2020_paper.pdf

[11] A. H. Farzaneh and X. Qi, “Facial Expression Recognition in the Wild via Deep Attentive Center Loss,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021. Available: https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_Suppressing_Uncertainties_for_Large-Scale_Facial_Expression_Recognition_CVPR_2020_paper.pdf

[12] Z. Wen, W. Lin, T. Wang, and G. Xu, “Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition,” *Biomimetics*, vol. 8, no. 2, Jun. 2023, doi: 10.3390/biomimetics8020199.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2201.03545

[14] L. Yuan *et al.*, “Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet.” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 558–567. DOI: 10.1109/ICCV48922.2021.00060.

[15] R. Jayaswal, M. A. Ansari, M. Dixit, D. K. Singh, and S. Ahmad, “Advances in facial expression recognition technologies for emotion analysis,” Dec. 01, 2025, *Springer Science and Business Media B.V.* doi: 10.1007/s10791-025-09699-8.

[16] B. Feng and H. Zhang, “Expression Recognition Based on Visual Transformers with Novel Attentional Fusion,” in *Journal of Physics: Conference Series*, Institute of Physics, 2024. doi: 10.1088/1742-6596/2868/1/012036.

[17] D. Song and C. Liu, “A facial expression recognition network using hybrid feature extraction,” *PLoS One*, vol. 20, no. 1, Jan. 2025, doi: 10.1371/journal.pone.0312359.

[18] E. H. Khujamatov, M. Abdullaev, and S. Umirzakova, “Analytical Modeling of Hybrid CNN-Transformer Dynamics for Emotion Classification,” *Mathematics*, vol. 14, no. 1, Jan. 2026, doi: 10.3390/math14010085.

[19] S. N. Yousafzai *et al.*, “A multi-scale simplicial transformer with graph attention for facial emotion recognition,” *Ain Shams Engineering Journal*, vol. 16, no. 10, Oct. 2025, doi: 10.1016/j.asej.2025.103584.

[20] F. Xue, Q. Wang, and G. Guo, “TransFER: Learning Relation-aware Facial Expression Representations with Transformers,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. DOI: 10.1109/ICCV48922.2021.00358.

[21] C. Zheng, M. Mendieta, and C. Chen, “POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition.” in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2023. DOI: 10.1109/ICCVW60793.2023.00339.

[22] Z. Zhao, Q. Liu, and F. Zhou, “Robust Lightweight Facial Expression Recognition Network with Label Distribution Training,” 2021. Available: https://doi.org/10.1609/aaai.v35i4.16465

[23] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, “LibreFace: An Open-Source Toolkit for Deep Facial Expression Analysis,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024. Available: https://arxiv.org/abs/2308.10713

[24] Q. Li *et al.*, “Optimizing Class Imbalance in Facial Expression Recognition Using Dynamic Intra-Class Clustering,” *Biomimetics*, vol. 10, no. 5, May 2025, doi: 10.3390/biomimetics10050296.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” Feb. 2018. Available: http://arxiv.org/abs/1708.02002

[26] Q. T. Ngo and S. Yoon, “Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset,” *Sensors (Switzerland)*, vol. 20, no. 9, May 2020, doi: 10.3390/s20092639.

[27] Y. Zhang, C. Wang, and W. Deng, “Relative Uncertainty Learning for Facial Expression Recognition,” in Advances in Neural Information Processing Systems (NeurIPS), 2021. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/9332c513ef44b682e9347822c2e457ac-Paper.pdf

[28] N. Le, K. Nguyen, Q. Tran, E. Tjiputra, B. Le, and A. Nguyen, “Uncertainty-aware Label Distribution Learning for Facial Expression Recognition,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023. Available: https://doi.org/10.48550/arXiv.2209.10448

[29] Z. Wu and J. Cui, “LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. DOI: 10.1109/ICCV51070.2023.01892

[30] Y. Zhang, Y. Li, L. Qin, X. Liu, and W. Deng, “Leave No Stone Unturned: Mine Extra Knowledge for Imbalanced Facial Expression Recognition.” in Advances in Neural Information Processing Systems (NeurIPS), 2023. Available: https://doi.org/10.48550/arXiv.2310.19636.

[31] H. A. Shehu, W. N. Browne, and H. Eisenbarth, “Emotion categorization from facial expressions: A review of datasets, methods, and research directions,” Apr. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.neucom.2025.129367.

