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 Affect-Net 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 captures the research strategy for classifying six emotions from 64×64 inputs, where every design choice balances the limited signal with the need for interpretable predictions of happiness, surprise, sadness, anger, disgust, and fear. Each section then presents the technical rationale from the supporting documents—data sourcing, architecture design, optimisation, and evaluation—so future work inherits the same assumptions and metrics.

Rather than recounting implementation steps, the document builds a shared vocabulary that frames architectural choices within dataset heterogeneity, regularisation, and explainability, and the conclusion revisits the discussed alternatives to guide the ongoing implementation.

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 [6, 10, 20].

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 [1, 5, 12]. 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 [3]. 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 [7, 17].

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 tar-

get 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 [2, 4].

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 work first compares three options—basic CNN, ResNet-18, and AlexNet—before settling on ResNet-18 due to its superior accuracy-to-parameter trade-off and its smoother internal representations that pair well with explainability tools [15, 19, 21]. ResNet's identity-preserving skips also make it easier to capture discriminative features even when only 64×64 pixels are available. Because the canonical ResNet-18 expects 224×224 inputs, the initial 7×7 stride-2 convolution is replaced with a 3×3 stride-1 kernel to keep the receptive field growth in check and to safeguard horizontal and vertical detail inside each low-resolution face [11]. The downstream residual stages keep their batch normalisation and ReLU activations, global average pooling reduces the spatial footprint, and the head is a six-node softmax emitting probabilities for happiness, surprise, sadness, anger, disgust, and fear.

Hyperparameters mirror the research recommendations: Adam ($\beta_1{=}0.9, \beta_2{=}0.999$) 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 [14]. These settings aim for stable learning rather than chasing aggressive fine-tuning results that could overfit to the more frequent classes.

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 sur-
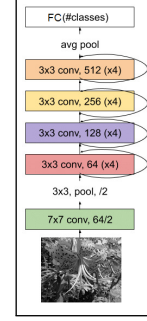

Figure 1. Adapted ResNet-18 backbone with a smaller initial kernel, residual stages, global average pooling, and six-output softmax tailored to the emotion taxonomy.

prise, mouth for happiness, brows for anger) [16]. 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, while Adam keeps the gradient steps adaptive to the heterogeneous samples [8]. 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). Adam remains the primary solver so that parameter updates adapt to the noisy, varied inputs without manual tuning [6]. Rectified Linear Unit (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 [9, 13, 18]. 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.

References

[1] Shuvo Alok. Raf-db dataset, 2025. Kaggle dataset. Accessed December 18, 2025. 1

[2] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2019. 2

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5203–5212, 2020. 1

[4] Zhongying Deng, Da Li, Yi-Zhe Song, and Tao Xiang. Robust target training for multi-source domain adaptation. *arXiv preprint arXiv:2210.01676*, 2022. 2

[5] Fatihkgg. Affectnet-8 dataset (yolo format), 2025. Kaggle dataset. Accessed December 18, 2025. 1

[6] K. Ruwani M. Fernando and Chris P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7): 2940–2951, 2021. 1, 2

[7] Bahadir K. Gunturk, Aziz U. Batur, Yucel Altunbasak, Monson H. Hayes III, and Russell M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–600, 2003. 1

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 2

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. Published as a conference paper at ICLR 2019. 2

[10] Fatma Mazen, Ahmed Nashat, and Rania Ahmed Abul Seoud. Real time face expression recognition along with balanced fer2013 dataset using cyclegan. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(6):153–158, 2021. 1

[11] Rohit Modi. Resnet — understand and implement from scratch. *Medium*, 2021. Reference from the Model Architecture research draft. 2

[12] Msambare. Fer-2013 dataset, 2025. Kaggle dataset. Accessed December 18, 2025. 1

[13] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010. 2

[14] Gaurav Kumar Pandeya, Khushi Mittal, Antra Bansal, and Sumit Srivastava. Fire detection with resnet 18: Comparative analysis across different hyperparameters. *ScienceDirect*, 2025. Reference from the Model Architecture research draft. 2

[15] Leonardo Reginato. A practical comparison between cnn and resnet architectures: A focus on attention mechanisms. *Medium*, 2024. Reference from the Model Architecture research draft. 2

[16] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 2

[17] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 1

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 2

[19] Milind Talele and Rajashree Jain. A comparative analysis of cnns and resnet50 for facial emotion recognition. *Engineering, Technology & Applied Science Research*, 15(2), 2024. 2

[20] Nursel Yalçın and Muthana Alisawi. Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques. *Heliyon*, 10(20):e38913, 2024. 1

[21] Wenxuan Zhang. Comparison of alexnet and resnet models for remote sensing image recognition. *Transactions on Computer Science and Intelligent Systems Research*, 5, 2024.