SEP Emotion Classifier - Preliminary Report

Isabela Labus Julius Porzel Necati Deniz Baykus Mehmet Berberoglu
{I.Labus,julius.porzel,n.baykus,M.Berberoglu}@campus.lmu.de

Abstract

We outline an emotion-classification pipeline that adapts a ResNet-18 backbone to the fixed 64×64 *input size mandated by the SEP task, couples it with a multi-source, preprocessing-first data strategy, and prioritizes balanced, explainable predictions for six canonical expressions. The preprocessing pipeline stabilizes faces, intelligently upscales FER-2013 frames, and progressively merges AffectNet and RAF-DB sources so that the limited pixel budget is devoted to expressive detail rather than nuisance pose variation. Weighted cross-entropy, regularisation, and Grad-CAM saliency maps keep the training stable while enabling interpretable feedback, and macro-averaged metrics plus a hold-out validation split ensure performance is measured reliably despite the amplified class imbalance.*

1. Introduction

This report distills the collective research into a coherent strategy for the six-emotion classification task under the 64×64 constraint. The problem statement drives every design decision: the limited resolution reduces the signal available to represent subtle facial expressions, yet the downstream demo requires robust, interpretable outputs for happiness, surprise, sadness, anger, disgust, and fear. Each subsequent section unpacks one of the research threads that were developed in parallel—data sourcing and preprocessing, architecture design, training/optimisation, and evaluation—so that the next project milestones can inherit the same assumptions, trade-offs, and metrics.

The following pages therefore do not attempt to summarize implementation status, but rather present the technical rationale gathered from the research documents. This creates a shared vocabulary for the final paper and situates every architectural choice within the broader discussion of dataset heterogeneity, regularisation, and explainability. The conclusion then revisits the alternatives discussed in the research and suggests how the ongoing implementation work could leverage those insights.

2. Data strategy and preprocessing

(Research lead: Julius Porzel)

Training a six-class emotion classifier on fixed 64×64 inputs raises two intertwined constraints: the images are too small to resolve fine-grained micro-expressions, and the taxonomy (happy, surprise, sadness, anger, disgust, fear) amplifies the already strong class imbalance once the neutral category is removed. The research therefore sets these constraints as first principles: geometric nuisance such as pose, scale, and skew must be eliminated before the model sees any pixels, the scarce discriminative information must not be wasted on uninformative details, and every optimisation step must try to prevent the model from collapsing onto the majority classes.

To reach competitive generalisation, multiple datasets are merged instead of relying on a single source. FER-2013 is retained for its structured, grayscale 48×48 frames but requires a careful upscaling step to match the 64×64 requirement. AffectNet-8 contributes hundreds of thousands of 96×96 in-the-wild color images with landmark annotations that are helpful for standardizing the face geometry, while RAF-DB adds further variance from real-world captures. Joining these datasets also creates secondary challenges: the distribution of samples across emotions becomes even more uneven (AffectNet supplies the bulk of the data) and color spaces or illumination statistics differ from FER-2013. Rather than attempting a naive concatenation, the pipeline orchestrates the sources so that the largest dataset is introduced last, allowing the model to first learn from data that closely resemble the FER-2013 target before being exposed to increasingly harder domains.

Every image is processed through a strict three-stage pipeline. RetinaFace locates the face bounding box and five landmarks (eyes, nose, mouth corners), also providing a dependable foundation for later demo tasks that overlay saliency on the live video. A similarity transform then stabilizes the face so that key points occupy canonical pixel locations, reducing the need for the subsequent network to be invariant to pose. For FER-2013 inputs, two specialised super-resolution options are considered: eigenface-domain SR prioritizes discriminative coefficients within a reduced dimensional face space, and zero-shot SR trains a tiny CNN

on the single input image at test time to adapt to its specific noise/artefacts. Both minimize blurring from naïve interpolation while retaining the richest features possible in the 64×64 budget.

Combining multiple domains also triggers "negative transfer," where highly divergent samples might degrade the core FER-2013 accuracy. The adopted solution is a progressive multi-source domain adaptation flow: start training with the subset of source samples most similar to the target distribution and progressively introduce harder domains only once the new data surpasses a relevance threshold. A density-aware memory keeps track of previously seen, beneficial samples so the network does not catastrophically forget earlier knowledge, ensuring that the training path remains smooth as new domains are added. The final stage, BORT^2 (Bi-level Optimization based Robust Target Training), fine-tunes the model on pseudo-labeled target samples, explicitly modeling label uncertainty via an entropy maximization regularizer so that noisy self-labels do not drown out the true expressions.

This multi-faceted strategy demands time and engineering effort, which is why the current implementation plan starts with a single dataset to validate the preprocessing and training loops. Once the baseline is stable, the progressive domain introduction and BORT^2 steps will be activated to assess whether the increased data variance actually translates into robustness gains for the downstream live classifier.

3. Model architecture and interpretability

(Research lead: Necati Deniz Baykuş)

The architecture chapter synthesizes the decisions described in the model architecture research. We base the classifier on a ResNet-18 backbone because its identity-preserving skip connections make it easier to learn discriminative features from constrained inputs, and the 11M parameter budget keeps inference lightweight for the planned video demo. Since the original ResNet is configured for 224×224 samples, the initial $7 \times 7,\ \text{stride} = 2$ convolution is replaced with a $3 \times 3,\ \text{stride} = 1$ kernel so that the receptive field grows more cautiously, preserving horizontal and vertical detail within the 64×64 grid. The remaining four residual stages keep their batch normalisation and ReLU activations, followed by global average pooling and a six-node softmax that emits the probabilities for happiness, surprise, sadness, anger, disgust, and fear.