[32] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” Jan. 2019. Available: http://arxiv.org/abs/1711.05101

Appendices

Appendix A. Per-class metric tables

This appendix presents the detailed per-class evaluation results for the best CNN-only model and the best hybrid single model. Tables 9 and 10 report precision, recall, F1-score, and support for each expression class, and are used to support the class-wise interpretation presented in Section 3.7..

Appendix A.1. Per-class results of the best CNN-only model

Model: ConvNeXt-Tiny baseline + Focal Loss + class-aware alpha
Experiment ID: `convnext_tiny_baseline_focal_alpha_run1`
Validation accuracy: 0.5829
Validation macro-F1: 0.5772

Class	Precision	Recall	F1-score	Support
Neutral	0.4173	0.6660	0.5131	500
Happy	0.5526	0.8820	0.6795	500
Sad	0.5985	0.6380	0.6176	500
Surprise	0.6072	0.5040	0.5508	500
Fear	0.7315	0.5340	0.6173	500
Disgust	0.6798	0.5520	0.6093	500
Anger	0.5971	0.5720	0.5843	500
Contempt	0.7659	0.3146	0.4460	499

Table 9. Per-class results of the best CNN-only model

Appendix A.2. Per-class results of the best hybrid single model

Model: ConvNeXt-Tiny + lightweight Transformer + Focal Loss + class-aware alpha + EMA
Experiment ID: `convnext_tiny_transformer_focal_alpha_ema_run3`
Validation accuracy: 0.5799
Validation macro-F1: 0.5708

Class	Precision	Recall	F1-score	Support
Neutral	0.3996	0.7240	0.5149	500
Happy	0.5557	0.9080	0.6894	500
Sad	0.6351	0.6300	0.6325	500
Surprise	0.6092	0.5300	0.5668	500
Fear	0.7500	0.5640	0.6438	500
Disgust	0.7386	0.4520	0.5608	500
Anger	0.5653	0.5800	0.5726	500
Contempt	0.8333	0.2505	0.3852	499

Table 10. Per-class results of the best hybrid single model

Appendix A.3. Short comparison note

The appendix-level class-wise results confirm that the strongest CNN-only model and the strongest hybrid single model are very close overall, but they differ in class-specific behavior. The hybrid model performs slightly better on neutral, happy, sad, surprise, and fear, while the CNN-only model remains stronger on disgust, anger, and especially contempt. This class-level difference explains why the CNN-only model preserves a slightly higher final macro-F1 despite the competitive overall performance of the hybrid model.

This appendix contains the normalized confusion matrices used to support the class-wise interpretation of the final results.

Appendix B. Additional matrices

This appendix offers a few extra normalized confusion matrices for the two secondary models used in the thesis, the ResNet50 baseline model (Fig. 6) and the two-block hybrid model without EMA (Fig. 7). These numbers are provided for the convenience of the reader to complement the analysis presented in the main text, and to provide a view of the error structure of the speed-oriented CNN baseline and a more powerful hybrid variant prior to EMA-based stabilization.

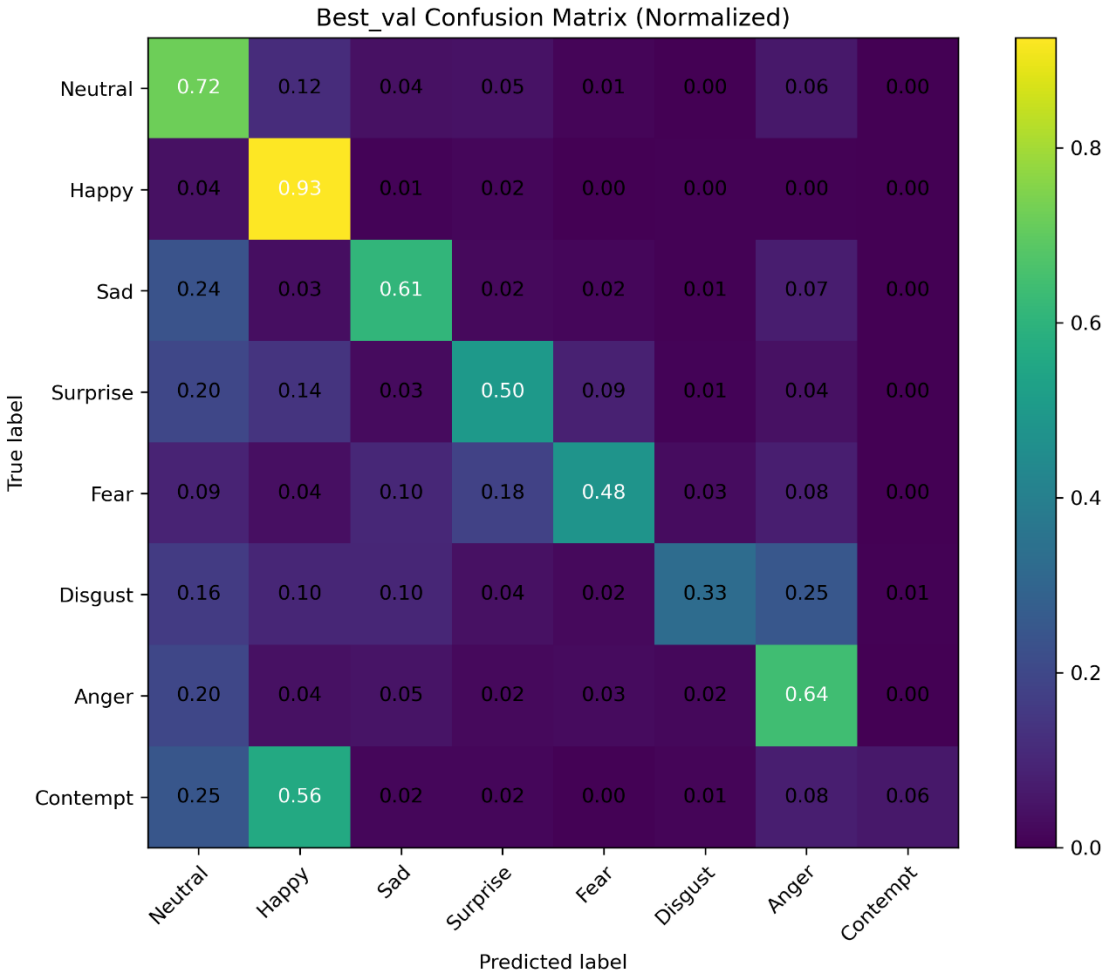

Fig. 6. ResNet50 Confusion Matrix

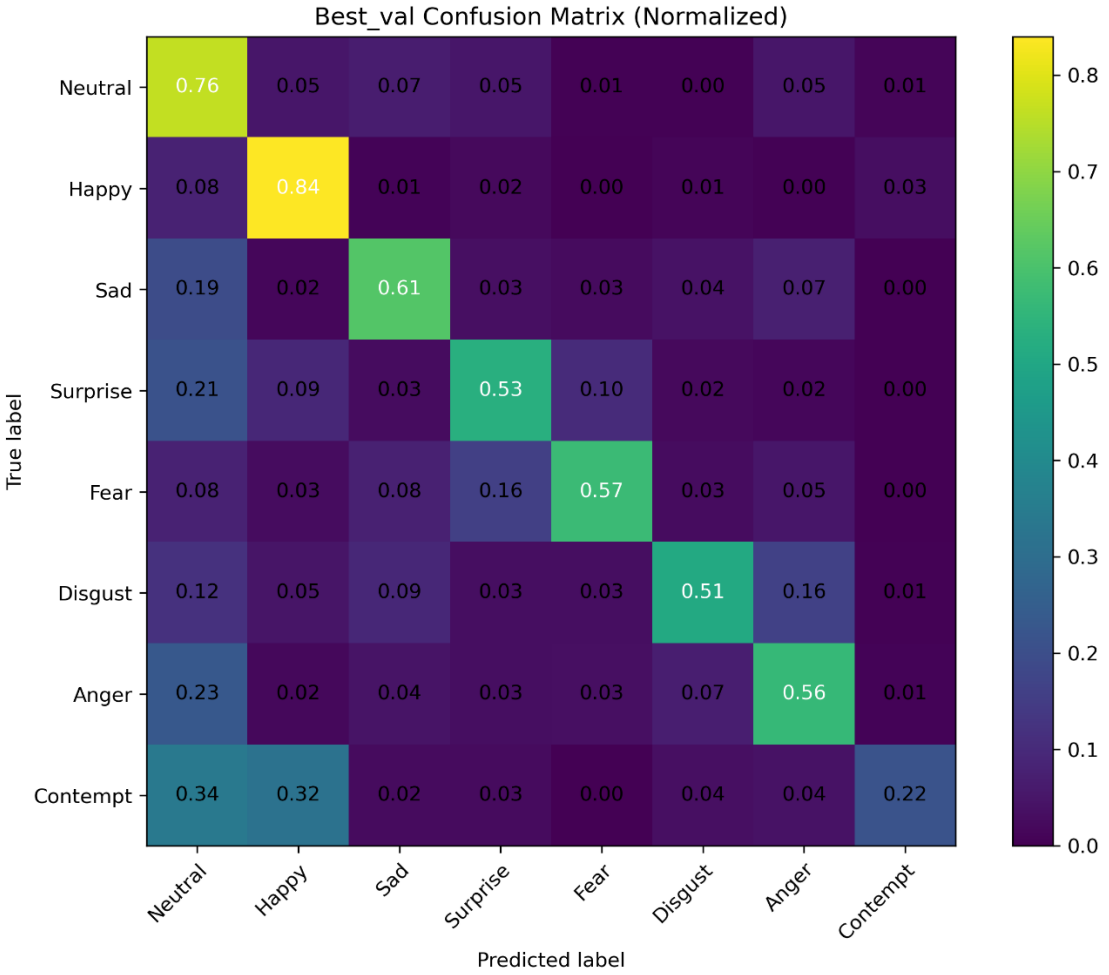

Fig. 7. Two Block Hybrid without EMA confusion Matrix

Appendix C. Reproducibility checklist

This appendix summarizes the main settings used to ensure the reproducibility of the experiments.

Appendix C.1. Data and protocol

- Dataset: AffectNet-8
- Training split size: 287,651
- Validation split size: 3,999
- Train CSV: /home/nadbal/Thesis/Dataset/train_index_full.csv
- Validation CSV: /home/nadbal/Thesis/Dataset/val_index_full.csv
- Fixed class mapping:
  - 0 – Neutral
  - 1 – Happy
  - 2 – Sad
  - 3 – Surprise
  - 4 – Fear
  - 5 – Disgust

    - 6 – Anger
    - 7 – Contempt

Appendix C.2. Input and preprocessing

- Input size: 224 × 224
- Validation preprocessing:
    - resize
    - tensor conversion
    - ImageNet normalization
- Training preprocessing:
    - resize
    - horizontal flip
    - light color jitter
    - tensor conversion
    - ImageNet normalization

Appendix C.3. General training settings

- Framework: PyTorch + torchvision
- Device: CUDA-enabled GPU environment
- Mixed precision: enabled
- Random seed: 42
- Optimizer: AdamW
- Early stopping patience: 4

Appendix C.4. Best CNN-only model settings

Experiment: convnext_tiny_baseline_focal_alpha_run1

- Architecture: ConvNeXt-Tiny baseline
- Loss: Focal Loss + class-aware alpha
- Gamma: 2.0
- Learning rate: 1e-4
- Weight decay: 1e-4
- Scheduler: CosineAnnealingLR
- Batch size: 128

Appendix C.5. Best hybrid single model settings

Experiment: convnext_tiny_transformer_focal_alpha_ema_run3

- Architecture: ConvNeXt-Tiny + lightweight Transformer