Hyperparameters mirror the research recommendations: Adam (β_1=0.9, β_2=0.999) is the default optimiser coupled with ReduceLROnPlateau, while SGD with momentum stands ready if a slower, more controlled descent is needed. Weighted cross-entropy compensates for the imbalanced expressions, dropout (0.3) regularises the late blocks, and L_2 weight decay keeps magnitudes in check. These settings aim for stable learning rather than chasing aggressive fine-tuning results that could overfit to the more frequent classes.

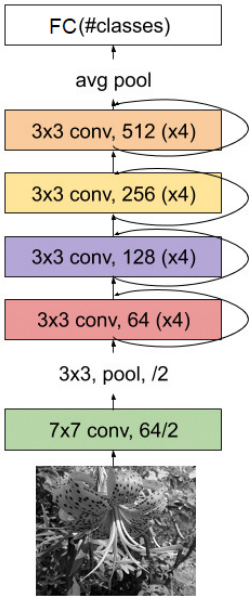

Figure 1. Adapted ResNet-18 backbone with a smaller initial kernel, residual stages, global average pooling, and six-output softmax tailored to the emotion taxonomy.

Grad-CAM is the explainability tool of choice because preserving spatial structure up to the final pooling simplifies gradient backpropagation, producing saliency maps that highlight semantically meaningful areas (eyes for surprise, mouth for happiness, brows for anger). These heatmaps are designed to be overlaid on the live video stream as described elsewhere, providing immediate interpretability.

4. Training and optimisation

(Research lead: Mehmet Berberoğlu)

Training mirrors the staged pipeline described in the optimisation research. After the preprocessing stack produces normalized 64×64 faces, mini-batches travel through the modified ResNet-18 and the softmax outputs feed into weighted cross-entropy. Training runs for multiple epochs with checkpointing to enforce a dedicated validation split, thereby exposing overfitting tendencies that often arise from dataset-specific artefacts or the limited expressive content of fixed-resolution faces.

Class imbalance is mitigated through the loss weights: each rare emotion is assigned a greater contribution so that its misclassifications change the gradient significantly. If weighting proves insufficient to enforce parity, the plan includes more advanced sampling strategies (e.g., oversampling minority classes or using focal loss variants).

ReLU activations throughout the network keep gradients well-behaved, while dropout in the final residual blocks and L_2 weight decay curb reliance on single neurons or runaway parameters when the training data are noisy or diverse. Together these measures prioritise stable, balanced learning rather than maximal but brittle accuracy.

5. Metrics and validation

(Research lead: Isabela Labus)

Depending on the chosen data sets, emotion classes like disgust, fear and anger might be underrepresented, making the class distributions imbalanced. Additionally, as some classes such as disgust and anger might be classified by the model as the same emotion, data difficulty factors such as class overlap contribute to imbalanced classes (Carvalho et al., 2025).

Taking class imbalance into account as well as class overlap possibly exacerbating class imbalance and worsening model evaluation (Carvalho et al., 2025), the performance of our classifier should be additionally measured by a macro-averaged F1-score, class specific macro-precision and macro-recall.

Accuracy will be one of the classification metrics used for our classification model as accuracy is one of the most utilized metrics for evaluating classifier performance, however, for imbalanced data sets, accuracy would not be sufficient (Juba and Le, 2019). The major problem is that in imbalanced classes its score is dominated by majority classes (Carvalho et al., 2025). Similarly, Juba and Le (2019) emphasize that a high accuracy score could be misleading when a dominant class has a stronger influence on the metric, which is why a poor performance of a classifier can likely not be caught by that score.

Therefore, accuracy will be included as a baseline metric in our project, but for reasons stated above, other metrics will be used as well.

Macro-averaged F1 could be a useful additional evaluation score because it makes no distinction between classes with high or low samples and all classes, regardless of size, contribute equally to the metric (Grandini et al., 2020). This would help our case tremendously. Macro-precision and macro-recall are computed first, then they are combined as a harmonic mean, which results in said F1-score (Grandini et al., 2020). Since precision and recall are calculated per class, I view them as additional insight where we can see which classes are over- or underpredicted; information that is neither provided by accuracy nor macro-F1. For this reason, they will be used in the evaluation step.

Lastly, we use hold-out as our validation strategy. Since the convolutional neural network classification model is trained from scratch on newly chosen data sets and the project has to be completed in a limited time frame, training the model and then testing it once is the most practical solution and an efficient evaluation strategy.

In hold-out validation, the data is divided into different training and validation subsets, while the validation set is not used during training (Bami et al., 2025). Testing the model on data that was not used during training helps prevent overfitting and provides a realistic estimate of how well the model performs on new data (Bami et al., 2025). And although hold-out validation can give slightly lower accuracy than k-fold cross-validation (Bami et al., 2025), it is a good choice for our project as it is easier to work with.

6. Conclusion

The preliminary report now captures the blueprint for the SEP Emotion Classifier: a multi-source data strategy that stabilizes and balances FER inputs, a modified ResNet-18 architecture that preserves spatial detail and supports Grad-CAM saliency, an optimisation pipeline that penalises misclassification of rare emotions, and a validation suite driven by macro-averaged metrics. The research emphasises discussion over final verdict: while combining FER-2013, AffectNet, and RAF-DB should increase variance, it also invites negative transfer, which is why the progressive MSDA and BORT^2 techniques exist alongside the more conservative single-dataset fallback. Similarly, the architecture and metric choices keep options open for lighter backbones or additional explainability layers once the current prototype matures. This document should therefore serve as a reference for future drafts, allowing implementation work to decide which of the discussed alternatives to activate, and helping the final report justify why certain research paths were chosen or deferred.