- Transformer depth: 2 blocks
- Loss: Focal Loss + class-aware alpha
- Gamma: 2.0
- EMA: enabled
- Learning rate: 5e-5
- Warmup: enabled
- Gradient clipping: 1.0
- Batch size: 128

Appendix C.6. Alpha values used in focal-loss experiments

The following class-aware alpha values were used:

- class 0: 0.403369
- class 1: 0.301054
- class 2: 0.691747
- class 3: 0.929849
- class 4: 1.382056
- class 5: 1.789801
- class 6: 0.699721
- class 7: 1.802404

Appendix C.7. Benchmarking settings

- Input size: 224 × 224
- Batch size: 1
- Warmup iterations: 50
- Timed iterations: 200
- Mixed precision inference: enabled
- Metrics:
    - parameter count
    - latency per image
    - FPS

Appendix D. AI tools usage statement

Artificial intelligence tools were used in a limited supportive role during the preparation of this thesis. They were only used to help in refining the language, structuring the draft and explaining the technical concepts as well as supporting the organization of the written presentation of the research process. The author also carried out all the implementation decisions, the execution of the experiment, debugging, verification of results, and finally determination of findings.

There was no AI tool to produce experimental results, create data, change measured outputs, or substitute the scientific content of the thesis with an AI. All of the reported metrics, tables, confusion matrices, and benchmark values were created by the implemented experimental pipeline and verified by the author prior to being included in the thesis.

The author accepts all responsibility towards the accuracy, integrity, interpretation and ultimate presentation of the work.